Proceedings of the 39th Chinese Control Conference
July 27-29, 2020, Shenyang, China

Research on Automatic Music Recommendation Algorithm Based on Facial Micro-expression Recognition

Ziyang Yu[1], Mengda Zhao[1], Yilin Wu[1], Peizhuo Liu[1], Hexu Chen[1]

1.College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
E-mail: 1113161395@qq.com

Abstract: In recent years, with the development and application of big data, deep learning has received more and more attention. As a deep learning neural network, convolutional neural network plays an extremely important role in face image recognition. In this paper, a combination of micro-expression recognition technology of convolutional neural network and automatic music recommendation algorithm is developed to identify a model that recognizes facial micro-expressions and recommends music according to corresponding mood. The facial micro-expression recognition model established in this paper uses FER2013 with a recognition rate of 62.1%. After identifying the corresponding expression, a content-based music recommendation algorithm is used to extract the feature vector of the song and a cosine similarity algorithm is used to make the music recommendation. This research helps to improve the practicality of the music recommendation system, and the related results will also serve as a reference for the application of the music recommendation system in areas such as emotion regulation.

Key Words: Deep learning, Facial micro-expression recognition, CNN, FER2013, CB, Music recommendation algorithm

1 Introduction

With the advent of the information age, deep learning is widely used in image recognition, image processing, and especially facial expression recognition. Face recognition has become a research hotspot in the field of human-computer interaction, but it still has limitations on the application of image processing results. Image research often focuses on improving the accuracy of recognition, and the data in the image lacks the use of secondary processing, that is, in the actual production and life process, the image information has not been completely and efficiently used [1].In this paper, a deep learning method is used to design and train a convolutional neural network expression recognition model. The results of image processing are combined with a music recommendation algorithm, and the music that adjusts the mood is recommended by judging the mood shown by the person. Music data sets are created by crawling the playlists and manual annotations of major music websites. The scope of application of image processing results has been appropriately expanded.

2 Facial Micro-expression Recognition Based on CNN Neural Network

2.1 Basic Steps for Micro-expression Recognition

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

1. Obtain micro facial expression images of human faces and pre-process the images;

2. Perform micro-expression detection and related feature extraction;

3. Perform micro-expression classification.

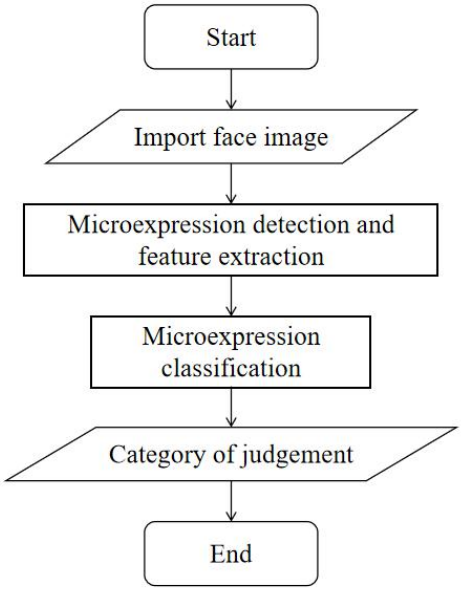

Figure 1: Flow Chart of Micro-expression Recognition

2.1.1 Image Pre-processing

An important step in the facial micro-expression recognition system is the pre-processing of facial images. Due to various factors, the quality of the input image (size, pixels, gray value, etc.) is not uniform, and the image cannot be directly used in the subsequent face recognition process, which greatly affects the recognition performance. Therefore, the image pre-processing has standardized and unified standards, which can eliminate the impact on the recognition performance in terms of size, posture, brightness and darkness. Minimize the interference of irrelevant information as much as possible, thereby maximizing the recognition rate of micro-expressions. There are many preprocessing methods, such as histogram equalization, median filtering, grayscale stretching, homomorphic filtering, nearest neighbor algorithm, bilinear interpolation, etc [2-4]. Different micro-expression recognition systems require different image sources and recognition algorithms,

1. Project 201910145145 Supported by National Training Program of Innovation and Entrepreneurship for Undergraduates.

2. Supported by "the Fundamental Research Funds for the Central Universities(N182410001)".

and different pre-processing methods. This paper uses the FER2013 facial micro-expression data set to eliminate the image preprocessing process.

2.1.2 Micro-expressions and Feature Extraction

Feature extraction is a vital part of micro-expression recognition. An effective feature extraction method can obtain complete and accurate feature information for distinguishing different types of objects. It can also reduce the dimensionality of data and avoid the interference of irrelevant information. The micro-expression recognition model in this paper uses the deep feature extraction method.

With the popularization and development of big data, deep learning technology has gradually entered our field of vision, and at the same time, deep learning networks are also of great significance to achieve micro-expression recognition. Traditional facial expression recognition algorithms have poor robustness and insufficient expression feature extraction capabilities. Convolutional neural network-based facial expression recognition algorithms can make up for these shortcomings [5]. 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.

2.2 Convolutional Neural Network (CNN)

CNN neural network refers to convolutional neural network, more complex features of the image can be extracted and learnt by it . It is similar to other neural networks in that it uses forward propagation to input data and activate the hidden layer; backpropagation loss values modify the parameters of the hidden layer. The activation of the hidden layer depends on the activation function, which introduces non-linear factors so that the neural network can arbitrarily approach any non-linear function.The difference between CNN and other neural networks is that it is a deep neural network. After several convolutional layers and pooling layers are alternately connected, the fully connected layer is connected, thereby replacing the hidden layer within the network [6]. However, as the number of network layers and learning parameters increase, the risk of fitting during the training process will increase, and the generalization ability of the model will become worse. Two 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 [7].

Convolutional layer: several convolutional units make it, and the parameters of each convolutional unit are optimize d by back propagation algorithm. The convolution layer per forms a convolution operation on the original image and a s pecific feature filter. During multiple convolution processe s, each operation uses a different filter to map different feat ures.

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 parameters and calculations will also decrease. To some extent, it also controls overfitting.

Fully connected layer: After flattening the previous results, each node is connected to all nodes in the previous layer, which is used to integrate the features extracted before.

2.3 Model Design

Based on the training data volume and training category information of facial micro-expression tasks, this paper designs an 8-layer small CNN model with 4 convolutional layers, 2 pooling layers, and 2 fully connected layers. The model design is based on the improvement of AlexNet. It uses a convolution method of two convolution layers followed by a pooling layer to improve the feature extraction capability of the CNN network, and performs feature dimensionality reduction to improve the model training speed. Add Relu as the activation function after the layer, and finally pass the two fully connected layers and the Softmax classifier to map the extracted features into the classification probability.

2.3.1 Model Structure

The facial micro-expression CNN model structure designed in this paper is shown in Figure 2.

Among them, 48 * 48 represents the grayscale image data with the input of the network as 48 * 48 pixels, and 64 & 48 * 48 represents the output of the first convolution layer as 64 feature map data with the size of 42 * 42 pixels. The output of the accumulation layer and pooling layer are also multiple feature map data. The last two fully connected layers indicate that the number of neurons in this layer is 512 and 7, respectively. The model parameter settings are shown in Table 1.

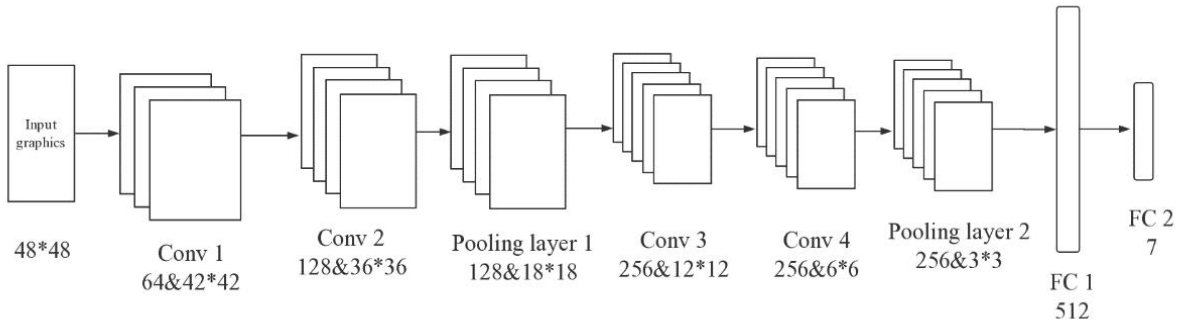

Figure 2: Structure of Facial Micro-expression CNN Model

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

2.3.2 FER2013 Data Set

This model uses FER2013 data set for model training. The FER2013 data set consists of 35886 facial expression pictures. Among them, there are 28708 training pictures, 3589 public test pictures and 3589 private test pictures. Each picture is composed of a grayscale image with a fixed size of 48 * 48, which is a pose. And unposed avatars, there are a total of 7 expressions, corresponding to the number tags 0-6, the specific expressions corresponding to the tags and Chinese and English are as follows: 0 Angry; 1 Disgust; 2 Fear; 3 Happy; 4 Sad; 5 Surprise; 6 Neutral [8]. Among them, happy mood is the most common.

2.3.3 Model Training

CNN is a feed-forward neural network, so the training of CNN is divided into two processes of forward propagation and back propagation [9]. During the forward propagation process, each neuron in each convolution kernel of the convolutional layer and the forward the local and local visual fields of the input feature map of the layer are connected, and the convolution operation is performed to extract the features of this part. After adding a bias, the result is passed through the activation function as an output to form the neurons of the current layer. These neurons constitute the current layer. Multiple feature maps for different features. The calculation method of the convolution layer is as follows:

$$y_j^l = f(\sum_{i=1}^{N_j^{l-1}} w_{i,\ j} x_i^{l-1}) + b_j^l, j = 1,2,...,M \quad (1)$$

Among them, l is the current layer number, $l-1$ is the previous layer number, y_j^l is the j-th feature map of the current layer, N_j^{l-1} is the number of feature maps of the previous layer connected to the j-th feature map of the current layer, $w_{i,\ j}$ is the j-th feature map of the current layer Convolution kernel of the feature map and the i-th feature map of the previous layer, x_i^{l-1} is the i-th feature map of the previous layer, b_j^l is the offset of the j-th feature map of the current layer, M is the number of feature maps of the current layer, and f is a non-linear activation function, the ReLU activation function is used in this paper, and its formula is as follows:

$$f(x) = \begin{cases} 0, x \le 0 \\ x, x > 0 \end{cases} \quad (2)$$

Practice has proved that the ReLU activation function can well solve the problem of disappearance of gradients in backpropagation and can accelerate the convergence speed [10].After the neural network has performed convolution, activation, and pooling operations for many times, the output of the final forward propagation can be expressed by the following formula:

$$y = \mathrm{Softmax}(wx + b) \quad (3)$$

Among them, x is the neuron input, y is the neuron output, w is the weight matrix, and b is the bias matrix. The Softmax function in TensorFlow is used to assign probability to different objects.

The back-propagation phase is mainly used to calculate the model error loss, and then the error is back-propagated and the optimizer is used to optimize the error. This paper selects the cross-entropy loss(Cross Entropy， CE)as the model's loss function [11], and its formula is as follows:

$$Loss = -\sum y\_ * \log(y) \quad (4)$$

Where y_ represents the known label and y represents the predicted value. To optimize the loss value, this article uses the mini-batch gradient descent (Mini Batch Gradient Descent, Referred to as MBGD) method. The idea is to use batch_size (the amount of data for each batch during training) to update the parameters for each iteration. The optimization method mainly has 2 Advantages: Through matrix operations, optimizing neural network parameters on one batch at a time is not much slower than a single data; using one batch each time can greatly reduce the number of iterations required for convergence, and at the same time can make convergence to The results are closer to the effect of gradient descent [12]. The entire model training process is to iteratively extract features, calculate losses, and optimize losses in batches until the model reaches the optimization stop condition.

3 Music Recommendation Algorithm

3.1 The Establishment of Song Library

Use Python to crawler music website to store song files and song information. The song information is stored in Excel form. Based on the seven kinds of emotions(anger, dispute, bear, happy, sad, secure, neutral) that can be recognized by the emotion recognition model, the emotion analysis and classification are carried out for the crawled songs, and then the song library corresponding to different emotions is established.

The analysis method of song emotion in this paper can be divided into two steps: lyrics data mining and song emotion analysis.

3.1.1 Data Mining of Lyrics

The steps of data mining for lyrics[13] are as follows (take Chinese songs as an example, the same treatment for English songs):

a.Word segmentation

In this paper, we use the open-source Chinese word breaking system Jieba as a tool to segment the lyrics.

b.Comparison with emotional lexicon

In this paper, we draw lessons from the lexicon established by others — — a Chinese emotional lexicon integrated with HowNet and Sogou as a comparison model.

In the lexicon, words are first divided into two sets with opposite meanings, that is, the set containing positive emotional words (such as "happy", "happy", "warm", etc.) and the set containing negative emotional words (such as "degenerate", "abandonment", "despair", etc.). In order to distinguish the emotion of words more accurately, this paper integrates emotion words, degree words (such as "good" in "good happiness") and negative words to compare the lyrics and find the emotional characteristic values in the lyrics.

3.1.2 Emotional Analysis of Songs

This paper uses SVM model in emotion classification. The general principle of SVM is as follows (schematic diagram is shown in Figure3):

We need to classify the black and white points in the graph. SVM will find the boundary between the plane where the black point is and the plane where the white point is, such as the dark line in the graph. The two light lines are used to indicate the distance w between the points and the black line in the two planes, and the goal of SVM is to find the black line with the largest distance Max (w). This paper uses LIBSVM (Chang et al.,2011), a SVM tool set developed by Professor Zhiren Lin of Taiwan University,to classify the training with R language.

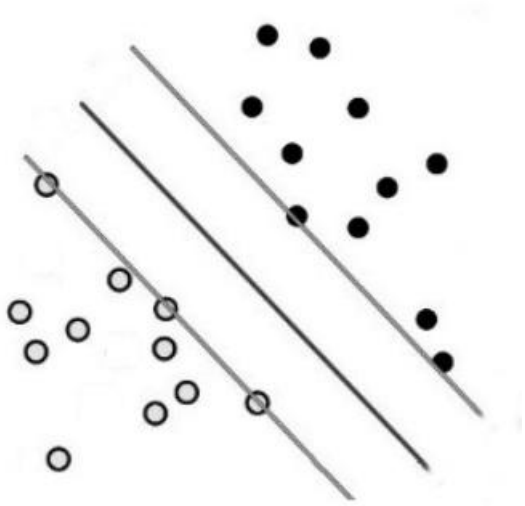

Figure 3: Schematic Diagram of SVM Classification

Because this paper involves seven emotions, it needs to be classified many times. First, all categories are divided into two subcategories (such as positive emotion and negative emotion), and then the subcategory is further divided into two sub categories, so as to cycle until a separate category is obtained.

3.2 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 items to users based on their past preferences.

Under the condition that the song library has been established and facial emotions have been determined, due to the difference in the emotions expressed by different songs in the same emotion, the random recommendation of the existence of the recommended songs cannot relieve negative emotions or promote The role of positive emotions, so the original random recommendation algorithm was improved to a content-based recommendation algorithm.

3.2.1 General Steps of Content-based Recommendation

The general steps of content-based recommendation are as follows:

1. Item Representation：Some features are extracted for each object (in this paper, each song) to represent the object;

2. Profile Learning：The feature data of several objects that the user likes is used as the user's preference feature;

3. Recommendation Generation： By comparing the characteristics of users and candidate objects obtained in the previous step, Recommend the most relevant objects for users.

The algorithm flow chart applied in this paper is shown in Figure 4:

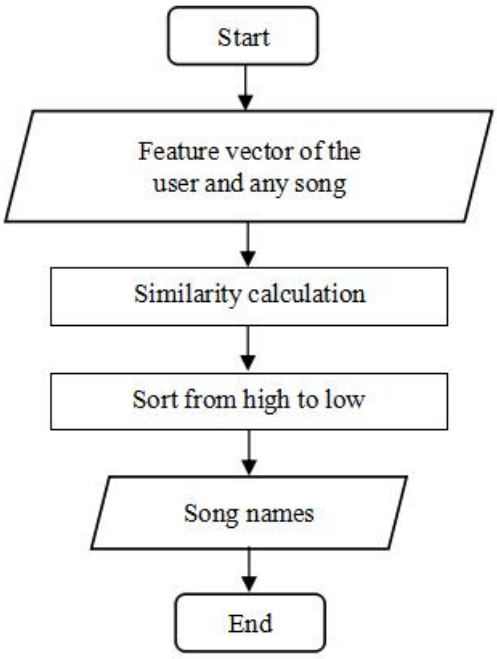

Figure 4: Flow Chart of Content-based Recommendation Algorithm

3.2.2 Feature Vector Extraction of Songs

The user selects three typical songs corresponding to each emotion in the song library. Organize the feature labels of the three songs and record them without repetition as：A_1, A_2, A_3, ..., A_n， record the occurrence times of each feature as：a_1, a_2, a_3, ..., a_n, then get the feature vectors of the user:

$$[(A_1, a_1), (A_2, a_2), (A_3, a_3), \ldots, (A_n, a_n)],$$

the feature vectors of any other song：

$$[(A_1, b_1), (A_2, b_2), (A_3, b_3), \ldots, (A_n, b_n)]$$

3.2.3 Cosine Similarity Algorithm

Note the similarity as Y. Use cosine similarity algorithm [14]. The calculation formula is as follows：

$$\gamma = \cos^{-1}\left(\frac{\sum_{i=1}^{n}(a_i \times b_i)}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}}\right) \quad (5)$$

The larger the final calculation result is, the higher the similarity is, indicating that the song is more suitable for users' current emotions. All songs corresponding to different emotions in the song library are calculated and recommend -ed according to the order of similarity from high to low.

4 Experiment and Result Analysis

This experiment is carried out on the platform of TensorFlow deep learning framework, and the experimental data set is FER2013 [15] public data set.

4.1 Experimental Results Analysis of CNN Model of Facial Micro-expression

The number of training rounds is set to 500, and the learning rate is 0.0001. After several parameter adjustment experiments, the CNN model of facial micro expression gets 62.1% recognition rate on FER2013 data set. Looking up the related literature [16-17], we found that most of the accuracy is about 60%. The visualization operation after network training is shown in the following figure:

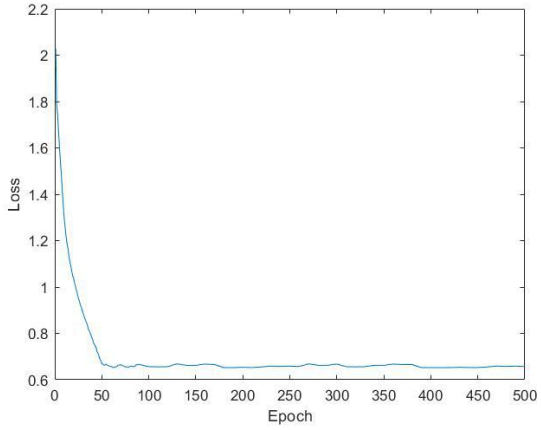

Figure 5: Loss Function Curve

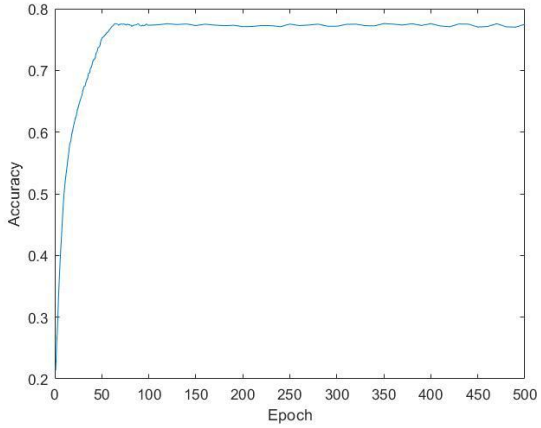

Figure 6: Recognition Accuracy Curve

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 batch. Figure 6 shows the training set recognition accuracy curve, the horizontal axis Epoch represents the number of training iterations, and the vertical axis Accuracy represents the batch recognition accuracy. It can be seen from the training curve that the network training process eventually stabilizes, the model converges well, the final loss value approaches 0.65, and the training set model recognition accuracy rate is above 77%.

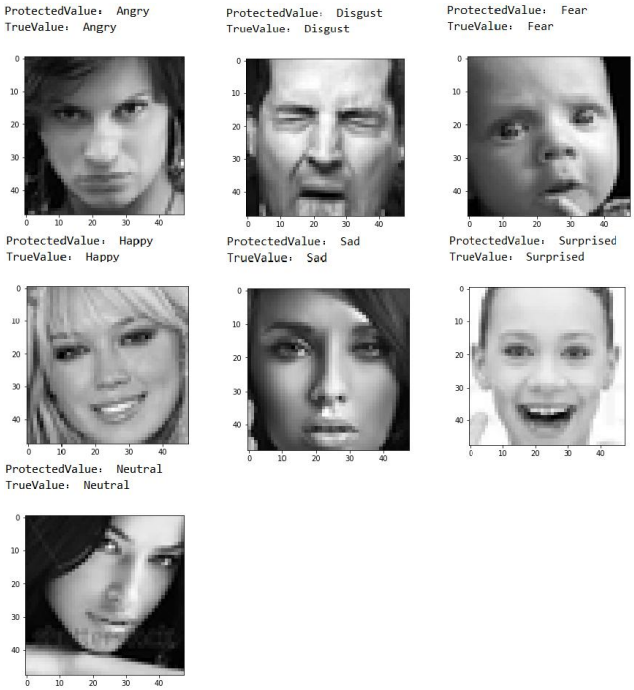

Figure 7: Examples of 7 Expressions Recognized by the Model

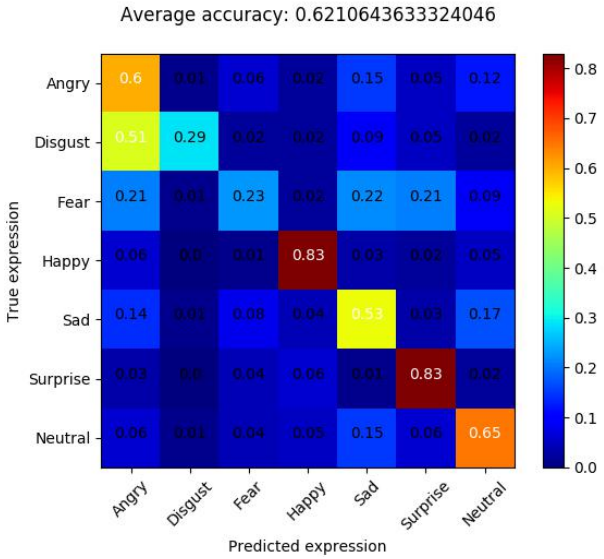

Figure 8:The Confusion Matrix on the FER2013 Dataset

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:

(1)There are wrong labels in FER2013 and its image resolution is low(48*48pixels). The accuracy of human eyes in this data set is also only 65% ± 5%[18].
(2)As can be seen from Figure 7-Figure 9, the similarity between 'Disgust' and 'Angry' is very high. The number of 'Disgust' in the training set is the least and 'Fear' has the problem of low recognition, so the training effect is not very good.

Figure 9: Number of expression labels in training set

In order to verify the performance of the model in this paper, the method of this paper is compared with other methods on the FER2013 test set. The experimental results are shown in Table 2.

Table 2. Comparison of Recognition Rate Between this Model and Traditional Model

Method	Test Pictures Number	Correct Number	Recognition Accuracy
Gabor+SVM[19]	200	99	49.5%
LBP+SVM[20]	200	107	53.5%
Transfer Learning[21]	200	103	51.5%
Mapped LBP[22]	200	92	46%
Inception[23]	200	113	56.5%
Model in this paper	200	125	**62.5%**

The first two types in the table are the traditional machine learning expression recognition method that first extracts the feature expression and then combines the feature classifier. Under the same test data, the model in this paper has a recognition accuracy of 13% higher than Gabor + SVM and 9 higher than LBP + SVM. %. Compared with other deep learning models, this model also has a better recognition effect, which verifies the effectiveness of this model.

4.2 Results Analysis of Music Recommendation Experiment

4.2.1 Design of Experiment

The music song recommendation experiment in this paper is mainly carried out in PyCharm integrated development environment (Python version is 3.6.8), and the data set used is the song library crawled from the network in 3.1. Process of experiment refers to Section 3.1,3.2 above.

4.2.2 Analysis of Experiment

Take the Happy Song Library as an example, select three typical songs in the Happy Song Library, and analyze and sort the feature labels using the methods of 3.1.1, 3.1.2. The results are shown in Table 3: ('1' means the song contains the label, '0' means not).

Table 3. Typical Songs and Labels

Label / Songs	Cheerful	Wild	Keyboard	Absolute Music	Interenet
Back wave	1	1	1	1	1
Rise	0	0	1	1	1
Lanterns	1	0	1	1	1

The feature vectors of the user are as follows: [(Keyboard,3), (Absolute Music,3), (Internet,3), (Cheerful, 2), (Wild,1)].

Then cosine similarity algorithm is applied to calculate cosine similarity of other songs. The recommended results are shown in Table 4('C' stands for Cheerful label, 'W' stands for Wide label, 'K' stands for Keyboard label, 'I' stands for Internet label, and 'Cos' stands for Cosine Similarity):

Table 4. Results of Song Recommendation

Label / Songs	C	W	K	A	I	Cos
Brass	1	0	1	1	1	0.972
Rise	1	0	1	1	1	0.972
PDD	1	0	1	1	0	0.816
Higher	1	0	0	1	0	0.625
All in	1	0	0	1	0	0.625
Adventures	0	0	0	1	0	0.531
Temptation	1	0	0	0	0	0.177
Light	1	0	0	0	0	0.177

Cosine similarity is used to measure the difference between a song and a typical song. Theoretically, the closer the value is to 1, the more applicable the song is to the user's current mood. After cosine similarity is obtained, songs are recommended according to the order of similarity from high to low.

5 Summary

In this paper, we proposed a model of facial micro-expression recognition based on convolutional neural network (CNN). After training the model on FER2013 data set, we got a recognition rate of 62.1%. On the basis of the state that facial expression and emotion were both recognized, the content-based recommendation algorithm was applied to automatically recommend music for users. Compared with the existing algorithms that only recommend music according to the users' past listening preferences, the algorithm proposed in this paper increases the user's emotion recognition, so that the recommended music can better meet the users' listening needs. Therefore this algorithm has a relevantly promising application market. Although we have made some achievements, still some problems need to be solved . For example, the accuracy of microexpression emotion recognition needs to be improved. In the follow-up work, the recognition rate of tags with low recognition will be improved, and the music recommendation algorithm will be further optimized and improved.

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 Siquan, 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. Multi-angle 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 LeNet-5 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 FRR-CNN [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 Zhaoet al. Facial expression recognition based on transfer learning from deep convolutional networks[C]. 2015 11th Int. Conf. Nat. Comput.,2015: pp.702 - 708

[16] Xu Linlin, Zhang Shumei, Zhao Junli. Expression recognition algorithm for constructing parallel convolutional neural networks [J]. Journal of Image and Graphics, 2019, 24 (02): 0227-0236.

[17] Zhai Junkui, Liu Jian. Research on Transfer Convolutional Neural Network for Facial Expression Recognition [J]. Signal Processing, 2018, 34 (6): 729-738.

[18] Kung, Sheng H. Facial expression recognition using Opti- cal Flow and 3D HMM and human action recognition u- sing cuboid and topic models[D]. Oakland USA: Oak- land University. 2016.

[19] Li Wenshu, He Fangfang, Qian Yitao, et al. Facial expression recognition based on Adaboost-Gaussian process classification [J]. Journal of Zhejiang University (Engineering Science), 2012 (1): 84-88.

[20] Xu C，Dong C，Feng Z，et al．Facial Expression Pervasive Analysis Based on Haar-Like Features and SVM// Contemporary Research on E-business Technology and Strategy[M]. Springer Berlin Heidelberg, 2012: 521-529.

[21] Ng H W,Nguyen V D,Vonikakis V,et al. Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning[C]//ACM International Conference on Multimodal Interaction. Shanghai China: ACM, 2015: 443-449.

[22] Levi G, Hassner T. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns[C]//ACM on International Conference on Multimodal Interaction. Shanghai China: ACM, 2015:503-510.

[23] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks [J]. Computer Science,2015:1-10.