The Visual Computer (2020) 36:1635–1649
https://doi.org/10.1007/s00371-019-01759-7

ORIGINAL ARTICLE

Improved curriculum learning using SSM for facial expression recognition

Xiaoqian Liu[1] · Fengyu Zhou[1]

Published online: 9 October 2019

Abstract
Facial expression recognition is an important research issue in the pattern recognition field. However, the generalization of the model still remains a challenging task. In this paper, we apply a strategy of curriculum learning to facial expression recognition during the stage of training. And a novel curriculum design method is proposed. The system first employs the unsupervised density–distance clustering method to determine the clustering center of each category. Then, the dataset is divided into three subsets of various complexity according to the distance from each sample to the clustering center in the feature space. Importantly, we develop a multistage training process where a main model is trained by continuously adding harder samples to training set to increase the complexity. To solve the problem that the model has a poor recognition accuracy for anger, fear and sadness, a self-selection mechanism is introduced in the test stage to make further judgment on the result of the main model. Experiment results indicate that the proposed model can achieve a satisfactory recognition accuracy of 72.11% on FER-2013 and 98.18% on CK+ dataset for 7-class facial expressions, which outperforms the other state-of-the-art methods.

Keywords Curriculum learning · Density–distance clustering · Facial expression · Recognition

1 Introduction

Over the past few years, there has been an increasing interest in machine understanding and recognition of affective and cognitive mental states, especially based on facial expression analysis [1]. Facial expression is one of the most powerful, natural, and universal signals for human beings to convey their emotional states and intentions [2], which is a method of identifying human emotion. The task of facial expression recognition (FER) is an image classification of six basic expressions (anger, disgust, fear, happiness, sadness, surprise) [3] and neutral. Numerous studies have been conducted on automatic facial expression analysis [4–8] due to its wide application in human–machine interaction (HMI), psychological analysis, and emotion-based recommendation system [9]. In fact, FER has been regarded as one of the fundamental technologies for human–machine interaction, where human can communicate with machine just like human.

✉ Fengyu Zhou
zhoufengyu@sdu.edu.cn

[1] School of Control Science and Engineering, Shandong University, Jinan 250061, China

Facial expression recognition can be divided into two main categories according to the feature representations: static image FER and dynamic sequence FER. The static-based methods [7,10,11] only extract spatial information as feature representation from the current single images, whereas the dynamic-based methods [12–14] take the temporal relation among contiguous frames in the input facial expression sequence. The majority of the traditional methods have used hand-crafted features, such as pixel intensity [15], local binary pattern (LBP) [10,16], histogram of oriented gradients (HOG) [17], and Gabor-wavelet. Nevertheless, owing to emotion recognition competitions, such as FER-2013 [18] and Emotion Recognition in the Wild (EmotiW) [19,20], some relatively sufficient training data are collected from challenging real-world scenarios, which implicitly promotes the transition from lab-controlled to in-the-wild settings. In the meanwhile, benefited by the GPU units and well-designed network architecture, studies have begun to transfer to deep learning method. Recently, deep learning technology has made great progress in many areas of computer vision, such as object detection, image segmentation and image classification. Deep neural networks can automatically extract advanced semantic features and effective facial rep-

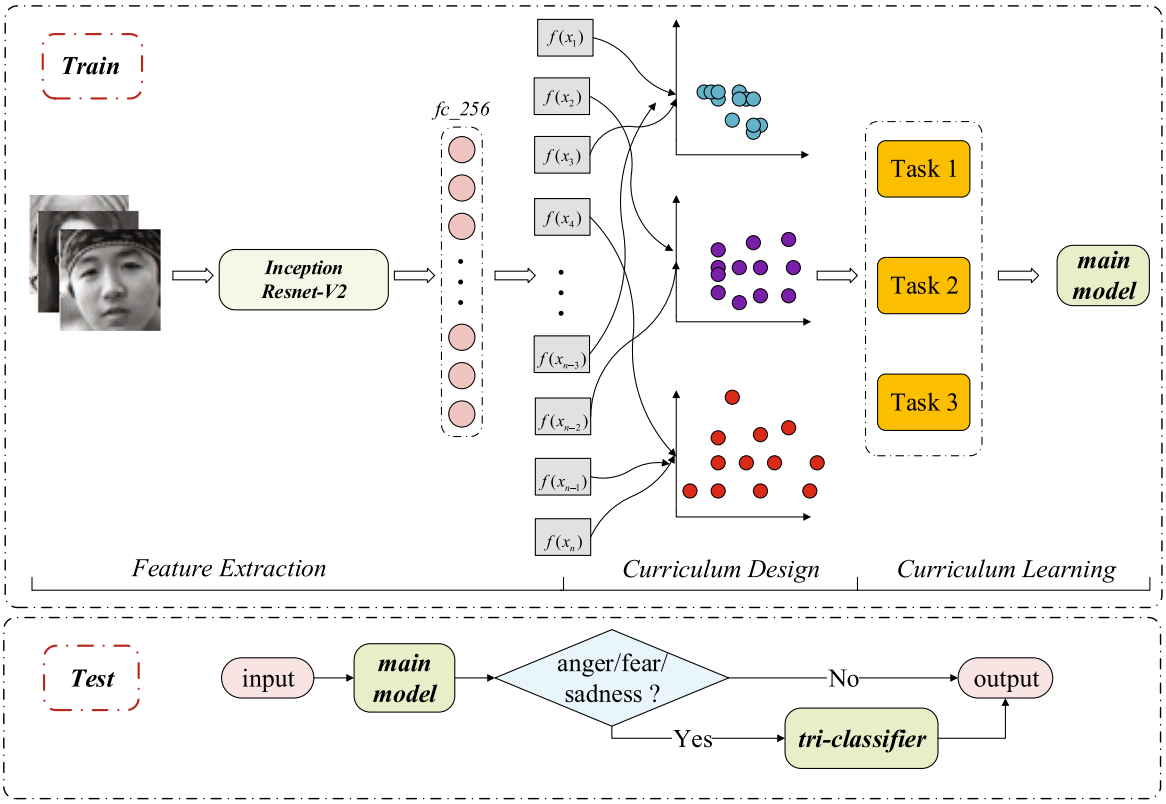


Fig. 1 Pipeline of the proposed networks. The training phase mainly consists of three parts: feature extraction, curriculum design, and curriculum learning. In the test stage, we introduced a self-selection mechanism to further judge the expressions of anger, fear and sadness. The directly output channel is called the main channel, and the other channel is called the second channel

resentations from raw images, which are important for FER. However, because of the subtlety of facial expression, there are still a lot of challenges in facial expression recognition, such as generalization and the poor ability to recognize the expressions of fear, anger and sadness. Owing to sharing similar appearance and movement of facial muscles, these three expressions are easy to be confused with each other.

In this paper, we apply a novel training strategy to facial expression recognition which can lead to better generalization performance and propose a self-selection mechanism (SSM) in the test phase. Curriculum learning allows the model to learn simple tasks early so they can be used as building blocks to learn more complex ones. In our work, a curriculum is designed by measuring the complexity of data according to the distance between each sample and the clustering center which is determined by the unsupervised density–distance clustering method. Then, the curriculum learning training strategy is used to optimize the main model. In addition, in the test phase, a self-selection mechanism is introduced to make further judgment on anger, fear and sadness expressions. Extensive experiments are conducted on the well-known FER-2013 and CK+ datasets. Figure 1 illustrates the system architecture.

The rest of the paper is organized as follows. Section 2 briefly overviews the literature about curriculum learning and facial expression recognition. Our proposed method is detailed in Sect. 3. The experimental results on FER-2013 dataset and discussion are presented in Sect. 4. Some conclusions and future works are drawn in Sect. 5.

2 Related works

In this section, we introduce the progress of the curriculum learning and elaborate the research process of facial expression recognition.

2.1 Curriculum learning

Bengio et al. [21] first proposed the concept of curriculum learning. He demonstrated that it can result in better generalization and faster learning on a synthetic vision and word representation learning tasks. Pentina et al. [22] applied curriculum learning to a multi-task learning and proposed a model to learn the order of multiple tasks. Their experimental results showed that learning tasks sequentially is better

than learning them jointly. Vanya et al. [23] investigated the effect of curriculum learning in image classification by training a CNN from scratch. Huang et al. [24] presented a CurriculumNet to handle massive amount of noisy labels and data imbalance effectively. He found that those images with highly noisy labels could improve the generalization capability of model.

2.2 Facial expression recognition

The facial expression recognition focuses on extracting features from raw images which is an important step and then recognizing different facial expressions with a trained classifier. Features are mainly divided into appearance-based and geometric-based features. Traditional facial expression recognition bases on hand-crafted features. Bartlett et al. [25] showed that Gabor-wavelet features derived from 48×48 face images have the high dimensionality of which computation is costly. Shan et al. [10] found that LBP features perform robustly and stably over a range of low-resolution images which are widely used to extract texture features. However, features based on appearance are extracted from the entire facial region and local regions that are highly related to expression changes, such as the nose, eyes, and mouse, are ignored [26]. Facial Action Coding System (FACS) [27] which is based on geometric features defined the basic deformation Action Units (AUs) according to the facial muscle type and movement characteristics and facial expressions can finally be decomposed and corresponded to each AU. Liu et al. [28] proposed AU-inspired deep networks (AUDNs) inspired by the psychological theory that expressions can be decomposed into multiple facial AUs.

Existing facial expression recognition based on hand-crafted features has limited ability of extracting advanced semantic features. Deep convolutional neutral networks have recently obtained state-of-the-art performance for FER tasks which extract features end-to-end. There are some well-known CNN architectures, such as AlexNet [29], VGGNet [30], ResNet [31], used as pre-trained model and then fine-tune on the target datasets which is called transfer learning. Khorrami et al. [32] showed that the convolutional neutral network is effective, and they introduced a method to decipher which part of the face image influences the CNNs predictions. Tang et al. [33] demonstrated that switching from softmax to SVM is simple and beneficial for classification by optimizing the margin-based loss rather than cross-entropy loss. Dehghan et al. [34] proposed a network which can automatically identify age, gender, and facial expressions by using several convolutional neural networks. Nguyen et al. [9] proposed a VGG-similar network composed of a stack of convolutional blocks. Due to the capability of recollecting information about the past inputs, RNN has the ability to learn relative dependencies with images, which is advantageous in comparison with CNN. Therefore, RNN is generally combined with CNN in order to achieve better performance in image processing tasks such as image recognition and segmentation. Jain et al. [35] presented a hybrid convolution-recurrent neural network method for FER in images to unify RNN to extract the temporal dependencies which exist in the images. Chernykh et al. [36] developed CNN+RNN model for video and speech recognition. In addition, Gui et al. [37] proposed a novel curriculum learning technique which leads to better generalization for emotion recognition from facial expressions. To our knowledge, it was the first to apply curriculum learning to facial expression recognition. Differing from [37], this paper proposes a new curriculum design method where samples are divided into subsets with different complexities according to the distance to the cluster center, while [37] defines a complexity function to measure the expression intensity of expression where higher intensity indicates lower complexity. Building on top of curriculum learning and CNN work for facial expression recognition, our work applies curriculum learning to pre-trained model which differs from the previous approaches.

3 Our proposed approach

In this section, the proposed curriculum learning and self-selection mechanism are detailed. The model is trained by the curriculum learning training strategy, which is called the main model. And the self-selection mechanism is used for model prediction. During the test phase, the main channel or the second channel is automatically selected based on the prediction of main model. In the following sections, we demonstrate in detail: (1) The design of learning curriculum which is how to distinguish between simple and complex data. (2) The reasons for the introduction of SSM and its working mechanism.

3.1 Curriculum learning

In the case of facial expression recognition task, two assumptions are defined: (a) Facial expressions images have different complexity. (b) In the feature space, the closer the feature vectors are to the clustering center, the more likely they are to be simple data. Curriculum learning is motivated by human learning, in which the model starts from learning simple samples and then gradually takes more complex tasks into training process. It contains two main steps: curriculum design and curriculum learning. In the curriculum design phase, an unsupervised density–distance clustering method is utilized to determine the clustering center of each category and the dataset is divided into three subsets with different complexity according to the distance between each sample and the clustering center in the feature space. In the cur-

riculum learning phase, based on the subsets that have been subdivided, the model was optimized from simple subset which combines the simple subsets over all categories. Then, the capability of model is improved gradually by continuously adding the data with increasing complexity during the training process.

(1) *Curriculum design*

The goal of curriculum design is to divide the dataset into three subsets of varying complexity. Inspired by recent clustering algorithm described in [38], we conduct a density–distance clustering algorithm to determine the clustering center of each category using the product of the local density value and the distance value of each sample.

Firstly, the initial *InceptionResnet-V2* model is utilized as pre-trained model to train all training sets. Then, we can get their feature vectors in deep feature space by using the *fc_256* layer features of the initial model. For each image x_i,we can get $x_i \rightarrow f(x_i)$.

Secondly, we calculate a Euclidean distance of every sample in each category, and we can get a Euclidean distance matrix M as,

$$M_{ij} = \left\| f(x_i) - f(x_j) \right\|^2, 0 \le i \le n, 0 \le j \le n \tag{1}$$

where n is the number of samples in the current category, and M_{ij} indicates a similarity value between x_i and x_j (a smaller M_{ij} means higher similarity between x_i and x_j).

Thirdly, we calculate a local density value ρ_i of each image x_i in each category.

$$\rho_i = \sum_{j=1}^{n} S(M_{ij} - t_c) \tag{2}$$

$$S(x) = \begin{cases} 1, & x < 0 \\ 0, & \text{other} \end{cases} \tag{3}$$

where $S(x)$ is a threshold function, and t_c is a distance threshold. We determine t_c by sorting $\frac{n \times (n-1)}{2}$ distances in M ascending order and select a number which is ranked at $k\%$. In our experiments, we set k =50 and ρ_i is the number of samples whose distance to x_i is smaller than t_c . Naturally, we assume that simple samples have similar visual appearance between each other, and these images projected closely to each other, so it will have a large local density value. On the contrary, complex samples have a significant visual diversity, leading to a sparse distribution with a smaller local density value.

Fourthly, a distance value d_i of each image x_i is defined.

$$d_i = \begin{cases} \min\limits_{1 \le j \le n, j:\rho_j > \rho_i} (M_{ij}) & \text{if} \exists j s.t. \rho_j > \rho_i \\ \max\limits_{1 \le j \le n} (M_{ij}) & \text{otherwise} \end{cases} \tag{4}$$

As for image x_i , if there exists an image x_j having $\rho_j > \rho_i$, the distance d_i is the minimum M_{ij} of all samples that satisfy $\rho_j > \rho_i$. Otherwise, d_i is the distance between x_i and the sample which is the farthest from x_i.

$$z_i = \rho_i \times d_i \tag{5}$$

Fifthly, a sample with maximum value of z_i is selected as clustering center for this category. The determination of clustering center is shown in Fig. 2, and the cluster diagram is shown in Fig. 3.

Finally, according to the Euclidean distance from each sample to the clustering center, the dataset is divided into three subsets. As we have computed a clustering center for the category, a closer sample to the cluster center has a higher confidence to be a simple one. Therefore, after the Euclidean distance from all samples to the clustering center is sorted ascending, all data are divided into three subsets according to the appropriate data ratio: simple subset, hard subset, and complex subset.

(2) *Curriculum learning*

The curriculum learning is a process of model optimization. The designed curriculum is able to discover underlying data structure based on visual appearance. Following the basic ideas of curriculum learning, we design an optimized strategy with increasing complexity of the training set, and training is proceeded sequentially from easier task to harder ones. A multistage training process is developed where a convolutional model is trained by continuously adding harder samples to training set to increase the difficulty of training set. For comparison experiments, we also design a multistage training process by replacing training set with harder samples. Figure 4 shows the training details.

In the training process by mixing harder samples to training set, firstly we train the model by only using the easiest data (simple subset), where images within each category have similar visual appearance. In this way, the model learns basic and clear features of each category and lays the foundation for the subsequent process of more robust features. Secondly, when the model trained in the first stage converges, we start the second stage model training by adding harder data where images have more significant diversity. The model learns more discriminative features, which improves the performance of model. Thirdly, after the model converges, we add more complex data to training set where images have more discrete and indistinguishable features. We observe that the accuracy does not decrease due to indistinguishable features. In the contrast, it leads to better generalization of the model and allows model to avoid over-fitting over the simple data. Finally, when the model in the last stage is converged where three subsets are all combined, the basic learning rate of the model reduced tenfold to achieve fine-tuning. The optimization procedure is shown in Algorithm 1.

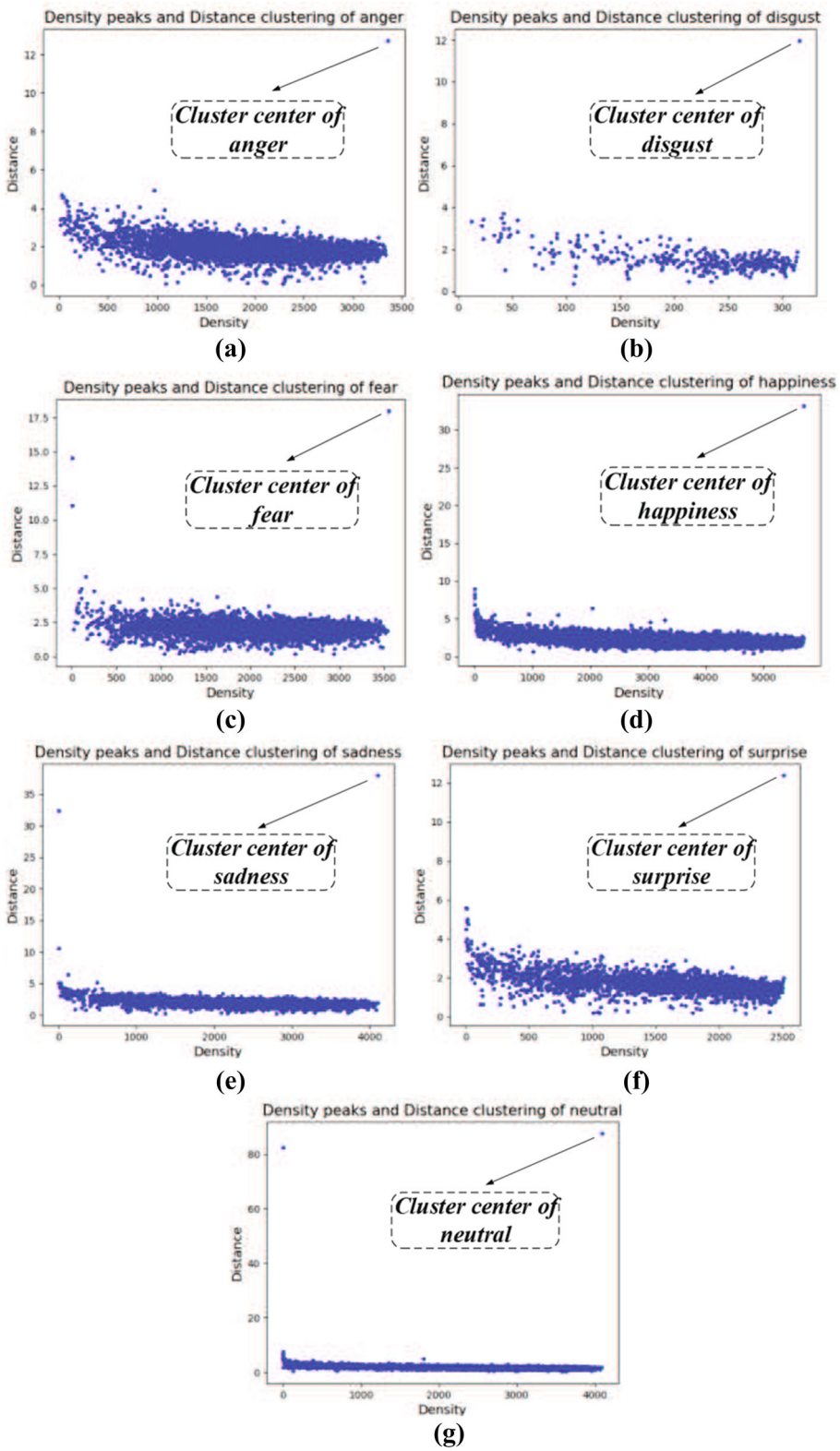


Fig. 2 The determination of the clustering center of each category. The point in the upper right corner is selected as the clustering center

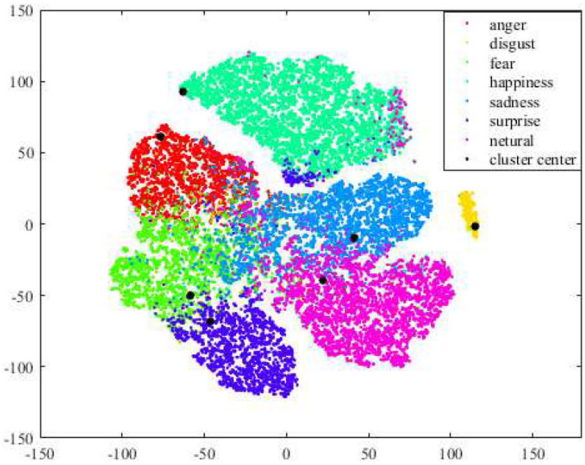


Fig. 3 Cluster diagram of features. The clustering center of each class is represented by black dots

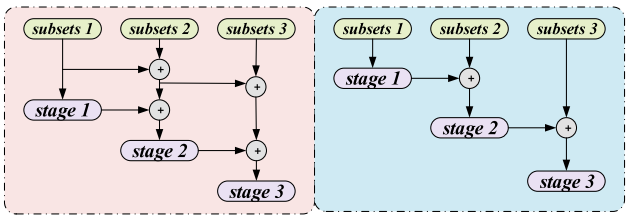


Fig. 4 Training process with designed curriculum. To increase the difficulty of training set, left is the training process by mixing harder samples to training set and right is the training process by replacing training set with harder samples

Algorithm 1 Optimization with curriculum learning

Input: Input dataset $D = \{D_i\}_{i=1}^{d}$ ordered by designed learning curriculum
Output: Optimal model parameter W^*

```
D_train = ∅
for i = 0 → d do
    D_train = D_train ∪ D_i
    for epoch = 1 → k do
        train model(W, D_train)
    end for
end for
update learning rate γ
train model(W, D_train)
```

3.2 Self-selection mechanism

Although FER tasks can achieve satisfactory accuracy under ideal conditions, the recognition ability of a single model for seven facial expressions is different. According to the clustering centers of each category, the Euclidean distances between cluster centers are calculated according to (6).

$$\mathrm{dist}(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

From Tables 1 and 2, it is observed that fear is the most confusing expression. Anger and surprise are also relatively confusing expressions. According to the confusion matrix analysis of some model, we find that the accuracies of anger, fear, and sadness are relatively poor among the seven expressions which are easy to confuse with each other, but surprise expression presents a satisfactory recognition accuracy. Taking these two situations into consideration, we conclude that at the level of expression, anger, fear, and sadness are complex expressions which are relatively difficult to be recognized. Furthermore, in many cases, a single model alone could only identify seven expressions with limited ability and has some difficulty in recognizing the three expressions which have similar movement characteristics and AUs, such as eyebrows together, corners of the mouth down, so further judgment is necessary to make on the more complex ones. Based on this, the self-selection mechanism is introduced which is different from model fusion. The pipeline of SSM is shown in the test phase of Fig. 1.

First, we use the three datasets to train a tri-classifier whose structure is same as main model except the number of neurons in the output layer to classify the three expressions. Cross-entropy and squared-hinge loss functions are used to optimize the tri-classifier. Second, in the test phase, main model and tri-classifier work together in the form of self-selection mechanism, namely main model and tri-classifier are used for the main channel and the second channel, respectively, for further judgment of complex expressions. If the output of the main model is one of the four expressions of disgust, happiness, surprise, and neutral(contempt for CK+), the prediction is directly output. Otherwise further judgment will be made through the tri-classifier. The introduction of SSM takes the poor recognition accuracy into account which reduces the sample space.

4 Experimental results and discussion

4.1 Datasets

Our algorithm is evaluated on FER-2013 and CK+ datasets, some of which are shown in Fig. 5.

FER-2013 [18]: All images in the dataset have been registered and resized to 48×48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region. It contains 28709 training images, 3589 validation images and 3589 test images which are labeled with any of seven expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

Table 1 The Euclidean distance between clustering centers under FER-2013 dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Anger	0	6.367	4.905	7.090	6.273	7.300	9.919
Disgust	6.367	0	7.154	8.870	9.445	8.634	9.957
Fear	4.905	7.154	0	5.233	5.290	4.576	6.268
Happiness	7.090	8.870	5.233	0	7.242	6.262	7.325
Sadness	6.273	9.445	5.290	7.242	0	9.134	8.185
Surprise	7.300	8.634	4.576	6.262	9.134	0	6.157
Neutral	9.919	9.957	6.268	7.325	8.185	6.157	0

Table 2 The Euclidean distance between clustering centers under CK+ dataset

	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0	12.968	16.141	15.060	6.258	15.136	16.156
Contempt	12.968	0	15.109	13.818	15.358	13.551	14.609
Disgust	16.141	15.109	0	12.277	17.642	11.361	12.408
Fear	15.060	13.818	12.277	0	16.569	3.437	3.164
Happiness	6.258	15.358	17.642	16.569	0	16.876	17.658
Sadness	15.136	13.551	11.361	3.437	16.876	0	3.544
Surprise	16.156	14.609	12.408	3.164	17.658	3.544	0

Fig. 5 Samples of FER-2013 dataset (above) and CK+ dataset (below)

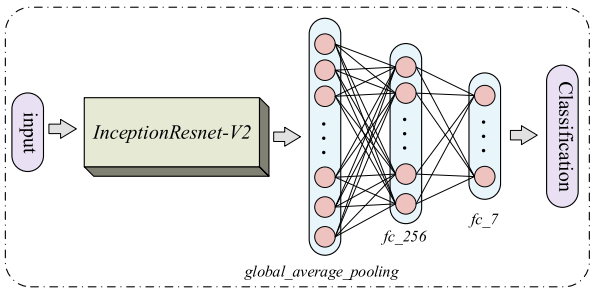


Fig. 6 The main network architecture. Following the pre-trained model is a global average pooling layer, followed by a fully connection layer with 256 neurons, and finally an output layer with 7 neurons. There is a dropout layer with a probability of 0.5 in both the global average pooling layer and the fully connection layer (except the output layer)

Extended Cohn-kanade (CK+) [6]: In contains 593 sequences across 123 subjects, including seven expressions: anger, contempt, disgust, fear, happy, sadness, and surprise. In this paper, we extract the last one to three frames with peak formation of each sequence. Hence, a total of 981 images are utilized for all experiments in a tenfold cross-validation.

4.2 Experiments setup

Our experiments are conducted using python based on keras and tensorflow frameworks on the computer with the following specifications: Inter(R) Core(TM) i5-8500 CPU@3.00GH, Ubuntu Operating System 16.04.4 64 bit, 64GB RAM.

4.3 Train details

Data preprocessing: Firstly, we normalize data per image. We subtract the mean value from each image and then set the standard deviation to 3.125 [9]. Secondly, we normalize data per pixel. For each image, we subtract each pixel from its mean value which is computed from the average of all corresponding pixels and then set the standard deviation of each pixel over all training images to 1.

Training: The *InceptionResnet-V2* is applied as pre-trained model, and the model architecture is shown in Fig. 6. In the training process, we adopt Adam optimizer to minimize the loss function where the batch size is set to 32. To reduce the risk of over-fitting, we apply data augmentation technologies which include random rotation, vertical and horizontal offset, random cropping, random scaling, and random flip

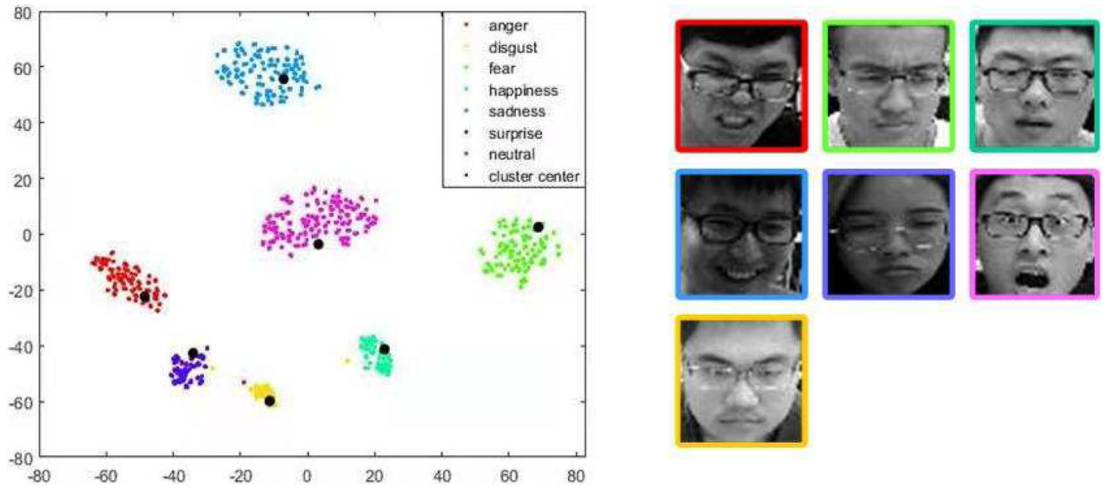


Fig. 7 Some clustering results using real faces as examples

Table 3 The number of expressions contained in each subset

Dataset	Subset	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral (Contempt)	Total
FER-2013	Simple	2398	262	2459	4330	2899	1904	2980	17232
	Hard	799	87	819	1443	966	633	993	5740
	Complex	798	87	819	1442	965	634	992	5737
CK+	Simple	75	96	42	116	44	131	29	533
	Hard	25	32	14	39	14	44	9	117
	Complex	24	31	13	38	14	43	9	172

horizontal when all datasets are used for training and dropout technique to the fully connected with a dropout probability of 0.5. In the training of the three-stage model, the early stopping technique is used to prevent the over-fitting when the loss on validation set does not reduce in t epochs where t is set to 5. After each epoch, the training data is randomly shuffled.

In our experiments, two loss functions are utilized: cross-entropy and squared-hinge. The cross-entropy loss function is defined as follows,

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_{\text{true}}^{(i)} \log y_{\text{pred}}^{(i)} \tag{7}$$

The squared-hinge loss function is defined as follows,

$$L = \frac{1}{N}\sum_{i=1}^{N} \left(\max(0, 1 - y_{\text{pred}}^{(i)} \times y_{\text{true}}^{(i)})\right)^2 \tag{8}$$

where N is the number of all samples, y_{true} is the ground truth and y_{pred} is the prediction of the model, respectively, in (7) and (8).

4.4 Results and evaluation

To evaluate the effectiveness of our proposed method, extensive experiments are conducted, and we compare our approach with baseline models. Firstly, some clustering results using real faces are utilized to emphasize the effective of clustering center selection. Secondly, the effects of curriculum learning strategy are investigated by comparing five different models. Thirdly, the effect of self-selection mechanism is demonstrated by using different main models. Finally, the proposed approach is compared with the baseline models.

(1) ***The clustering results using real faces***

In order to prove the practicability of this clustering method in facial expressions, the clustering center is verified on real faces. We collect 1050 images of 15 subjects in the laboratory to form a real face dataset and verify the clustering method on this dataset. As shown in Fig. 7, the universality of this clustering method on real faces can be proved.

(2) ***The affection of curriculum learning***

According to the designed curriculum, all training sets are divided into three subsets: simple subset, hard subset, and complex subset in a ratio of 6:2:2. The number of expressions contained in each subset is shown in Table 3.

Table 4 The training accuracy of the five models on the test set.We have performed 10 runs for each model

Dataset	Method	Max (%)	Average (%)
FER-2013	Model-1	71.47	70.92
	Model-2	71.69	70.95
	Model-3	71.36	70.74
	Model-4	71.66	70.95
	Model-5	70.55	69.88
CK+	Model-1	97.98	94.14
	Model-2	99.99	98.18
	Model-3	98.59	97.37
	Model-4	98.59	97.37
	Model-5	97.58	96.57

To explore the effect of curriculum on training expression recognition models, we compare five models using different training strategies which are described as follows. For a fair comparison, all models are evaluated using softmax classifier. We design Model-2 which has a higher accuracy when comparing with the Model-1 directly training. This makes us to guess that the order of complexity of the training sample affects the performance. Based on this, Model-3 is designed (complex samples first) whose performance is poor compared with Model-2 (easy samples first). Furthermore, in order to explore more details about single subset on training process and verify the effectiveness of the curriculum we designed, Model-4 and Model-5 are designed which change the complexity of the training set by replacing current training set using easier or harder samples.

—Model-1: The model is trained directly by using the whole training set.

—Model-2: The model is trained from simple subset to complex ones by adding complex dataset to training set.

—Model-3: The model is trained from complex subset to simple ones by adding simple dataset to training set.

—Model-4: The model is trained from simple subset to complex ones by replacing training set with more complex ones.

—Model-5: The model is trained starting from complex subset to simple ones by replacing training set with simpler subset.

The average and max accuracies of five models are shown in Table 4. We can see that the classification result of Model-4 is comparable to those of Model-2 on FER-2013 dataset. Furthermore, Boxplot diagrams for the distribution of classification results of each model are depicted in Fig. 8. It is obvious that Model-2 achieves superior performance, and the result of Model-2 indicates the overall smallest variation, compared with other larger variations of the other results.

The results of the Model-4 and Model-5 in the training process are recorded in Fig. 9. Model-4 (replacing current training set with more complex subset) starts training from simple samples, so its initial accuracy is higher than Model-

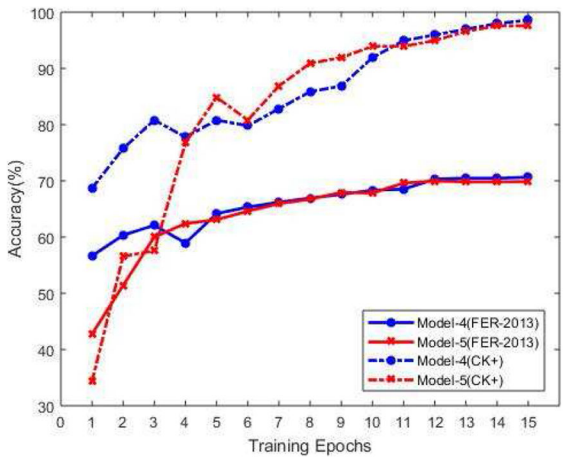


Fig. 9 Accuracy of the Model-4 and Model-5 in the training process

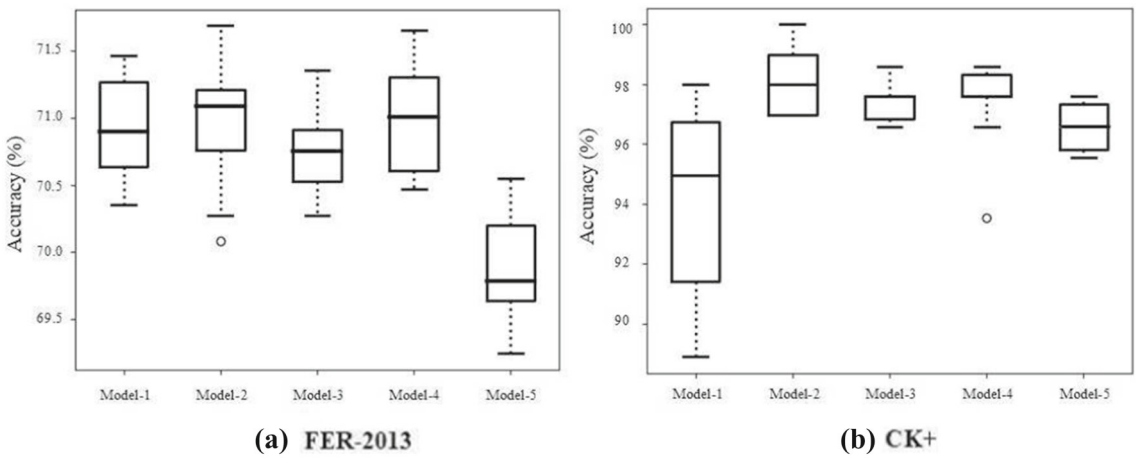


Fig. 8 Boxplot diagrams for the distribution of classification results for each model based on cross-entropy loss function over 10 runs

Table 5 The tri-classifier model accuracy using different loss function

Dataset	Loss function	Max (%)	Average (%)
FER-2013	Cross-entropy	69.50	68.10
	Squared-hinge	69.13	67.76
CK+	Cross-entropy	93.10	77.59
	Squared-hinge	89.66	80.01

We have performed 10 runs for each model

5 which starts training from complex samples. In the fourth epoch, there was a significant decrease in the accuracy of Model-4, which was caused by the complex samples training alone. However, Model-5 is trained with increasingly simple samples, so this phenomenon will not occur. From Fig. 9, we notice that Model-4 achieves better performance than Model-5. On the one hand, it shows the effectiveness of the curriculum designed. On the other hand, it shows that the training sequence of simple and complex samples will greatly affect the generalization of the model. Experiment results show that the curriculum that we designed with appropriate training order has a positive effect on the generalization ability of the model. In the following experiments, we adopt the curriculum learning training strategies of Model-2.

(3) ***The affection of self-selection mechanism***

After analysis of confusion matrix of the single model, it can be easily seen that almost all the models have a poor recognition accuracy for anger, fear, and sadness facial expressions. On the expression level, these three expressions are the most difficult expressions of the seven to identify, so a self-selection mechanism is proposed to make further judgment in the test phase.

To prove the proposed SSM, experiments are conducted on two main models (with/without CL) on two tri-classifier models (cross-entropy/squared-hinge). Table 5 shows the tri-classifier accuracy with two loss functions, and the tri-classifier (cross-entropy) obtains the better result. According

Table 6 The model accuracy after the introduction of SSM on FER-2013 dataset

Dataset	Time	Model (%)	Model + SSM (%)		Model + CL (%)	Model + CL + SSM (%)	
			Tri is 69.50%	Tri is 69.13%		Tri is 69.50%	Tri is 69.13%
FER-2013	1	70.35	70.80	70.83	70.08	70.35	70.44
	2	70.47	70.84	70.86	70.27	70.66	70.71
	3	70.58	70.95	70.97	70.72	70.88	70.93
	4	70.80	70.96	71.01	70.88	71.02	71.08
	5	70.80	70.94	71.00	71.02	70.88	70.94
	6	70.99	71.16	71.07	71.16	71.22	71.30
	7	71.08	71.14	71.22	71.19	71.26	71.22
	8	71.33	71.37	71.44	71.22	71.28	71.33
	9	71.38	71.42	71.49	71.25	71.45	71.50
	10	71.47	71.41	71.52	71.69	71.91	72.11

We have performed 10 runs for each model

Table 7 The model accuracy after the introduction of SSM on CK+ dataset

Dataset	Time	Model (%)	Model + SSM (%)		Model + CL (%)	Model + CL + SSM (%)	
			Tri is 93.10%	Tri is 89.66%		Tri is 93.10%	Tri is 89.66%
CK+	1	88.89	92.93	91.92	96.97	97.98	96.97
	2	89.90	91.92	90.91	96.97	96.97	95.96
	3	90.91	94.95	93.94	96.97	97.98	96.97
	4	92.93	93.94	92.93	96.97	97.98	96.97
	5	93.94	95.96	94.95	97.98	97.98	96.97
	6	95.96	97.98	96.97	97.98	98.99	97.98
	7	95.96	96.97	95.96	98.99	98.99	97.98
	8	96.97	96.97	95.96	98.99	99.99	98.99
	9	97.98	96.97	95.96	99.99	98.99	97.98
	10	97.98	97.98	96.97	99.99	98.99	97.98

We have performed 10 runs for each model

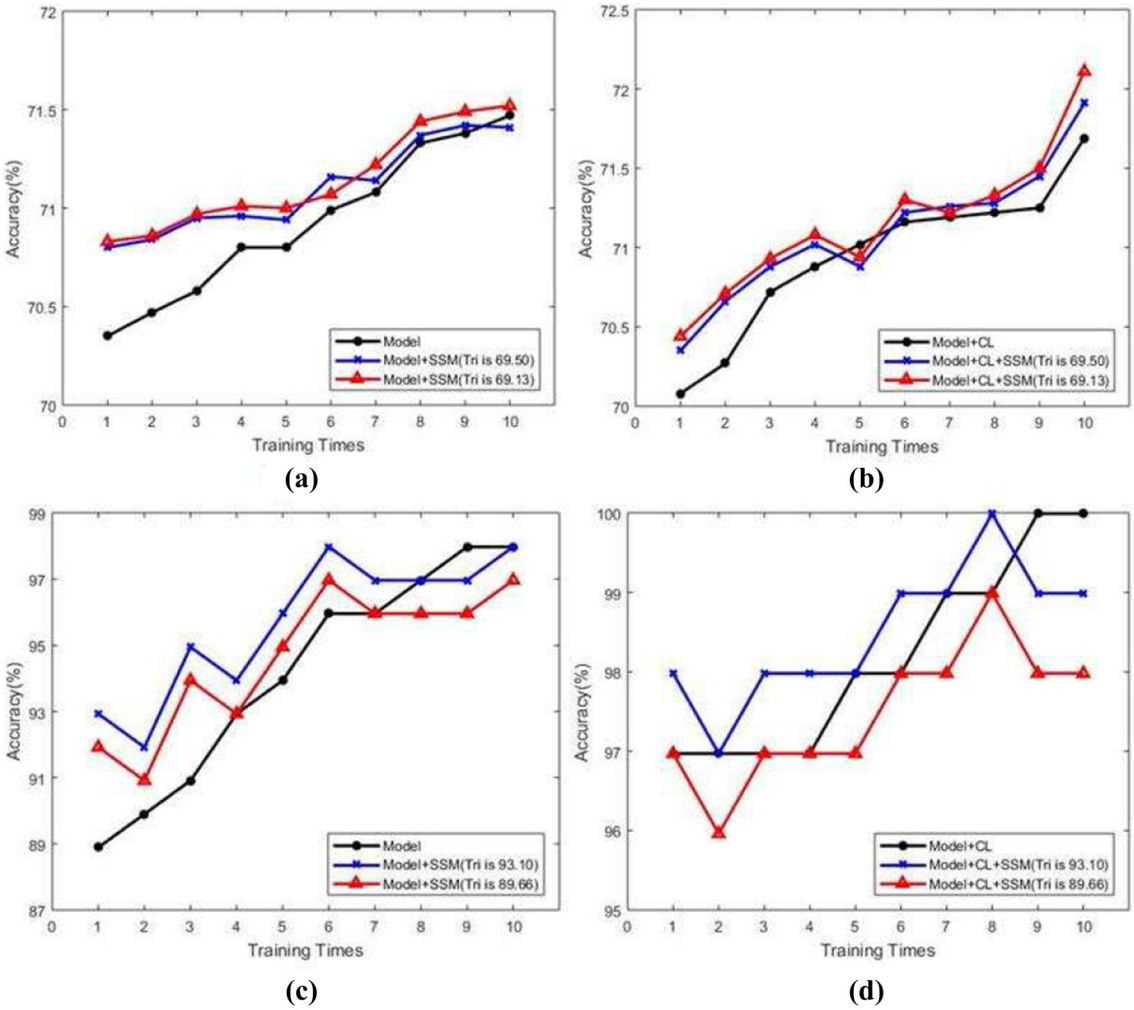


Fig. 10 The result of *model* and *model* + *CL* after the introduction of SSM. **a**, **b** FER-2013 dataset. **c**, **d** CK+ dataset

to Tables 6 and 7, the accuracy of the model was significantly improved after SSM was introduced except the *model* + *CL* on CK+ dataset. This may be due to the excellent performance 99.99% of *model* + *CL* on CK+ dataset. Figures 10 and 11 show the performance of the model with/without CL after the introduction of SSM under two tri-classifiers (cross-entropy/squared-hinge), respectively. Tri-classifier using the squared-hinge loss worked with SSM is effective on FER-2013 dataset even though individual training has poor accuracy. This may be due to the different abilities of the two tri-classifier models to recognize different expressions. There is an experiment in which the SSM did not work. We assume that this may be due to the relatively good recognition capability of the main model for these three expressions compared with the tri-classifier. However, the accuracy of the main model is poor, which indicates that the recognition ability for the other four expressions is limited. In this paper, the SSM is introduced for those three expressions because almost all the models are limited in recognizing these three expressions.

(4) *Comparison with baseline models*

The performances of our proposed method and baselines are shown in Tables 8 and 9. From Table 8, the 1-4 rows are the top four teams in the *Kaggle* competition [39]. BKVGG14 and BKVGG12 were proposed in [9] to build a VGG-similar network, of which results were 71.4% and 71.9%, respectively. The experimental results show that our proposed method outperforms BKVGG12 network which sets the state-of-the-art on FER-2013 dataset so far. From Table 9, the performance of *model* + *CL* is better than other state-of-the-art algorithms.

Furthermore, Tables 10 and 11 list the confusion matrix for FER-2013 and CK+ datasets. From Table 10, among seven expressions, happiness and surprise achieve excellent performances with the accuracy of 90.67% and 81.73%, owing to their distinctive features in the regions of eye and mouth. In fact, they are the most distinguishable expressions for human. Although the SSM is introduced to make the second judgment of the three expressions which are anger (65.38%), fear

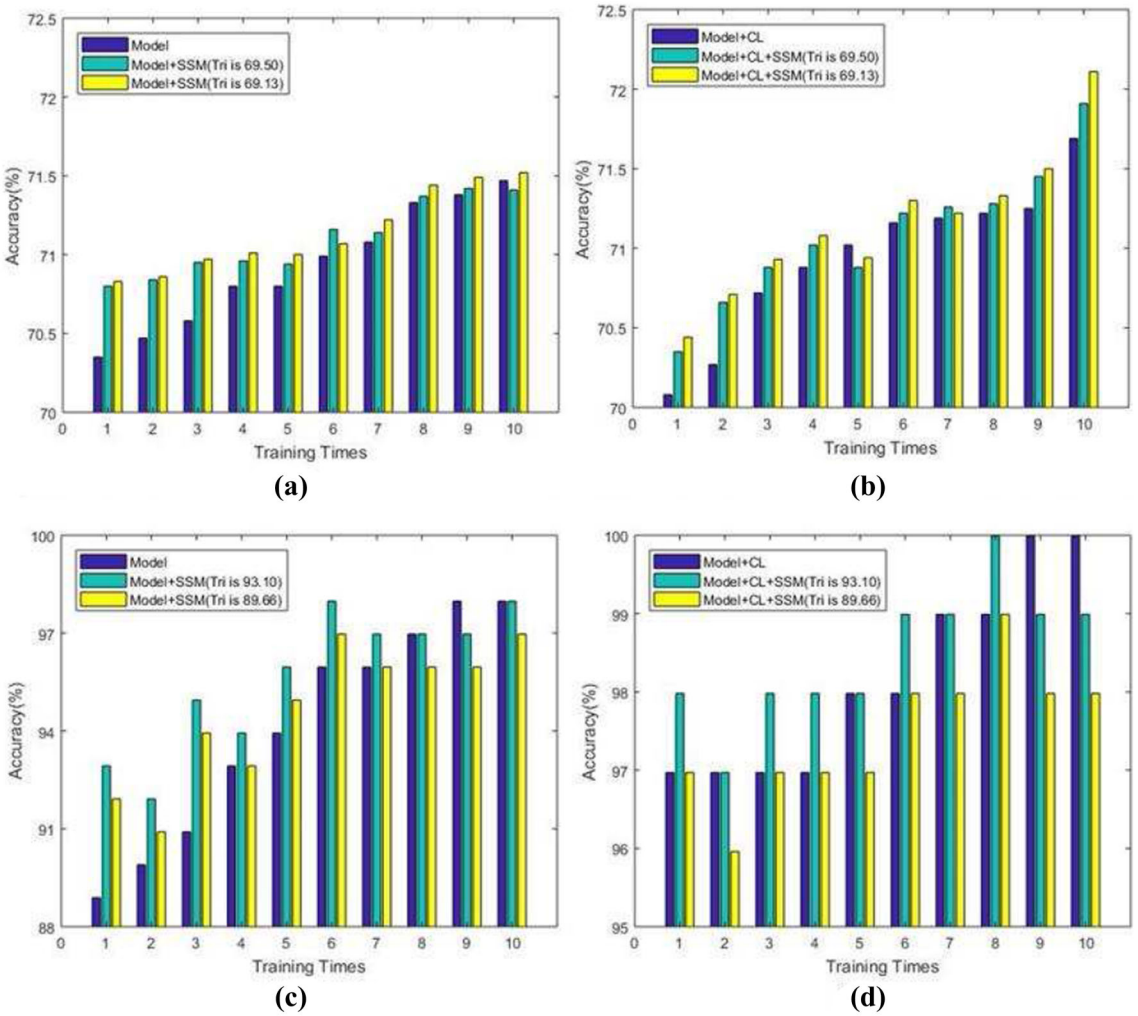


Fig. 11 The improvement of the SSM based on *model* and *model* + *CL*. **a**, **b** FER-2013 dataset. **c**, **d** CK+ dataset

Table 8 Comparison of our result on FER-2013 dataset with some baseline models

Model architecture	Accuracy (%)
SIFT+MKL [40] (*Radu+Marius+Cristi*)	67.5
CNN (*Maxim Milakov*)	68.8
CNN (*team Unsupervised*)	69.3
CNN+SVM Loss [33] (*team* RBM)	71.2
BKVGG14 [9]	71.4
BKVGG12 [9]	71.9
Model	71.47
Model + CL	71.69
Model + SSM	71.52
Model + CL + SSM	72.11

The *Model* is directly trained without the introduction of CL and SSM

Table 9 Comparison of our result on CK+ dataset with some baseline models

Model architecture	Accuracy (%)
Ouellet [41]	94.40
STM-ExpLet [42]	94.19
3DCNN-DAP [43]	92.40
DTAGN (weighted sum) [13]	96.94
DTAGN (Joint) [13]	97.25
Li et al. [14]	97.38
Model	94.14
Model + CL	98.18
Model + SSM	95.66
Model + CL + SSM	97.48

The *Model* is directly trained without the introduction of CL and SSM

(53.41%), and sadness (61.45%), they remain the three most confusing of the seven expressions. In addition, although the number of disgust samples is very small, it has achieved relatively satisfactory recognition accuracy 69.09%. More precisely, the neutral expression is easily misclassified as sadness and the percentage of the sadness expression falsely

Table 10 The confusion matrix on FER-2013 dataset with the introduction of SSM

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Anger	65.38	1.02	9.37	2.44	12.02	1.83	7.94
Disgust	21.82	69.09	1.82	1.82	1.82	0	3.64
Fear	12.12	0.19	53.41	2.46	16.67	7.39	7.77
Happiness	2.39	0	0.68	90.67	1.48	2.05	2.73
Sadness	8.92	0.34	9.43	2.02	61.45	1.35	16.50
Surprise	4.09	0	6.97	4.09	1.68	81.73	1.44
Neutral	4.15	0.16	4.95	3.83	14.86	0.96	71.09

Table 11 The confusion matrix on CK+ dataset with the introduction of SSM

(%)	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Anger	100	0	0	0	0	0	0
Contempt	0	100	0	0	0	0	0
Disgust	0	0	94.44	0	0	5.56	0
Fear	0	0	0	100	0	0	0
Happiness	0	0	0	0	100	0	0
Sadness	0	0	0	8.33	0	91.67	0
Surprise	0	0	0	0	0	0	100

classified as neutral is 16.50%. From Table 11, our algorithm performed well except for disgust (94.44%) and sadness (91.67%) expressions. In general terms, expressions are easily confused due to the similarity in shape and appearance features, and the individual variations for the same expression.

5 Conclusions

In this paper, inspired by curriculum learning, we present a novel curriculum design method and apply it to facial expression in the training phase. The dataset is divided into three subsets with different complexity according to the distance from the sample to the clustering center in the feature space. Then, the model is trained by adding complex samples to training set which leads to better generalization. Particularly, in the test phase, a self-selection mechanism is introduced to further judge the output of the main model. Only if the prediction of the main model is one of the anger, fear, and sadness, the judgment of the tri-classifier will be enabled. In the end, the experiment results demonstrate that the proposed approach outperforms other state-of-the-art methods in terms of the accuracy for 7-class expressions on the well-known FER-2013 and CK+ datasets.

For future work, we would like to extend our method from images to videos and exploit different curriculum design strategies to facial expression recognition tasks. More specifically, we also aim at investigating the relationships between subsets of different complexity. In addition, we would like to investigate the most distinguish features of anger, fear, and sadness expressions.

Acknowledgements This study was funded by the Key Program of Scientific and Technological Innovation of Shandong Province (Grant No. 2017CXGC0926), Key Research and Development Program of Shandong Province (Grant No. 2017GGX30133), National Key Research and Development Program of China (Grant No. 2017YFB1302400), National Natural Science Foundation of China (Grant No. 61773242).

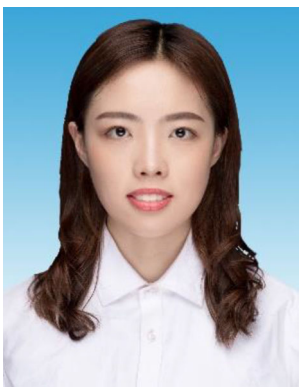
Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
2. Li, S., Deng, W.: Deep facial expression recognition: a survey (2018)
3. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
4. Zeng, Z., Pantic, M., Roisman, G.I., et al.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
5. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
6. Lucey, P., Cohn, J.F., Kanade, T., et al.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotionspecified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, pp. 94–101 (2010)
7. Agrawal, A., Mittal, N.: Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. Vis. Comput. (2019). https://doi.org/10.1007/s00371-019-01630-9
8. Liang, D., Liang, H., Yu, Z., et al.: Deep convolutional BiLSTM fusion network for facial expression recognition. Vis. Comput. (2019). https://doi.org/10.1007/s00371-019-01636-3
9. Ayata, D., Yaslan, Y., Kamasak, M.E.: Emotion based music recommendation system using wearable physiological sensors. IEEE Trans. Consum. Electron. **64**, 196–203 (2018)
10. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
11. An, F., Liu, Z.: Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM. Vis. Comput. (2019). https://doi.org/10.1007/s00371-019-01635-4
12. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
13. Jung, H., Lee, S., Yim, J., et al.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the 2015 IEEE International Conference on Computer Vision.

Santiago, CentroParque Convention Center, Chile, pp. 2983–2991 (2015)
14. Li, K., Jin, Y., Akram, M.W., et al.: Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis. Comput. (2019). https://doi.org/10.1007/s00371-019-01627-4
15. Mohammadi, M.R., Fatemizadeh, E., Mahoor, M.H.: PCA-based dictionary building for accurate facial expression recognition via sparse representation. J. Vis. Commun. Image Represent. **25**(4), 1082–1092 (2014)
16. Gogić, I., Manhart, M., Pandžić, I.S., et al.: Fast facial expression recognition using local binary features and shallow neural networks. Vis. Comput. (2018). https://doi.org/10.1007/s00371-018-1585-8
17. Mavadati, S.M., Mahoor, M.H., Bartlett, K., et al.: Disfa: a spontaneous facial action intensity database. IEEE Trans. Affect. Comput. **4**(2), 151–160 (2013)
18. Goodfellow, I.J., Erhan, D., Carrier, P.L., et al.: Challenges in representation learning: a report on three machine learning contests. Neural Netw. **64**, 59–63 (2015)
19. Dhall, A., Ramana Murthy, O.V., Goecke, R., et al.: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Seattle, Motif Hotel, USA, pp. 423–426 (2015)
20. Dhall, A., Goecke, R., Joshi, J., et al.: Emotiw 2016: video and group-level emotion recognition challenges. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo, Japan, pp. 427–432 (2016)
21. Bengio, Y., Louradour, J., Collobert, R., et al.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, pp. 41–48 (2009)
22. Pentina, A., Sharmanska, V., Lampert, C.H.: Curriculum learning of multiple tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts, USA, pp. 5492–5500 (2015)
23. Avramova, V.: Curriculum learning with deep convolutional neural networks (2015)
24. Guo, S., Huang, W., Zhang, H., et al.: CurriculumNet: weakly supervised learning from large-scale web images. arXiv:1808.01097 (2018)
25. Bartlett, M.S., Littlewort, G., Frank, M., et al.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, California, USA, **2**, pp. 568–573 (2005)
26. Yang, B., Cao, J., Ni, R., Zhang, Y.: Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access **6**, 4630–4640 (2018)
27. Ekman, P., Friesen, W.: Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists, San Francisco (1978)
28. Liu, M., Li, S., Shan, S., et al.: Au-inspired deep networks for facial expression feature learning. Neurocomputing **159**, 126–136 (2015)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Nevada, Lake Tahoe, USA, pp. 1097–1105 (2012)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
31. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, Caesars Palace, USA, pp. 770–778 (2016)
32. Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision Pattern Recognition Workshops. Santiago, CentroParque, Chile, pp. 19–27 (2015)
33. Tang, Y.: Deep learning using linear support vector machines. arXiv:1306.0239 (2013)
34. Dehghan, A., Ortiz, E.G., Shu, G., et al.: Dager: Deep age, gender and emotion recognition using convolutional neural network. arXiv:1702.04280 (2017)
35. Jain, N., Kumar, S., Kumar, A., et al.: Hybrid deep neural networks for face emotion recognition. Pattern Recognit. Lett. **115**, 101–106 (2018)
36. Chernykh, V., Sterling, G., Prihodko, P.: Emotion recognition from speech with recurrent neural networks. arXiv:1701.08071 (2017)
37. Gui, L., Baltrušaitis, T., Morency, L.P.: Curriculum learning for facial expression recognition. In: Proceedings of International Conference on Automatic Face and Gesture Recognition. Washington, DC, USA, pp. 505–511 (2017)
38. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
39. Challenges in representation learning: Facial expression recognition challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge (2013)
40. Ionescu, R.T., Popescu, M., Grozea, C.: Local learning to improve bag of visual words model for facial expression recognition. In: Workshop on Challenges in Representation Learning, ICML. Atlanta, GA, USA (2013)
41. Ouellet, S.: Real-time emotion recognition for gaming using deep convolutional network features. arXiv preprint arXiv:1408.3750 (2014)
42. Liu, M., Shan, S., Wang, R., et al.: Learning expressionlets on spatiotemporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
43. Liu, M., Li, S., Shan, S., et al.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian Conference on Computer Vision. Springer, Cham, pp. 143–157 (2014)



Xiaoqian Liu received the B.S. degree in automation in 2017 from Shandong University of Science and Technology, Qingdao, China. She is currently pursuing the M.S. degree in control engineering at Shandong University, Jinan, China. Her research interests include data mining, pattern recognition, computer vision, image processing and deep learning techniques.

Fengyu Zhou received the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2008. He is currently a professor of the School of Control Science and Engineering at Shandong University, Jinan, China. His research interests include service robotics, automation, pattern recognition, image processing, computer vision and cloud robotics.