IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 30, 2019, accepted December 16, 2019, date of publication December 20, 2019,
date of current version December 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2961161

The Application of a Hybrid Transfer Algorithm Based on a Convolutional Neural Network Model and an Improved Convolution Restricted Boltzmann Machine Model in Facial Expression Recognition

YINGYING WANG[1], **YIBIN LI**[1], **YONG SONG**[2], **AND XUEWEN RONG**[1]
[1]School of Control Science and Engineering, Shandong University, Jinan 250061, China
[2]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China

Corresponding author: Yong Song (songyong@sdu.edu.cn)

This work was supported by the National Nature Science Foundation of China Grant 61673245 and Grant 61573213.

ABSTRACT As an important part of emotion research, facial expression recognition is a necessary condition for intelligent interaction between human and machine, which has an important research significance and a potential commercial value. Convolutional neural network (CNN) is an effective method to recognize facial emotions, which can perform feature extraction and classification simultaneously, and can automatically discover multiple levels of representations in data. Due to the fact that there are millions of parameters involved in training the convolutional neural network model, and there is a large demand for marked samples, transfer learning is often used to fine-tune the pre-trained model for a small target sample set. However, there are often some content differences between data sets during deep transfer learning, which will affect the recognition ability of feature extraction. In order to improve the facial expression recognition ability in transfer learning, a hybrid transfer learning model based on an improved convolution restricted boltzmann machine (CRBM) model and a CNN model is proposed in this paper. This method is fused by two learning abilities of these two models. When the pre-trained CNN model is transferred to a small target set, the CRBM is used to replace the full connection layer in the CNN model, and the CRBM layer and the sofmax layer will be retrained on the target set. The added CRBM layer can not only fully connects all feature maps, but can also learn about the unique statistical characteristics about the target set, which eliminates the influence of content differences between data sets, and extracts higher-order statistical features of facial expression images from the target set. The proposed method is evaluated based on four publicly available facial expression databases: JAFFE, FER2013, SFEW and RAF-DB. The new method can achieve better performance than most state-of-the-art methods, and it can effectively prevent the negative influence of transfer learning features between different data sets.

INDEX TERMS Facial expression recognition, convolutional neural network, convolution restricted Boltzmann machine, transfer learning.

I. INTRODUCTION

Facial expression recognition plays a central role in human-computer interaction. In a basic communication course, 55% of the information is conveyed by different facial expressions, voice constitutes 38% of the information, and language only constitutes 7% of the communication information [1], therefore facial expression recognition has attracted much attention in recent years [2], and it has many important applications such as remote education, safety, medicine, psychology and human-robot interaction systems. Although great progress has been made [3], it is difficult to acquire a facial

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

expression recognition system with a satisfactory accuracy rate due to a variety of complex external conditions such as: head pose, image resolution, deformations, and illumination variations. Hence, facial expression analysis is still a challenging work. Generally, facial expression recognition is composed of three steps: preprocessing, feature extraction and classification. Feature extraction is a key step in the whole recognition work, and the feature that we expect should minimize the distance of within-class variations of expression while maximizing the distance of between-class variations [4]. If features are inadequate, even the best classifier would fail to achieve good performance. Among machine learning algorithms, features are extracted by hand, such as local binary patterns (LBPs) [5], Gabor [6], local Gabor binary patterns (LGBPs) [7], scale invariant feature transforms (SIFTs) [8], and histograms of oriented gradient (HOG). After feature extraction, the classification method should be applied to perform facial expression recognition, such as SVM, random forest, sparse coding, neural network, etc. Although these methods have achieved great success in specific fields, most methods can only obtain the low-level features, and can't obtain the high-level semantics.

To cope with the above disadvantages, deep learning methods [9] are considered, especially the emergence of convolutional neural networks [10]. Convolutional neural network is a very effective method to recognize facial emotions. They can perform the feature extraction and classification process simultaneously [11], and can automatically discover multiple levels of representations from data. This is why they succeed in breaking the most world records in recognition tasks. The structures of early convolutional neural networks are relatively simple. With the development of the relative research, the structure of convolutional neural network has been continuously optimized and its application fields have been extended. In recent years, the research on the structure of convolutional neural network is still very hot, and some network structures with excellent performance have been proposed. The research results of the convolutional neural network in various fields make it one of the most concerned research hotspots.

Ali Mollahosseini1 et al. [12] proposed a deep neural network architecture to address the facial expression recognition (FER) problem, and verified the proposed structure through multiple standard face datasets, viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013. Guihua Wen et al. [13] proposed a method for facial expression recognition by integrating many convolutional neural networks with probability-based fusion. Yu [14] proposed a method based on the ensemble of three state-of-the-art face detectors. Jung *et al.* [15] proposed a new CNN method based on two different models. The first deep network model can extract temporal appearance features from image sequences, while the other can extract temporal geometry features from temporal facial landmark points.

Although the CNN model has many advantages over the traditional machine learning algorithm for the facial expression recognition task, a large number of labeled samples are need for training the CNN model, which requires a lot of manpower and energy in reality. The above problem has been solved with the coming of transfer learning [16], which can take a CNN model that trained by a big data set as a feature extractor of the bottom or middle layer of the target set, and the last few layers were modified as the adaptive layer. In the training process, the transfer learning algorithm realizes a direct application of the pre-training model on a small sample set.

Choi *et al.* [17] presented a transfer learning approach for music classification and regression tasks. The convnet feature outperforms the baseline MFCC feature in all the considered tasks and several previous approaches that are aggregating MFCCs as well as low- and high-level music features. Phillip M et al. [18] evaluated transfer learning with deep convolutional neural networks for the classification of abdominal ultrasound images. HaijunLei [19] proposed a framework based on a very deep supervised residual network (DSRN) to classify HEp-2 cell images, and developed a cross-modal transfer learning strategy. They pre-trained ICPR2012 dataset to fine-tune ICPR2016 dataset based on the DSRN model since both datasets are similar.

Although the transfer learning algorithm solves the fitting problem of small samples in training one CNN model, the ability of the feature recognition will reduce because the content differences between the two data sets. In order to solve the above problems, this paper proposes a hybrid transfer learning model based on an improved Convolutional restricted Boltzmann machine (CRBM) and a CNN model. When the trained CNN model is transferred to the target set, the full connection layer of the traditional convolutional neural network model will be taken as the CRBM model.

The paper is arranged as follows: After this introduction, Section 2 presents the detailed structure of the CNN model. Section 3 focuses on the introduction of the traditional deep transfer learning algorithm. Section 4 presents the detailed structure of the CRBM model. Section 5 focuses on the new hybrid deep transfer learning method. Section 6 presents the experiments and its results. Finally, Section 7 summarizes and concludes this paper.

II. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network is a non-fully connected multilayer neural network, which is generally composed of convolution layer (Conv), down-sampling layer (or pooling layer) and full-connection layer (FC). Firstly, the raw image is convoluted by several filters on the convolution layer, which can get several feature maps. Then the feature is blurred by the down-sampling layer. Finally, a set of eigenvectors is get through a full connection layer. The architecture of convolutional neural network is shown in figure 1.

A. CONVOLUTIONAL NEURAL NETWORK

In convolutional layer, multiple convolutional kernels f_k with a kernel size n * m applied to the input x in order to calculate

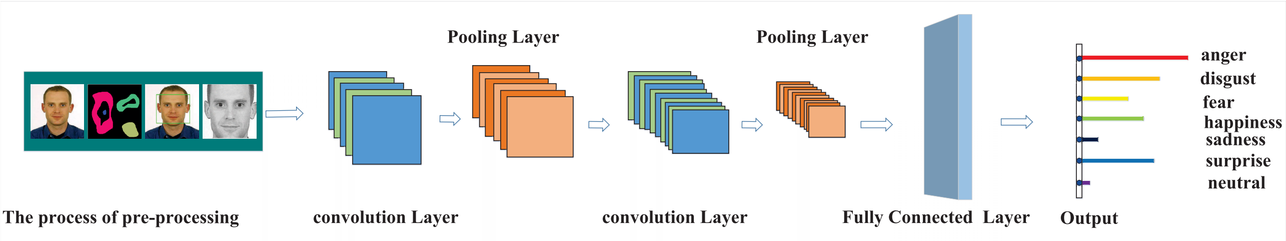


FIGURE 1. The basic structure of a convolutional neural network model.

a more rich and diverse representation of the input. It is not sufficient to have only one convolution kernel for feature extraction, hence multiple convolution kernels can be used in this step. If there are 50 convolution kernels, 50 features will be learned correspondingly. No matter how many channels there are in the input image, the total number of channels in the output image is equal to the number of convolution kernels.

B. POOLING LAYER

The main function of the pooling layer is to lower sampling, and further to reduce the number of parameters by removing unimportant samples in feature map. A large image can be downsized by the pooling layer, as well as can retain many important information. There are many methods for pooling operation: max pooling, mean pooling and etc, whereas max pooling is the most commonly used method. In fact, max pooling is to take the maximum value of n * n samples as the sample value.

C. ACTIVATION FUNCTION

In the process of facial expression classification based on convolutional neural network, the selection of activation function plays a great role in the whole system, which is mainly used to introduce nonlinear factors. The sigmoid function, tanh function and rule function are commonly used. And it turns out, the sigmoid function has a matter of disappearing gradients, and its calculation is expensive. When the sigmoid function approaches 0 or 1, the gradient approaches 0. Neurons with output values close to 0 or 1 will reach saturation state, therefore the weights of these neurons will not be updated, and the weights of neurons adjacent to such neurons also update slowly. When there are many neurons in this situation, the network will be unable to carry out the back propagation. Tanh function is an updated version of the sigmoid function on the range, and the disadvantage is also the gradient disappearance problem. ReLu function is more efficient than most other activation functions. It has a relatively cheap computation, because no exponential function has to be calculated. This function also can prevent the vanishing gradient error, since the gradients are linear functions or zero but in no case nonlinear functions.

D. FULLY CONNECTED LAYER

The fully connected layer connects all neurons of the prior layer to every neuron of its own layer.

III. THE TRANSFER LEARNING BASED ON DEEP CONVOLUTIONAL NEURAL NETWORK

Under the traditional machine learning framework, the learning task is to learn a classification model based on the given sufficient training samples, and then use the learned classification model to classify and predict the test samples. This learning process needs to be supported by a large number of labeled training samples. However, a large number of training data is very difficult to obtain in reality, because labeling training samples and test samples require a large amount of manpower and material resources. The transfer learning method allows transfer the existing knowledge to solve the learning problem of the target field with only a small number of labeled sample data, which can be seen in figure 2.

Transfer learning can be defined as: Given a source domain D_s and Learning task T_s, a target domain D_T and Learning task T_T. The purpose of transfer learning is to use the existing knowledge in D_s and T_s to improve the learning ability of prediction function $f_T(\cdot)$ in the target domain D_T

The transfer learning based on the deep convolutional neural network means that the pre-trained deep CNN model is re-trained on the data set of the new target task. This process can also be called network fine-tuning. In the network model that pre-trained by large-scale image data set, the hierarchical structure before the classifier is used as a general feature extractor, and the test images input into the CNN model will generate a depth feature vector, which has a strong generalization ability.

IV. THE INTRODUCTION AND IMPROVEMENT OF THE CONVOLUTIONAL RESTRICTED BOLTZMANN MACHINE

RBM [20] is a statistical neural network model based on probability, and its structure has two layers, which meet the full connection between layers and no connection within layers. RBM can encode the original m-dimensional input data and transform them into n-dimensional output data, which is another deep expression feature of original m-dimension data. This form is a nonlinear structure, and it can effectively extract features and obviously improve the recognition ability of the network. Figure 3 shows the general structure of RBM.

According to figure 3, the energy function can be defined:
$E(v, h) = -\sum_{i=1}^{N} v_i b_i - \sum_{i=1}^{N} h_j c_j - \sum_{i=1}^{N} \sum_{j=1}^{M} W_{ij} v_i h_i$

In which N represents the number of visible layers, and the status is v, M stands for the number of hidden layers, and the status is h. c_j represents the j_{th} bias value, v_i is the state of the i_{th} visible unit, h_j stands for the state of the j_{th} hidden unit.

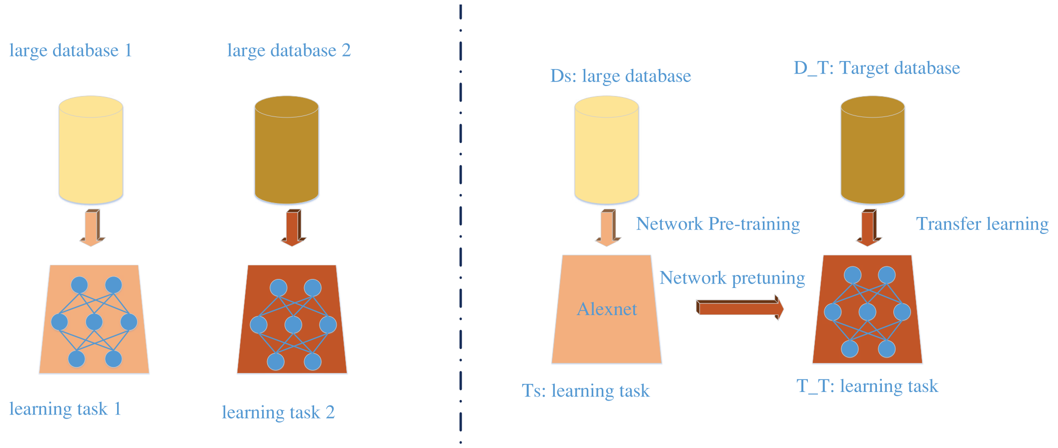


FIGURE 2. The flow charts of two different learning methods:the left is the traditional machine learning method, the right is the transfer learning algorithm.

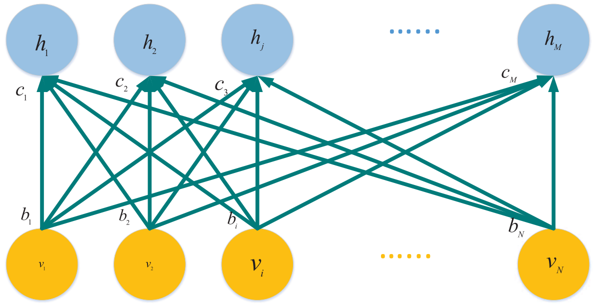


FIGURE 3. The structure of RBM.

W_{ij} is the weight value of the connection between the visible layer v_i and the hidden layer h_j.

RBM model has been widely used in many machine learning fields such as data dimensionality reduction, speech recognition and image processing due to its simple artificial neural network structure and fast learning algorithm. At present, the image processing methods based on RBM usually adopt two ways to construct the model: one is directly making each pixel in the image to correspond to a visible unit; the other is using multiple features of vectorization as visible elements.

The disadvantages of these two methods are as follows: (1) The model can only process small images, and it is difficult to process large images; (2)The selected features are greatly influenced by personal experience, which has poor flexibility; (3) The image is generally high dimensional, and the calculation of the algorithm should be simple, which is the opposite of RBM; (4) The useful objects are distributed locally in the image, and the feature representation is required to be invariant to the local transformation of the input.

To solve these problems, Lee et al. proposed the Convolution Ronstrained Boltzmann Machine (CRBM) model [21]. introduced. CRBM is a hierarchical generation model with translation invariance that combines CNN model with RBM model, which supports the probabilistic inference of top down and bottom up, and the probability maximum pooling method is used for dimension reduction and regularization operations. CRBM model effectively uses the convolution kernel, therefore it has better performance in image processing tasks. The features learned from the CRBM model will reflect the real information of the images more objectively, and improve the accuracy of image classification.

CRBM is a new breakthrough of RBM. It takes the advantage of weight sharing between filters and the image convolution operation to reduce the parameters of the model. Firstly, image features are extracted through convolution operation, and then features are further extracted features with translational invariance through probability maximum summation operation.

CRBM is very similar to RBM, and this model consists of three layers: visible layer, hidden layer, and pooling layer. Figure 4 shows the general CRBM structure and the detailed operation process. Suppose that the visible layer of CRBM is the binary matrix V of $N_v \times N_v$, and contain K convolution kernels of $N_W \times N_W$. The hidden layer is composed of K feature mapping surfaces with the size of $N_h \times N_h$, and there are ${N_h}^2K$ neurons in the hidden layer. Each binary array $N_h \times N_h$ is connected to a convolution kernel of size $N_W \times N_W$ ($N_W \triangleq N_v - N_h + 1$). Weights are shared in cells of the same sub-hidden layer. b and c are shared biases in sub-hidden layers and the visible layer, respectively. v_{ij} represents the input value of the i visible layer unit and the j hidden layer unit. h^k represents the k sub-hidden layer, and h^k_{ij} represents the value of the i visible layer and the value of the j unit in the k sub-hidden layer. W^k represents the convolution kernel of the k hidden unit, and b_k is its bias. Figure 4 shows the CRBM model. The following energy function of CRBM can be defined: $E(v, h) = -\sum_{k=1}^{K} h^k(W^k * v) - \sum_{k=1}^{K} b_k \sum_{i,j} h_ij^k - c \sum_{i,j} v_{ij}$. Where '*' stands for the convolution operation.

CRBM uses conditional probability distribution for sampling, and the function of the conditional probability can be defined as:

$$P(H^k_{ij} = 1 \,|V, \theta) = \sigma((w'^k \cdot v)_{ij} + b_k) \tag{1}$$

$$P(h^k_{ij} = 1 \,|H, \theta) = \sigma((w^k * h^k)_{ij} + \alpha) \tag{2}$$

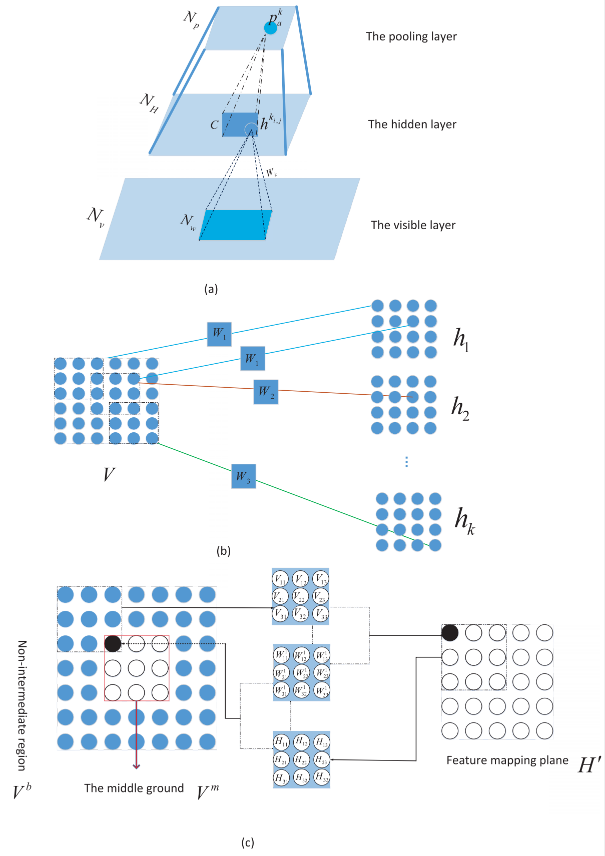

N_p
p_a^k
The pooling layer
N_H
C
$h^{k_{i,j}}$
The hidden layer
W_k
N_v
N_w
The visible layer
(a)
W_1
W_1
W_2
W_3
h_1
h_2
h_k
V
(b)
V_{11} V_{12} V_{13} V_{21} V_{22} V_{23} V_{31} V_{32} V_{33}
W_{11}^1 W_{12}^1 W_{13}^1 W_{21}^1 W_{22}^1 W_{23}^1 W_{31}^1 W_{32}^1 W_{33}^1
H_{11} H_{12} H_{13} H_{21} H_{22} H_{23} H_{31} H_{32} H_{33}
Non-intermediate region
V^b
The middle ground
V^m
Feature mapping plane
H'
(c)

FIGURE 4. CRBM model. (a) The general CRBM structure; (b) The mathematical process from the visible layer to the hidden layer; (c) The convolution operation in the visible layer of CRBM model.

Among them: $\sigma(x) = \frac{1}{1+\exp(-x)}$

Each group of pooling layers has $N_P \times N_P$ binary units. For each dimension, the hidden layer is divided into blocks of size $C \times C$ ($N_p = N_h/C$), where C is the size of the sampling block,and its value is generally 2 or 3. Each block a is connected to one binary unit p_a^k in the pooling layer The maximum probability pool is a layer that used to dimension reduce and regularized operation in this paper. In conclusion, the conditional probability of CRBM model with maximum probability pool can be defined as follows:

$$P(H_{ij}^k = 1\,|V,\theta) = \frac{\exp(I(H_{i,j}^k))}{\sum_{(i',j')\in B_a} \exp(I(H_{i',j'}^k))} \tag{3}$$

$$P(p_a^k = 0\,|V,\theta) = \frac{1}{1 + \sum_{((i',j'))\in B_a} \exp(I(H_{(i',j')}^k))} \tag{4}$$

where, $I(H_{i,j}^k) \triangleq b_k + (w'^k * v)_{ij}$

Contrastive Divergence (CD) algorithm [22], which was proposed by Hinton, is used to train the CRBM model. Firstly, the output of the visible layer is obtained through K convolution kernels of size $N_W \times N_W$, and then the activation probability of neurons in the hidden layer is calculated by the methods of equations (1) and (3). Next, select the point with the highest probability in a 3 × 3 region to form the pooling layer. Finally, equation (2) is used to reconstruct the visible layer, and calculate the reconstruction error and gradient value for updating the model parameters. Repeat this process until the model is basically stable.

A. THE IMPROVED CRBM MODEL

As can be seen from figure 4, only the posterior activation probability of the middle non-shaded region can be obtained from the equation (2). Aiming at the reconstruction of the visible elements in the CRBM model, an improved method has been proposed to carry out zero filling operation on the edge in order to include the edge region into the middle region and make the results more accurate. Suppose that v^b is the original edge area, which is changed as a new intermediate region. The new edge area is $v^{b'}$. The maximum likelihood probability is changed into the maximum likelihood probability of the intermediate region. The training objective of the improved model is as follows:

$$F_{t\arg et} = \max(\sum\nolimits_V \log p(V^{m'} \left| V^{b'} \right.) - P_E)$$

$V^{m'}$ refers to reconstruction values in the middle region of the new image, and the size is the original image size, which avoids the defect that the non-intermediate region of the image cannot be directly calculated by formula (2). $v^{b'}$ stands for the non-intermediate part of the new image, which is made up 0. $V^{m'} \left| V^{b'} \right.$ refers to $v^{b'}$ remains the same. $-P_E$ stands for minimizing the cross entropy sparse penalty factor.

V. THE NEW HYBRID TRANSFER LEARNING METHOD BASED ON A CNN MODEL AND THE IMPROVED CRBM MODEL

Transfer learning refers to transferring the trained CNN model to other similar data sets, and relearns the important features of the target set. Although the deep convolutional neural network model based on transfer learning can be directly used to extract features of a small target data set by fine-tuning the pre-trained model with very little training time, the recognition ability of the extracted features is affected in the transfer learning process due to the content differences between data sets. CRBM has a strong unsupervised characteristic learning ability. By learning complex rules of the input data, the statistical features with a high recognition ability of images are reconstructed.

In order to solve the negative influence brought by the traditional transfer learning method, a hybrid migration model based on CNN and CRBM is proposed in this paper in view of the advantages of CRBM. In order to solve the influence of negative migration in the process of deep transfer learning, in view of the advantages of CRBM, a hybrid transfer model based on CNN and CRBM is proposed in this paper. When the CNN model is transferred to the target set, the CRBM model is fused with the original pre-trained CNN model. The structure of the new hybrid deep transfer learning model is shown in figure 5.

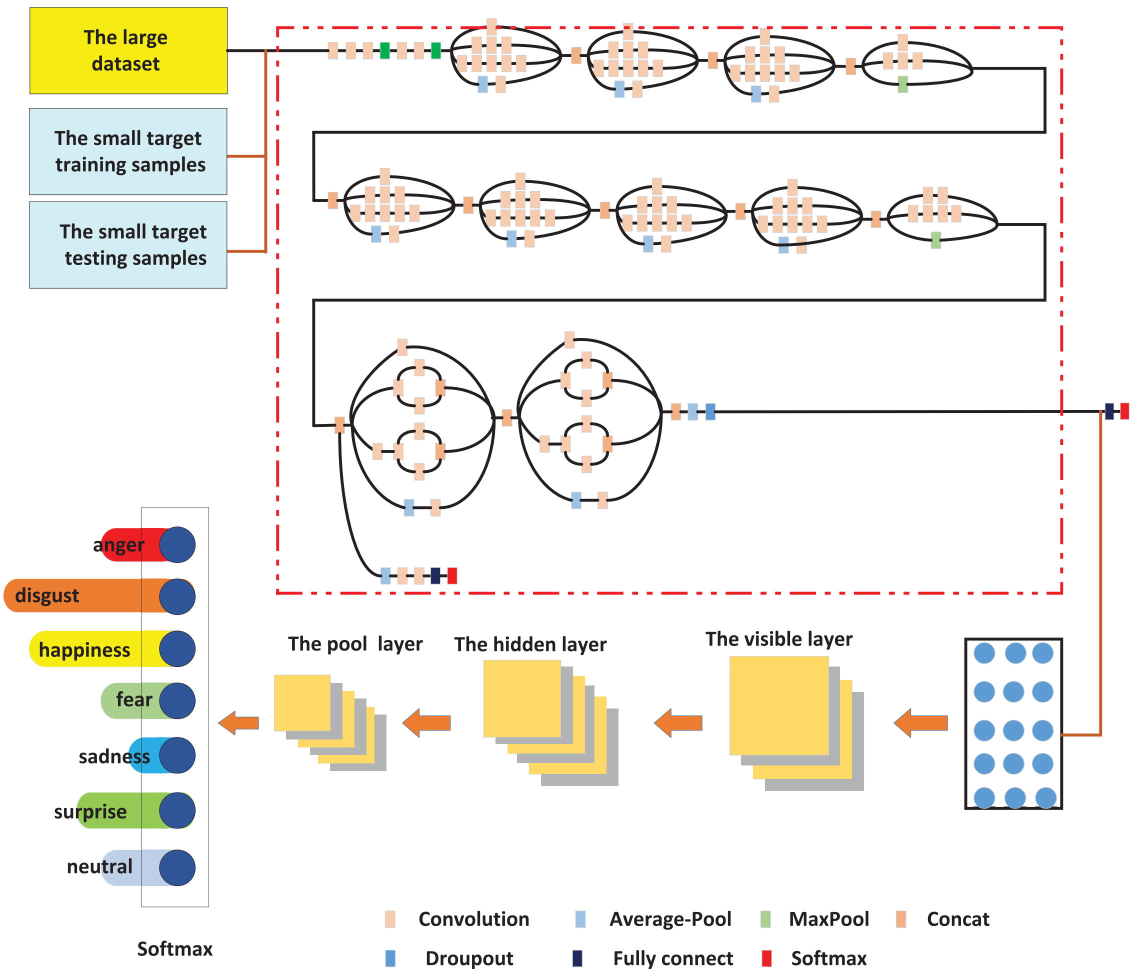


FIGURE 5. The new hybrid transfer learning model. The part of the red dotted line is the DCNN model that the full connection layer and the softmax layer have been removed. The output of the module in the region of the red dotted line is directly connected to the improved CRBM model in the first transfer learning stage for acquiring more favorable features for target classification.

A. THE PRE-TRAINING OF CONVOLUTIONAL NEURAL NETWORK

The training of the convolutional neural network model includes two steps: forward propagation and back parameter adjustment. In the training process, x^{l-1} is the input feature vector of the current layer. x^l is the output feature vector. In this layer. The weight and the bias of the convolution filter are w^l and b^l respectively, and then the input feature of each layer are as follows:

$$\begin{aligned} x^l &= f(u^l) \\ u^l &= w^l x^{l-1} + b^l \end{aligned} \tag{5}$$

Generally, $f(\cdot)$ is sigmoid function. Suppose m is the label number of the sample set.

$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ stands for the whole sample set, where $(x^{(i)}, y^{(i)})$ represents n dimensional input vector of one label. The overall cost function of the convolutional neural network model is:

$$J(w, b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}\left(x^{(i)} - y^{(i)}\right\|\right)^2\right) + \frac{\lambda}{2}\sum_{l=1}^{n_{l-1}}\sum_{i=1}^{s_l}\sum_{j=1}^{n_{l+1}}\left(w_{ji}^{(l)}\right)^2$$

where, λ stands for weight attenuation parameter; n_l is the total network layers; s_l represents the number of nodes in the network layer l. The specific model parameters are adjusted to realize gradient descent. Parameters for each layer are updated as follows:

$$\begin{aligned} w_{ij}^{(l)} &= w_{ij}^{(l)} - \alpha \frac{\partial}{\partial w_{ji}^{(l)}} J(w, b) \\ b_i^{(l)} &= b_i^{(l)} - \alpha \frac{\partial}{\partial b_{ji}^{(l)}} J(w, b) \end{aligned} \tag{6}$$

In equation (6), α refers to the learning rate. The partial derivative of the cost function is analyzed by calculating the residuals of each layer. When the cost error reaches the minimum, the convolution neural network model can be obtained and the training quality of the model can be guaranteed.

Google has trained an Inception-v3 model [23] on the large image database (ImageNet) [24] that can be used directly for image classification. Inception-v3 model has approximately 25 million parameters, and it takes 5 billion multiply and add instructions to classify one image. On a modern personal computer without a GPU, this model can classify

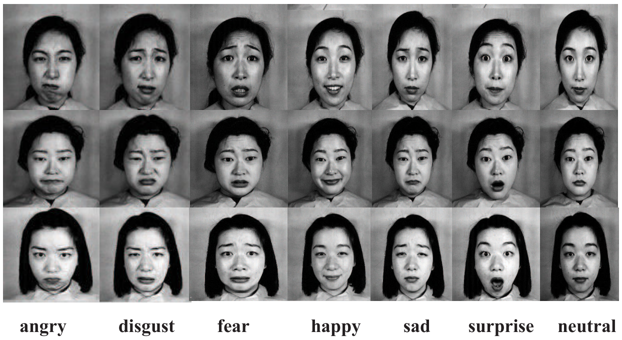


FIGURE 6. Examples of images in the JAFFE database. The emotions from left to right are: Anger, Disgust, Fear, Happy, Sad, Surprise, Neutral.

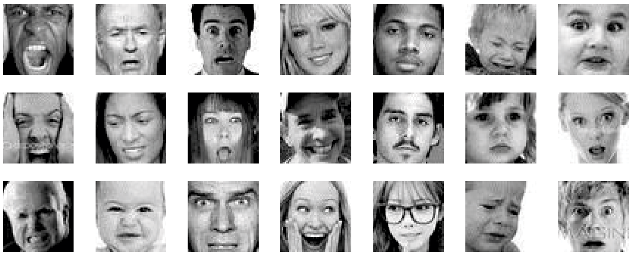


FIGURE 7. Examples of images in the FER2013 database. The emotions from left to right are: Anger, Disgust, Fear, Happy, Normal, Sadness, Surprise.

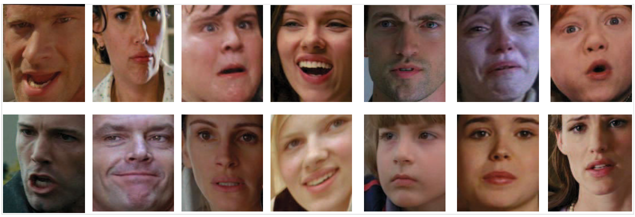


FIGURE 8. Examples of images in the SFEW database. The emotions from left to right are: Anger, Disgust, Fear, Happy, Neutral, Sadness, Surprise.

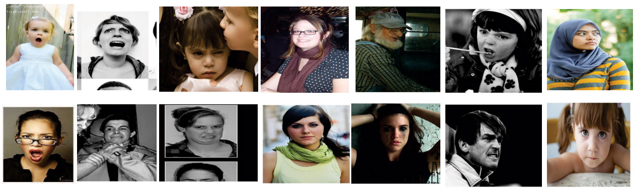


FIGURE 9. Examples of images in the RAF-DB database. The emotions from left to right are: 1, 2, 3, 4, 5, 6, 7.

an image quickly. The ImageNet dataset contains 15 million images, which belong to 22,000 categories. Its subset corresponds to the current most authoritative image classification competition, Large Scale Visual Recognition Challenge (LSVRC) [25], which contains 1 million images and 1000 categories. Several weeks may be spend on training the inception-v3 model by a normal personal computer (PC), hence it is not possible to train the deep model by using a normal PC. This paper uses a pre-trained Inception-v3 model for image classification. The pre-trained inception model will be downloaded and directly used to classify facial expression images.

B. THE NEW TRANSFER LEARNING ALGORITHM

The pre-trained convolutional neural network model has been transferred to the small target set, and the hybrid neural network model will be trained as follows. In the training process of the new hybrid model, CD algorithm is firstly used to train the improved convolution restricted boltzmann machine model to clarify the higher-order statistical characteristics of the input data. In the process of training the classifier, the cross entropy is used as a loss function to update the network parameters for getting accurate classification results. With the help of back propagation (BP) algorithm [26], the parameters of each layer can be effectively monitored and adjusted. In the process of adjusting parameters, the likelihood function and BP algorithm have been respectively used to carry out the process of the back propagation, which can ensure that the statistical feature extracted from CRBM remains unchanged. For the whole training sample, the loss function is defined as:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\log(_{j}^{(i)})$$

The detailed classification steps of the fused deep transfer learning model are as follows:

Step 1: The CNN model (Inception-V3) is pre-trained on the big data set (Imagenet);

Step 2: Transfer the CNN model to the training samples in a small target set;

Step 3: Use the image feature maps extracted by the feature extraction layer that in front of the full connection layer to concatenate all feature maps of each image into a new feature map.

Step 4: Remove the full connection layer and the softmax layer of the original pre-training CNN model, and use the CRBM model to fully connect the input feature map.

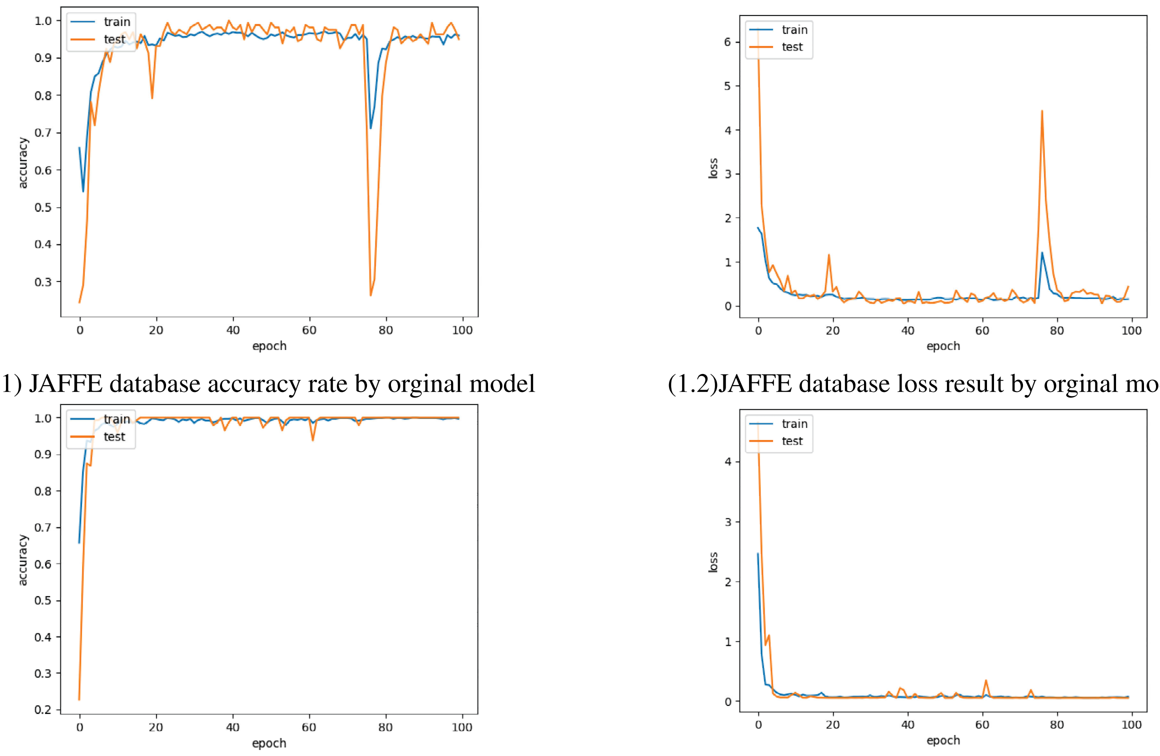
Step 5: The training samples of the target set are taken as the input information, and the parameters of each CRBM layer will be learned without supervision successively. The CD algorithm is used to fine-tune the parameters of softmax regression and each CRBM layer to obtain a better trained deep transfer learning classifier.

Step 6: The testing samples in the target set will be used as the input information, and the final image classification result can be get from the new trained model.

VI. EXPERIMENTS

A. DATABASE

JAFFE database was published in 1998 [27], and it is a relatively small database. This database includes 213 images that produced by 10 Japanese women, and each person has seven emotions: disgust, anger, fear, happy, sad, surprise and neutral. Figure 6 shows parts samples of JAFFE.



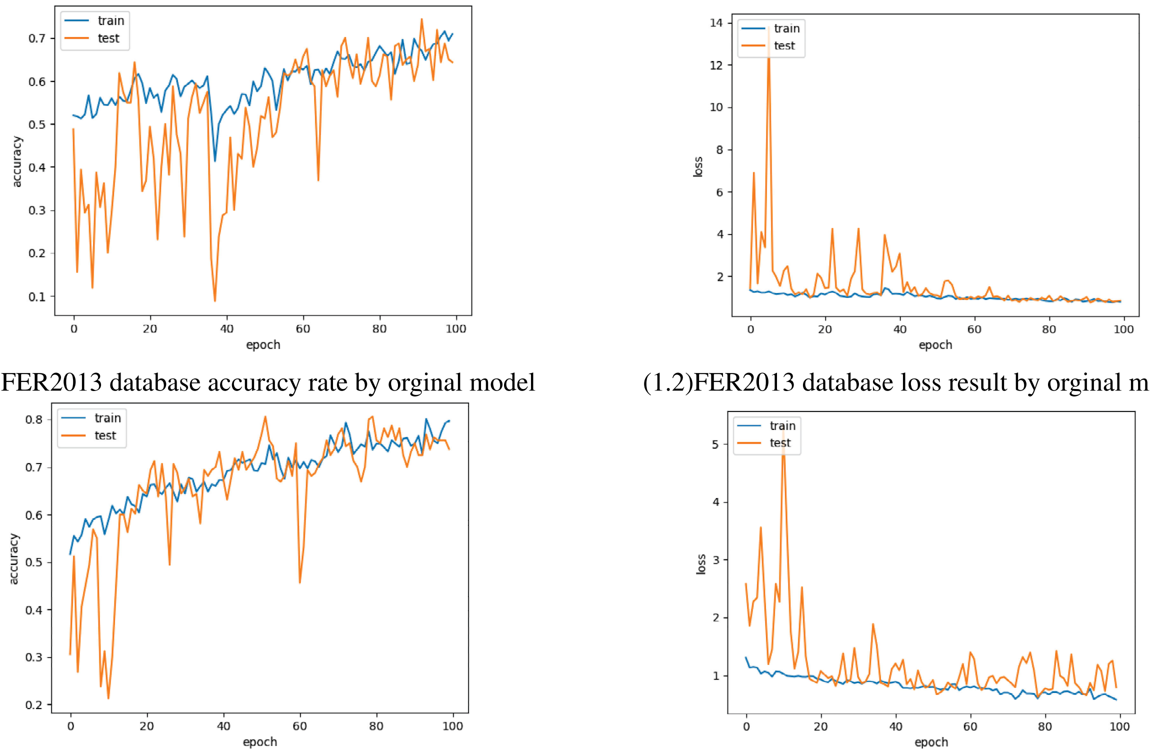
(1.1) JAFFE database accuracy rate by orginal model

(1.2)JAFFE database loss result by orginal model

(1.3)JAFFE database accuracy rate by new model

(1.4)JAFFE database loss result by new model

FIGURE 10. Comparison between the original model and the new hybrid model based on JAFFE database.



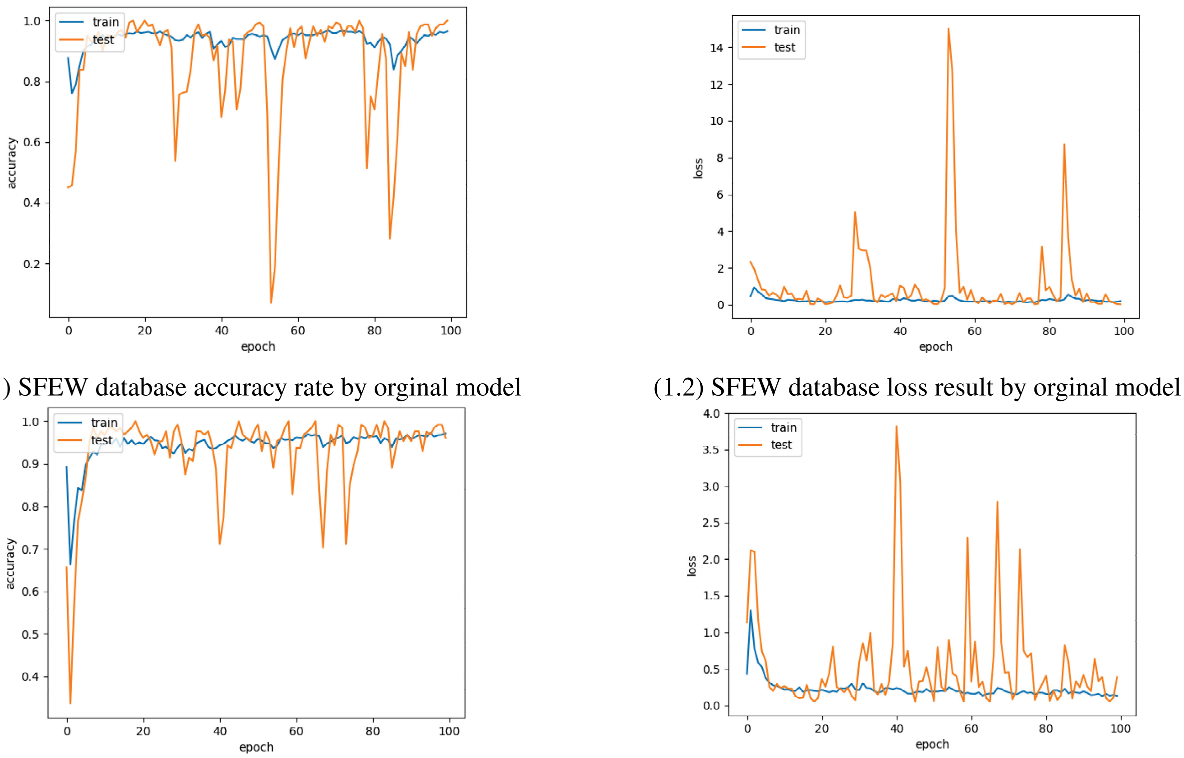
(1.1)FER2013 database accuracy rate by orginal model

(1.2)FER2013 database loss result by orginal model

(1.3)FER2013 database accuracy rate by new model

(1.4)FER2013 database loss result by new model

FIGURE 11. Comparison between the original model and the new hybrid model based on FER2013 database.



(1.1) SFEW database accuracy rate by orginal model

(1.2) SFEW database loss result by orginal model

(1.3) SFEW database accuracy rate by new model

(1.4) SFEW database loss result by new model

FIGURE 12. Comparison between the original model and the new hybrid model based on SFEW database.

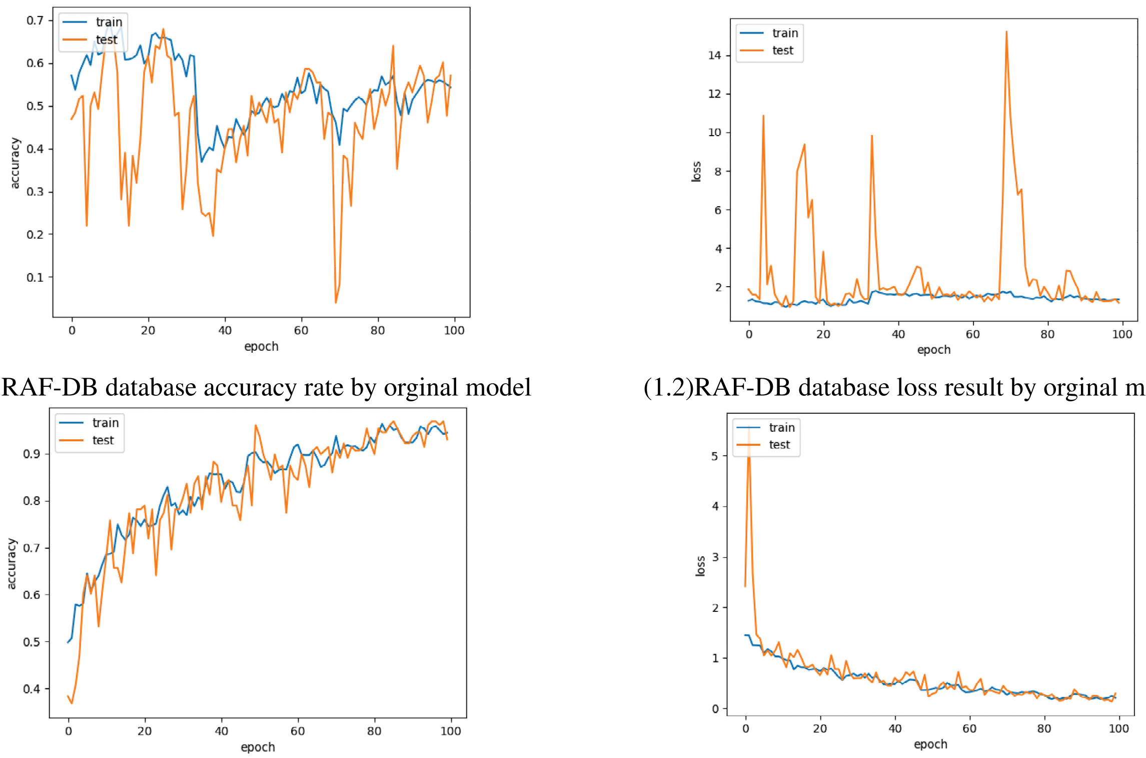
TABLE 1. The number of four experimental dataset: JAFFE, CK+, FER2013, SFEW and RAF-DB. where, AN=anger, DI=digust, FE=fear, HA=happy, SA=sad, SU=surprise, NE=neutral, NO=normal.

JAFFE	number	FER2013	number	SFEW	number	RAF	number
AN	30	AN	4953	AN	255	1	1619
DI	29	DI	547	DI	89	2	355
FE	32	FE	5121	FE	145	3	877
HA	31	HA	8989	HA	271	4	5957
NE	30	NO	6198	NE	236	5	2460
SA	31	SA	6077	SA	245	6	867
SU	30	SU	4002	SU	153	7	3204

The Facial Expression Recognition 2013 (FER-2013) database [12] includes 35,887 different images. The training set consists of 28,709 examples. The public test set used for the leaderboard consists of 3,589 examples. The private test set consists of another 3,589 examples. The data consists of 48x48 pixel grayscale images of faces. There are seven expressions are labeled in this database: normal, happy, sadness, surprise, anger, disgust, and fear. Some examples of the fer2013 database images are shown in figure 7.

SFEW database [28] is a part of a temporal facial expressions database acted facial expressions in the wild which we have extracted from movies. This database closes to the real world illumination. There are 958 images in the training set and 436 images in the validation set. Some examples of the sfew database images are shown in figure 8.

Real-world Affective Faces Database (RAF-DB) [29] is a large-scale facial expression database with around 30K great-diverse facial images downloaded from the Internet. Based on the crowdsourcing annotation, each image has been independently labeled by about 40 annotators. Images in this database are of great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g. glasses, facial hair or self-occlusion), post-processing operations (e.g. various filters and special effects), etc. RAF-DB has large diversities, large quantities, and rich annotations, including: 29672 number of real-world images, a 7-dimensional



(1.1)RAF-DB database accuracy rate by orginal model

(1.2)RAF-DB database loss result by orginal model

(1.3)RAF-DB database accuracy rate by new model

(1.4)RAF-DB database loss result by new model

FIGURE 13. Comparison between the original model and the new hybrid model based on RAF-DB database.

expression distribution vector for each image. Some examples of the RAF-DB database images are shown in figure 9.

B. RESULTS

In order to verify the superiority of the algorithm that proposed in this paper, it is necessary to verify the hybrid transfer learning method through the test of the classification accuracy and the analogy of some similar methods. The average accuracy was used to evaluate and compare the classification of facial expression images, and the detailed information of these experimental databases can be seen in figures 6, 7, 8, 9. The detailed numbers of the four experimental datasets are listed in table 1.

Figures 10, 11, 12, and 13 show the results of two different deep transfer learning models, and the differences between the new model and the traditional model can be obviously found in these figures.

From figures 10, 11, 12 and 13, the proposed model performs well in these four public datasets: JAFFE, FER2013, SFEW and RAF-DB, with accuracy of 99.2%, 73.75%, 96.09% and 92.97% respectively. Therefore, in the process of deep transfer learning, the output feature maps of the traditional deep CNN model can be connected with the CRBM model, and the new hybrid transfer learning method shows a better application effect than directly using the full connection layer. The classification accuracy is obviously improved by the new method.

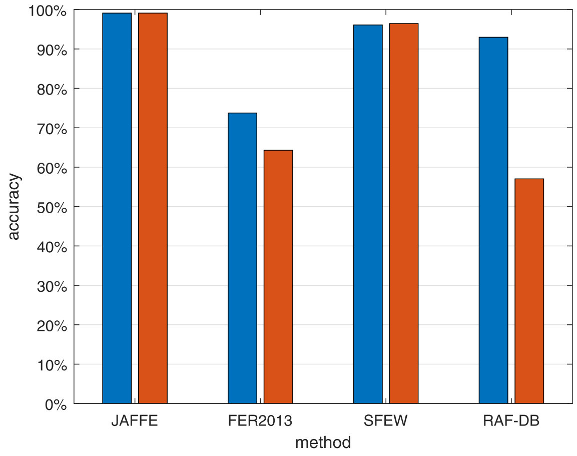


FIGURE 14. The comparison of the traditional transfer learning method (the right histogram) and the new hybrid transfer learning method (the left histogram) on four experimental databsets.

Table 2 shows the result comparison of the current state of the art in facial expression recognition on four databases introduced in this paper. Figure 14 shows the comparison between the traditional transfer learning method without CRBM mechanism and the new hybrid transfer learning method.

Experimental results show that the new method that proposed in this paper is still better than the current state-of-the-art in emotion recognition on these datasets: JAFFE, FER2013, SFEW and RAF-DB.

TABLE 2. This Table summarizes the current state of the art in facial expression recognition on the four databases: JAFFE, FER2013, SFEW and RAF-DB.

JAFFE dataset Author	Method	Acuuuracy (%)
Chen[29]	ECNN	94.3
Wen[13]	Probability-Based	50.7
AI Abdullah[30]	FLDA+KNN	95.09
Liu L[31]	KECA and SSVM	93.04
Minaee[32]	Attention CNN	92.8
Wang[33]	Fa-Net	95.7
This paper	new method	99.2
FER2013 dataset Author	Method	Acuuuracy (%)
Chen[29]	ECNN	69.96
Minaee[32]	Attention CNN	70.6
Wang[33]	Fa-Net	71.1
Wang[34]	VGG+SVM	66.3
M Shin[35]	Hist+CNN	66.67
Amani Alfakih[36]	Multi-view DCNN	72.27
This paper	new method	73.75
SFEW dataset Author	Method	Acuuuracy (%)
Li[37]	attention mechanism	53
Liu[38]	Adaptive Deep Metric	54
Liu[39]	CNN+soft label	55.73
Ji Y[40]	ICID fusion network	51.2
Mitre-Ortiz A[41]	Galvanic Skin Response	61.6
Tong X[43]	DASOP	59.518
Acharya[44]	covariance pooling	58
This paper	new method	96.09
RAF-DB dataset Author	Method	Acuuuracy (%)
Liu[39]	CNN+soft label	86.31
Ji Y[40]	ICID fusion network	75.4
S Li[42]	DLP-CNN+msvm	74.2
Tong X[43]	DASOP	88.625
Acharya[44]	covariance pooling	87
This paper	new method	92.97

VII. CONCLUSION

The application principle of convolutional neural network is to extract the image features layer by layer to obtain structural features, which can be used to express the high-level semantics of a single image. The higher the level is, the more abstract the model features will be, which greatly improves the recognition ability in the classification process. In the training process of the convolutional neural network model, there are millions of parameters involved, which requires a large number of labeled samples. Generally, feature that extracted from the pre-training model is often directly applied on a small target sample set. But there are content differences between the original data set and the target data set based on the traditional transfer learning method of convolutional neural network, which will affect the ability of feature extraction and recognition. Considering the advantages of CRBM model, this paper proposed a new hybrid transfer learning method based on the advantages of these two models to reduce content differences between the two different data sets before using the transfer learning algorithm, and proposed an improved method based on the visible unit reconfiguration problem of the original CRBM model to acquire more important feature. When the target set is taken as the input of the deep transfer CNN model, the improved CRBM is used to replace the full connection layer of the traditional convolutional neural network model. The input of the improved CRBM model is get by combining all kinds of feature maps, which will be existed in the form of overall structural features. On the basis of the existing theories, the advantages of the above model are given full play. Through re-training the CRBM model, the special higher-order statistical characteristics of the target data set are extracted by effective methods to ensure more accurate image classification. The content difference of data sets will not affect the feature recognition ability. A number of experiments have shown the effectiveness and feasibility of the new hybrid transfer learning method.

REFERENCES

[1] A. Mehrabian, "Communication without words," *Psychol. Today*, vol. 2, no. 4, 1968.

[2] H. Sadeghi and A. A. Raie, "Suitable models for face geometry normalization in facial expression recognition," *J. Electron. Imag.*, vol. 24, no. 1, 2015, Art. no. 013005.

[3] L. Gui, T. Baltrušaitis, and L. P. Morency, "Curriculum learning for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture*, May 2017, pp. 505–511.

[4] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," *Image Vis. Comput.*, vol. 24, no. 3, pp. 239–248, 2006.

[5] A. Majumder, L. Behera, and V. K. Subramanian, "Facial expression recognition with regional features using local binary patterns," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2013, pp. 556–563.

[6] L. Shu and A. C. S. Chung, "Face recognition with salient local gradient orientation binary patterns," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2010, pp. 3317–3320.

[7] S. K. Wang, L. Liu, and X. Xu, "Vehicle logo recognition based on local feature descriptor," *Appl. Mech. Mater.*, vol. 263, pp. 2418–2421, Dec. 2013.

[8] S. A. Korkmaz, A. Akçiçek, H. Bínol, and M. F. Korkmaz, "Recognition of the stomach cancer images with probabilistic HOG feature vector histograms by using HOG features," in *Proc. IEEE 15th Int. Symp. Intell. Syst. Inform. (SISY)*, Sep. 2017, pp. 339–342.

[9] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] D. Li, G. Wang, Q. Jin, and T. Song, "Improving convolutional neural network using accelerated proximal gradient method for epilepsy diagnosis," in *Proc. UKACC Int. Conf. Control*, 2016, pp. 1–6.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[13] G. HWen, H. Zhi, and H. Li, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cognit. Comput.*, vol. 9, no. 4, pp. 1–14, 2017.

[14] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 435–442.

[15] H. Jung, S. Lee, and J. Yim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2983–2991.

[16] J. Hong, J. Yin, and Y. Huang, "TrSVM: A transfer learning algorithm using domain similarity," *J. Comput. Res. Develop.*, vol. 48, no. 10, pp. 1823–1830, 2011.

[17] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," 2017, *arXiv:1703.09179*. [Online]. Available: https://arxiv.org/abs/1703.09179

[18] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *J. Digit. Imag.*, vol. 30, no. 2, pp. 234–243, 2017.

[19] H.-J. Lei, T. Han, and F. Zhou, Z. Yu, J. Qin, A. Elazab, "A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning," *Pattern Recognit.*, vol. 79, pp. 290–302, Jul. 2018.

[20] Z. J. Zhao and J. W. Gu, "Recognition of digital modulation signals based on hybrid three-order restricted Boltzmann machine," in *Proc. IEEE Int. Conf. Commun. Technol.*, Oct. 2016, pp. 166–169.

[21] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. ACM 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 609–616.

[22] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

[23] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *Proc. Int. Conf. Image*, 2017, pp. 783–787.

[24] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?" 2016, *arXiv:1608.08614*. [Online]. Available: https://arxiv.org/abs/1608.08614

[25] J. Dong, X. Li, and S. Liao, "Image retrieval by cross-media relevance fusion," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 173–176.

[26] J. Li, J. Cheng, F. Huang, and J. Shi, "Brief introduction of back propagation (BP) neural network algorithm and its improvement," in *Advances in Computer Science and Information Engineering*. Berlin, Germany: Springer, 2012, pp. 553–558.

[27] M. J. Lyons, S. Akamatsu, J. Gyoba, J. Budynek, and M. Kamachi, "The Japanese female facial expression (JAFFE) database," in *Proc. 3rd Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 14–16.

[28] A. Dhall, R. Goecke, and S. Lucey, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.

[29] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, Jul. 2016.

[30] A. I. Abdullah, "Facial expression identification system using Fisher linear discriminant analysis and k-nearest neighbor methods," *ZANCO J. Pure Appl. Sci.*, vol. 31, no. 2, pp. 9–13, 2019.

[31] L. Liu, L. Yang, X. Zhang, L. Hu, F. Deng, and Y. Chen, "Facial expression recognition based on SSVM algorithm and multi-source texture feature fusion using KECA," in *Recent Developments in Intelligent Computing, Communication and Devices*. Singapore: Springer, 2019, pp. 659–666.

[32] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," 2019, *arXiv:1902.01019*. [Online]. Available: https://arxiv.org/abs/1902.01019

[33] W. Wang, Q. Sun, and T. Chen, "A fine-grained facial expression database for end-to-end multi-pose facial expression recognition," 2019, *arXiv:1907.10838*. [Online]. Available: https://arxiv.org/abs/1907.10838

[34] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019.

[35] M. Shin, M. Kim, and D. S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *Proc. 25th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2016, pp. 724–729.

[36] A. Alfakih, S. Yang, and T. Hu, "Multi-view cooperative deep convolutional network for facial recognition with small samples learning," in *Proc. Int. Symp. Distrib. Comput. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 207–216.

[37] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[38] X. Liu, B. V. K. Vijaya Kumar, and J. You, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 20–29.

[39] Y. Gan, J. Chen, and L. Xu, "Facial expression recognition boosted by soft label with a diverse ensemble," *Pattern Recognit. Lett.*, vol. 125, pp. 105–112, Jul. 2019.

[40] Y. Ji, Y. Hu, Y. Yang, F. Shen, and H. T. Shen, "Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network," *Neurocomputing*, vol. 333, pp. 231–239, Mar. 2019.

[41] A. Mitre-Ortiz and H. Mitre-Hernandez, "Study of spontaneous and acted learn-related emotions through facial expressions and galvanic skin response," *Res. Comput. Sci.*, vol. 148, pp. 97–105, 2019.

[42] S. Li, W. Deng, and J. P. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2852–2861.

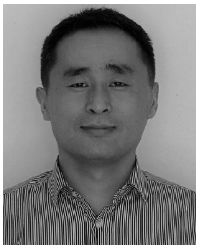
[43] X. Tong, S. Sun, and M. P. Fu, "Data augmentation and second-order pooling for facial expression recognition," *IEEE Access*, vol. 7, pp. 86821–86828, 2019.

[44] D. Acharya, Z. Huang, and D. P. Paudel, "Covariance pooling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 367–374.

YINGYING WANG was born in 1989. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent control with the School of Control Science and Engineering, Shandong University, Jinan, Shandong, China, in 2016.

YIBIN LI received the B.Sc., M.Sc., and Ph.D. degrees from Shandong University, China, in 2002, 2005, and 2012, respectively. His research interest includes algorithms for neural networks and gait planning of legged robots and so on.

YONG SONG received the Ph.D. degree in pattern recognition and intelligent system from Shandong University, Jinan, Shandong, in 2012. He is currently a Professor with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai. His current research interests include mobile robot navigation, machine learning, neural networks, control of intelligent robots, and swarm intelligence robotics.



XUEWEN RONG received the bachelor's and master's degrees from the Shandong University of Science and Technology, China, in 1996 and 1999, respectively. He is currently pursuing the Ph.D. degree with the School of Control Science and Engineering, Shandong University, China. He is also a Senior Engineer with the School of Control Science and Engineering, Shandong University. His research interests include robotics, mechatronics, hydraulic servo driving technology, and so on.