Neurocomputing 359 (2019) 130–138

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Action Units recognition based on Deep Spatial-Convolutional and Multi-label Residual network

Su-Jing Wang[a,*], Bo Lin[b], Yong Wang[b], Tongqiang Yi[b], Bochao Zou[c,d], Xiang-wen Lyu[d]

[a] Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[b] College of Software, Xi'an Jiaotong University, Xi'an 710000, China
[c] Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing 100054, China
[d] China Academy of Electronics and Information Technology, Beijing 100041, China

ARTICLE INFO

Article history:
Received 3 December 2018
Revised 26 March 2019
Accepted 11 May 2019
Available online 29 May 2019

Communicated by Dr. Qingshan Liu

Keywords:
Sample imbalance problem
AU recognition
Multi-label learning
Local convolution
Residual unit

ABSTRACT

Facial Action Unit (AU) recognition is an essential step in the facial analysis. A facial image has one or more AU(s). Given an AU, the number of images without the AU is far greater than that of images with the AU. So, AU recognition is not only a sample imbalance problem but also a multi-label learning problem. For the two problems, we proposed a novel Multi-label Slope Rate (MSR) loss function and an Advanced-MSR (Ad-MSR) loss function in deep network architecture to recognize AU. For other characters of AU recognition, a local convolution and residual units are used in the architecture. The experimental results on two expression databases labeled AU show that the proposed loss functions not only address overfitting of the network on the training set and enhancing the generalization ability on the test set. The proposed architecture also gets well performance in the databases.

1. Introduction

Facial Action Unit (AU) recognition is an essential step in the facial analysis. Without a robust AU recognition method, facial expression recognition, micro-expression recognition, and so on facial related action problems cannot be effectively solved. Especially, micro-expression recognition needs more robust AU recognition method, because of subtler muscle movements of micro-expression compared with common expressions [1]. Ekman and Friesen [2] proposed Facial Action Coding System (FACS), which is a comprehensive system for describing facial expression by Action Units (AUs). An AU or AU combinations describe a facial expression. For example, AU6 means rising the cheeks and AU12 means pulling lip corners obliquely. While the combination of AU6 and AU12 describe smile. However, this descriptive power is still costly, because manual FACS coding is a very time-consuming task. It often takes a long time for a coder to have an acceptable capability. Once a FACS coder achieves can meet the requirement, it can take a time or longer to encode a video of tens of seconds, and we must always pay attention to the reliability of the encoder. To be able to use FACS more efficiently, computer vision can meet this requirement and automate AU coding. Although significant progress has been made in achieving this goal [3–6], automatic AU recognition is a challenging problem.

For decades, many methods of AU recognition have proposed by many researchers. Zhong et al. [7] uses structured regional learning to improve the accuracy of general basic expressions, this method extracts important points of the face as local areas, which can reduce irrelevant features and increase the accuracy of identifying specific features. Facial landmark points play a crucial role in AU recognition. Many conventional methods extract texture features near the landmark points. Valstar and Pantic [8] extracted Gabor wavelet near 20 landmark points as features and put them into Adaboost and SVM to recognize AU. The facial structure information is obtained by measuring the normalized landmark distances, and the angles of the Delaunay mask formed by the landmark points. In AU detection, the methods of feature extraction are becoming more and more mature. In general, there are geometric features [5,9], texture features [10,11], dynamic features [12–14] or feature fusions [15]. These features are usually quantized by histograms. Tian *et al.* [16] claimed that the automatic face analysis system should recognize fine-grained changes in facial expression into AUs of the FACS, instead of a small set of prototypic expressions, such as happiness, anger, surprise, and fear. They extracted facial structure information as the input of the neural networks with one hidden layer to recognize AUs and AU combinations. Tong *et al.* [17] used a dynamic Bayesian network to model the relationships AUs and their temporal evolutions for AU recognition.

* Corresponding author.
E-mail address: wangsujing@psych.ac.cn (S.-J. Wang).

https://doi.org/10.1016/j.neucom.2019.05.018
0925-2312/

Table 1
Five samples with 11 AUs from EmotionNet Database. 0 means the sample has not the corresponding AU. 1 means the sample has the corresponding AU.

No.	Samples	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU17	AU20	AU25	AU26
1		0	0	0	0	1	0	1	0	0	1	0
2		0	0	0	0	1	0	1	0	0	1	1
3		0	0	0	0	1	0	1	0	0	0	1
4		0	0	0	0	0	0	1	0	0	1	0
5		0	0	0	0	0	0	0	0	0	1	0

The model makes AU recognition more reliable, robust, and consistent. Given an image or video sample, it maybe includes one or more AU(s). That means a sample has multi-label. So, AU recognition is a multi-label learning problem. However, it is difficult to obtain a complete label assignment for each example. There is a strong correlation [18,19] between AUs. For example, when a person smiles, AU6 and AU12 will often appear at the same time. The multi-label learning method in AU detection takes advantage of this strong association, which makes the model have higher classification accuracy [20]. Wu *et al.* [21] propose a method for multi-label learning that explicitly handles missing labels. Zhao *et al.* [22] leveraged group sparsity to identify important facial patches, and learns a multi-label classifier constrained by the likelihood of co-occurring AUs. They proposed Deep Region and Multi-label Learning (DRML) [23] to recognize AUs. In the part of the shallow network, they introduced a new region layer to extract the features of AUs, then combine the region features and used a deep network to extract higher-level features for classification of the expression.

AU recognition also suffers from the sample imbalance problem, that is, the frequency of one class with a given AU can be 100 times less than another class without the AU. The problem has a significant detrimental effect not only on traditional classifiers but also on recent deep learning technology. It affects both convergence during the training phase and generalization of a network on test sets. Methods for addressing the sample imbalance problem are generally divided into two categories [24].

One category is data level methods changing class distribution training set by replicating or removing some samples. In the category, undersampling and oversampling are two methods often used. Undersampling removes some samples from the majority class randomly. So it discards a portion of available samples and is not suitable for deep learning especially, in the small sample case. Oversampling directly replicates randomly samples the from minority class. Oversampling is one of the most commonly used and effective methods in deep learning. However, it maybe lead to overfitting of the network on the training set and reducing the generalization ability on the test set. Furthermore, to directly apply of oversampling is not suitable for AU recognition, because AU recognition is a multi-label task. A sample has more labels. For example, each image sample from EmotionNet Database has 11 labels. Table 1 lists 5 samples with 11 AUs from EmotionNet Database. In the table, 0 means the sample has not the corresponding AU and 1 means the sample has the corresponding AU. The three first samples have AU6, and the two last samples have not AU6. To balance samples based on AU6, oversampling replicates No. 4 sample or No. 5 sample. The replicate can lend to more imbalance for AU25, AU26, and so on. To relieve the problem, Charte *et al.* [25] produce synthetic samples instead of directly replicate.

Another category is model level methods adjusting models or algorithms while keeping the training set unchanged. Cost-sensitive learning is one of the methods in the category. It assigns different cost to misclassification of examples from different classes [26]. Borrowed from the idea, we use different cost to propose novel loss functions to address the sample imbalance problem in multi-label learning tasks.

In this paper, we use CNN and residual unit to build network architecture. To get better performance for AU recognition, the idea of the local convolution is introduced into the architecture. For the sample imbalance problem in multi-label learning, we proposed a novel Multi-label Slope Rate (MSR) loss function by using the proportions of positive and negative samples in each batch. Furthermore, we proposed an Advanced-MSR (Ad-MSR) loss function by using the loss value of the previous batch to enhance the performance of MSR. The proposed methods are evaluated two expression databases with AU labels.

2. Deep Spatial-Convolutional and Multi-label Residual network

2.1. DSCMR architecture

In some situations, CNN has greatly improved the performance of the vision system, including facial verification [27–29], object detection [30], and video tracking [31]. For AU recognition, deep CNN can extract not only each AU feature but also the relationships of AU. That leads to ensure the classification accuracy. There are many AUs which may interact with each other. Their relationship spatial structure can be effectively captured by deep CNN. For example, a surprise is coded as AU1+2+5+26. AU1+2+5 mean that brows and upper eyelids are raised and occur in forehead. AU26 means that lips are relaxed and parted; mandible is lowered. It occurs in chin. The distance between forehead and chin is related long. The one or more layers of convolutional layers are hard to activate neural unit in the same feature map.

So multiple convolutional layers are needed, but this leads to the number of layers in the network is very deep. Despite using of ReLU and optimization technology, sometimes it is impossible to avoid some problems such as vanishing gradient. To solve this problem, the residual learning [32] is used in our model. It has proven to be very effective to use residual learning for training ultra-depth neural networks with more than 1000 layers.

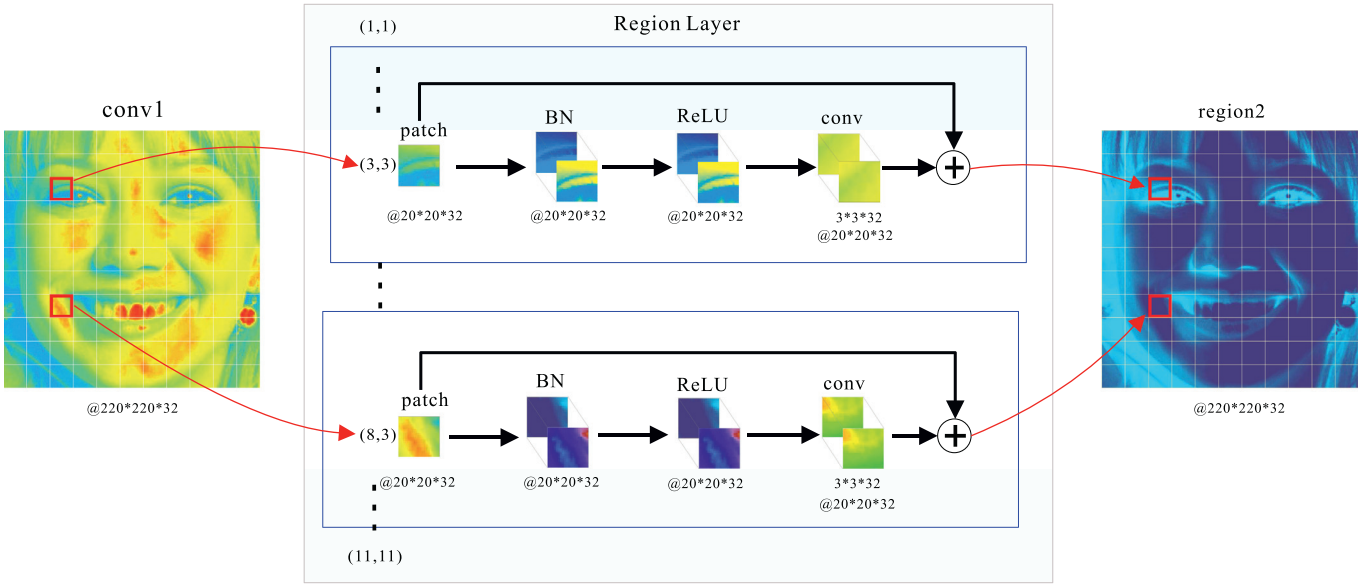


Fig. 1. An illustration of the local convolutional layer.

The local feature is very important not only for face recognition but also for expression recognition and AU recognition. A standard convolutional layer assumes that weights are shared across an entire image. However, facial images are non-grid and its structured information is more important than the holistic information. Based on that, we add the local convolutional layer [23] into the proposed architecture.

Fig. 1 shows that the local convolutional layer includes three parts: patch clipping, local convolution, and identity addition. The patch clipping part divides a 214×214 feature map into 8×8 blocks. Each block is followed by a Batch Normalization (BN) and ReLU. A local convolution part captures local appearance changes and forces the weights in each part to be updated independently. An identity addition part is introduced along with shortcut connection from the input block.

Suppose x is the input, the desired underlying mapping is $H(x)$. If x is directly regarded as the initialized output. Then we let the stacked nonlinear layers fit another mapping of $F(x) = H(x) - x$. The original mapping is recast into $F(x) + x$. It is easier to optimize the residual mapping $F(x) + x$ than to optimize the original, mapping $H(x)$. The formulation of $F(x) + x$ can be realized by feed forward neural networks with shortcut connections. Shortcut connections are those skipping one or more layers. Simply, the shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers. Identity shortcut connections add neither extra parameter nor computational complexity. In this paper, each residual unit includes two convolutional layers.

ResNet-18 is used in our architecture. The input is 224×224 color face images. Following the input, the first convolutional layer includes 64 filters of size 7×7 and is followed by a max-pooling layer. There are 4 residual groups follow the max-pooling layer. The numbers of filters in convolutional layers of each residual group are 64, 128, 256, and 512. Each residual group includes two residual blocks with two 3×3 convolutional layers. In the first block, a convolutional layer including the same number of filters of size 1×1 is used for the shortcut connection. In the second block, the identity shortcut connection is used. Each convolutional layer is followed by Batch Normalization (BN) [33] and Scale layers. The first convolutional layer in each residual block is also followed by ReLU. The second last layer is a fully connected layer with 2046 neural units. For a multi-label task with n labels, the last layer is a fully connected layer with n neural units. To address the sample imbalance problem, we use the proposed Multi-label Slope Rate (MSR) loss function and Advanced MSR loss function in the architecture. The proposed architecture without the local convolutional layer is called as DSCMR and the one with local convolutional layer is called L-DSCMR. L-DSCMR Network Architecture is shown in Fig. 2. In the experiments, we compare the performances of DSCMR and L-DSCMR.

2.2. Multi-label Slope Rate loss

In this section, we introduce Multi-label Slope Rate (MSR) loss function that can solve the sample imbalance problem in AU recognition. For the single label task, the sample space of ith AU is denoted as $S^{(i)}$,

$$S^{(i)} = \{S_-^{(i)}, S_+^{(i)}\} \tag{1}$$

where, $S_+^{(i)}$ denotes the space of positive sample with ith AU. And $S_-^{(i)}$ denotes the space of negative sample without ith AU. In reality, the size of the set $S_-^{(i)}$ is far greater than the size of the set $S_+^{(i)}$. This is the sample imbalance problem. We introduce the balance factor γ to the cross-entropy loss function to address the problem. The cross-entropy loss function can be written as follows

$$L_{cur}^{(i)} = -\frac{1}{m}\sum_{j \in S^{(i)}} \log P_j = -\frac{1}{m}\left(\sum_{j \in S_+^{(i)}} \log P_j + \sum_{j \in S_-^{(i)}} \log P_j\right) \tag{2}$$

where m is the number of all samples. P is the output of the network and indicates the probability of classification of ith AU occurrence. In order to address the sample imbalance problem, The balance factor $\gamma^{(i)}$ ith AU definite:

$$\gamma^{(i)} = \frac{\|S_-^{(i)}\|}{\|S_+^{(i)}\|} \tag{3}$$

where $\|\bullet\|$ denotes the cardinality of a set. If the number of negative samples is greater than the number of positive samples, $\gamma^{(i)}$ is greater than 1. We combine $\gamma^{(i)}$ with Eq. (2) and get

$$L_{cur}^{(i)} = -\frac{1}{m}\left(\sum_{j \in S_+^{(i)}} \gamma^{(i)I(\gamma^{(i)}>1)} \log P_j + \sum_{j \in S_-^{(i)}} \gamma^{(i)(-I(\gamma^{(i)}<1))} \log P_j\right) \tag{4}$$

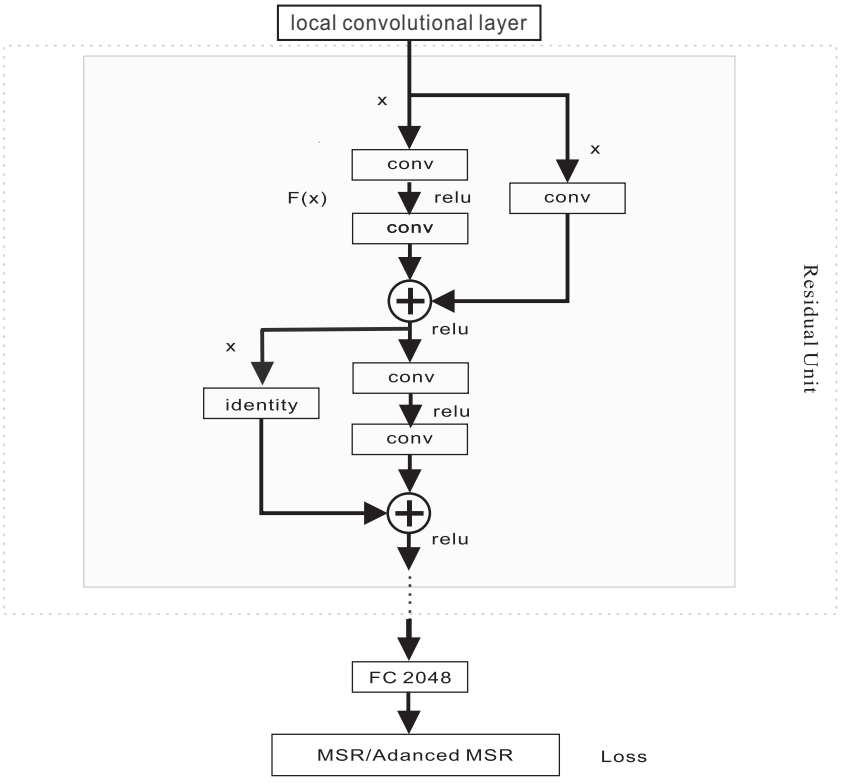


Fig. 2. L-DSCMR network architecture.

where

$$I(\text{condition}) = \begin{cases} 1 & \text{if condition is true;} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

When the number of negative samples is greater than the number of positive samples, we have $\gamma^{(i)} > 1$. So we get $I(\gamma^{(i)} > 1) = 1$ and $I(\gamma^{(i)} < 1) = 0$. Eq. (4) will be following

$$L_{cur}^{(i)} = -\frac{1}{m}\left(\sum_{j\in S_+^{(i)}} \gamma^{(i)} \log P_j + \sum_{j\in S_-^{(i)}} \log P_j\right) \tag{6}$$

Compare Eq. (6) with Eq. (4), we find that the influence of positive samples on loss is enlarged by multiplying $\gamma^{(i)}$ which is greater than 1. So, the sample imbalance problem caused by related less number of positive samples is alleviated.

For the multi-label task with n labels, the loss function can be written following

$$L_{multi} = \sum_{i=1}^{n} L_{cur}^{(i)} \tag{7}$$

2.3. Advanced MSR

In the multi-label task, all labels shares hidden layers and weights. In the procedure of network optimization, it suffers the problem of batch bias. The different training batch includes different samples with huge difference on the number of type of AU. Through several batch training, the network is good at certain AUs, and not good at others. Through more several batch training, the network is not good at the certain AUs, and good at others. In the view of entire training procedure, the accuracy of each AU will fluctuate. This phenomenon comes from the adjustment of the learning ability of the network to other labels, thus affecting the effect of the network on the overall multi-label task. Therefore, we proposed Ad-MSR on the basis of MSR to smooth the network training process for multi-label learning.

We define the $L_{pre}^{(i)}$ is the cost value of ith AU in the previous iteration. In our options, the procedure of the training network is the procedure of updating weight matrices to make the error approximate 0. So, we can draw the conclusion that the error includes the information of impact to classification results caused by the network. On generally, when weight matrices are updated, they are subtracted by the actual gradient value and move to the direction of the minimum gradient. However, the historical information of the training model has not been used during the entire training procedure. Therefore, we introduce the weight factor α during the update and consider the current loss error by using the historical training information by the weighted average.

$$L_S = \frac{1}{n}\sum_{i=1}^{n}\left(\alpha L_{cur}^{(i)} + (1-\alpha)L_{pre}^{(i)}\right) \tag{8}$$

where $L_{pre}^{(i)}$ is the loss value of ith AU in the previous iteration and includes the historical training information by introduced with a small proportion $(1-\alpha)$. In the continuous iteration process, the earlier training results have less influence on the current loss. In addition, because the network is required to finish a multi-label learning, the current loss value need to be calculated by combining the loss values of all labels. Here, we use the arithmetic mean of the all label loss values as the final loss value for this iteration. The historical cost value $L_{pre}^{(i)}$ is updated by Eq. (9) before proceeding to the next iteration:

$$L_{pre}^{(i)} = \alpha L_{cur}^{(i)} + (1-\alpha)L_{pre}^{(i)} \tag{9}$$

In addition, a loss gain coefficient $\beta^{(i)}$ is introduced to indicate the degree of deviation of ith AU for all AUs. After each calculation, we can get each AU's own error. For the AU with smaller errors, it means that the network has better learning ability for it. At this time, $\beta^{(i)} = 0$ means that the learning ability of network is good

Fig. 3. The samples come from EmotioNet database.

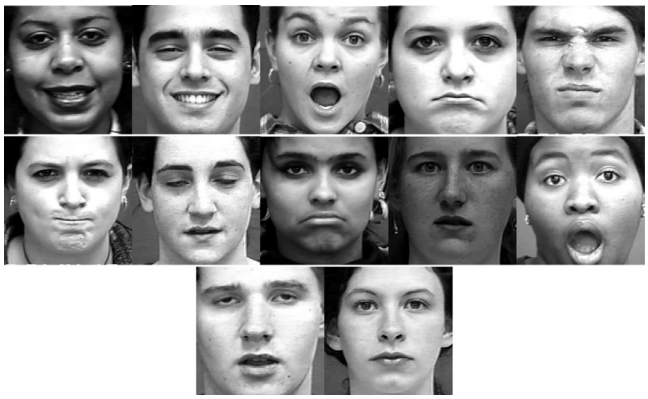


Fig. 4. The samples come from CK+ database.

for *i*th AU. For the AU with larger error, it indicates that the network has poor learning ability for it. So we increase the learning ability of the AU. At this time, the $\beta^{(i)}$ coefficient is:

$$\beta^{(i)} = \frac{(L_{cur}^{(i)} - L_S)}{(L_{cur}^{(i)} + L_S)} + 1 \tag{10}$$

In order to ensure that $\beta^{(i)} \in [1, 2]$, $\beta^{(i)}$ is calculated by the relative deviation of the loss values of each AU. The final loss function of Advanced MSR can be written as follows:

$$L = \sum_{i=1}^{n} \beta^{(i)} L_{cur}^{(i)} \tag{11}$$

3. Experiments

In the section, the proposed DSCMR framework is evaluated by recognizing single AU1 and multi-AUs on two expression databases: CK+ [34] and EmotioNet [35]. We commence by experimental settings and close by discussion.

3.1. Experimental settings

Database: We evaluated DSCMR in two databases that involve posed and spontaneous facial behaviors in varied contexts. Each database has been FACS-coded by well-experienced coders.

1. EmotioNet database includes 950,000 images with annotated AUs. AUs included in the database are AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU17, AU20, AU25, and AU26. The database contains a large number of images from the face of the network, and all of these images are marked with AUs. The samples in the database are shown in Fig. 3.
2. The CK+ database contains and 593 image sequences by captured from 123 subjects. The last frame of each image sequence is labeled with AU. In the 593 image sequences, 327 sequences have the label of the motion. The database is one of the popular databases in facial expression recognition (Fig. 4).

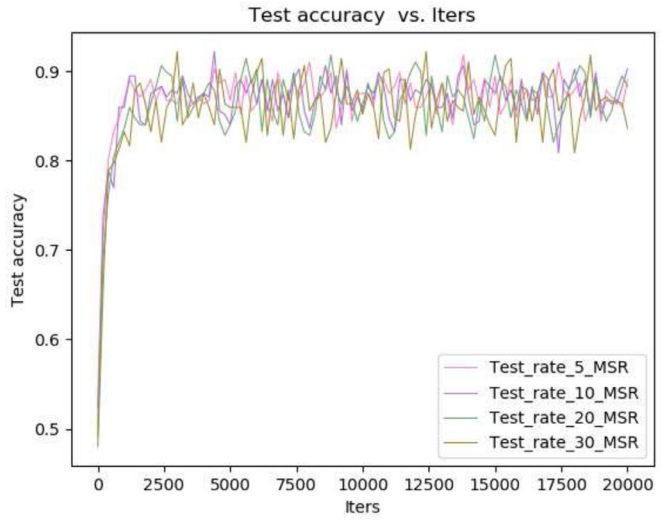
Preprocessing and configuration: We registered face images to 224×224 using similarity transform [27,28]. Each face was horizontally mirrored for data augmentation. All models were initialized with the learning rate of 0.0005, which was further reduced after 2000 iterations. For Eq. (8), $\alpha = 0.5$. A momentum of 0.9 and weight decay of 0.0005 was used. All implementations were based on the Caffe toolbox with modifications to support the region layer and multi-label cross-entropy loss.

Metrics: The performance was evaluated on two common frame-based metrics: *F*1-score and AUC. *F*1-score is the harmonic mean of precision and recall and widely used in AU recognition. AUC quantifies the relation between true and false positives. For each method, we computed the average metrics on all AUs.

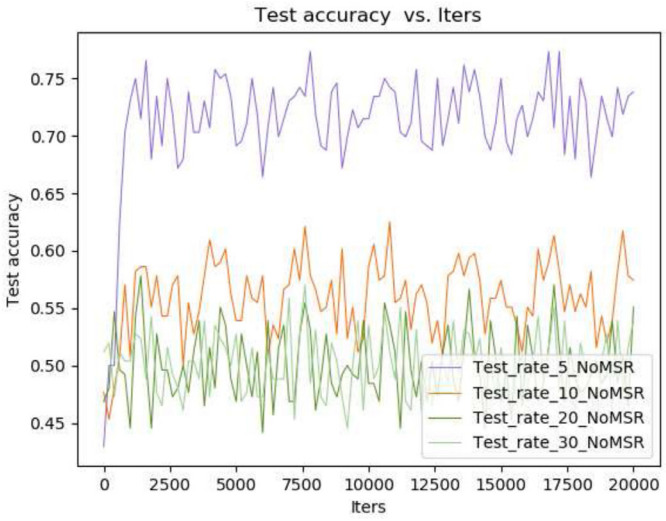
3.2. Experimental results

For the single label task for recognizing AU1, we vary proportion of the positive (with AU1) and negative (without AU1) samples by sampling from EmotioNet. In the process of manually sampling, we gradually increased the gap between the positive and negative samples, and proportions of negative and positive samples are 5, 10, 20, and 30 times. In order to ensure the accuracy of the final verification effect, the number of positive samples is the same as the number of negative samples in testing sets. Based on AlexNet, we use Eq. (7) as the loss function, where $n = 1$. Accuracies and losses are plotted in Fig. 5. For AlexNet without MSR, when the proportion of negative and positive samples is 5 times, the loss of the testing set converges during the training process, and also the recognition accuracy is more than 70%. However, when the proportion is 10 times, the network has been overfitting, and the loss is not as good as that in 5 times. The accuracy is also drastically reduced. The proportion is larger, and the performance is worse. For AlexNet with MSR, the convergence of losses are better, and the accuracy has been greatly improved. Accuracies in the proportion of both 5 times and 30 times archive are about 90%. Moreover, the curves of accuracy in varying proportions are almost overlapped. The curves of loss are also almost overlapped. That means that the proposed MSR is robust to the proportions of negative and positive samples.

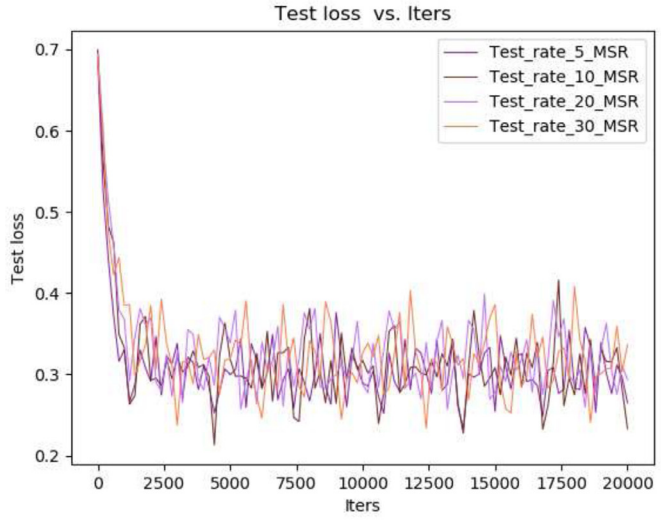
Based on the previous experiments, we compare the performance of MSR and Advanced MSR on multi-label learning. We still use AlexNet on EmotioNet database. The experiments are conducted on the database with AU labels of 3, 5, 8, and 11. Table 2 lists the proportions of negative and positive samples of each AU and the accuracies of MSR and Ad-MSR in each AU. Among the 11 AUs, the maximal proportion of positive and negative samples is 220.85 times, and the minimal proportion is 1.22 times. For every AU, the accuracy of Ad-MSR is higher than that of MSR. The biggest difference between the accuracies of MSR and Ad-MSR is 5.61%, and the smallest difference is 1.16%. Table 3 list accuracies of MSR and Ad-MSR of each AU when the numbers of AUs are 3, 5, 8, and 11. As shown in the table, the accuracy of each AU is decreased with the increase in the number of AUs. We think that this is caused by the fact that multi-label tasks share the network and the parameters have been shared. Unrelated tasks treat each other as noise and affect each other's convergence during training processes. When the number of AUs is increased from 3 to 11, we found that the accuracy of MSR was reduced by 1.36%, 0.2%, and



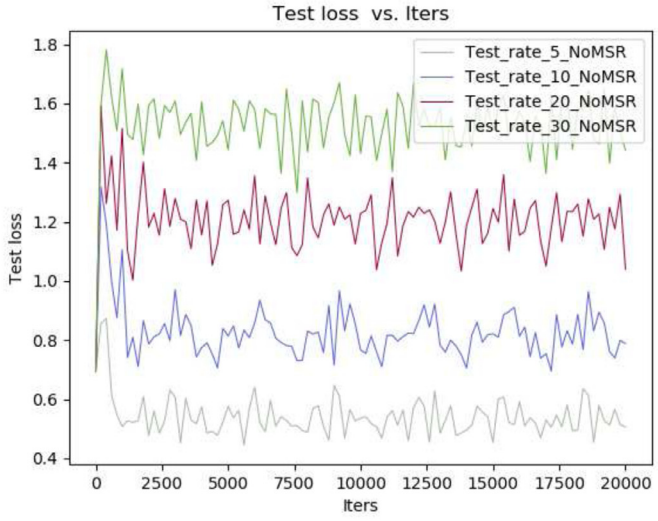
(a) Accuracies of AlexNet with MSR



(b) Accuracies of AlexNet without MSR



(c) Losses of AlexNet with MSR



(d) Losses of AlexNet without MSR

Fig. 5. Accuracies and losses of AlexNet with and without MSR in which the proportions of negative and positive samples are 5, 10, 20, and 30 times.

Table 2
The accuracies of MSR and Ad-MSR on EmotioNet database at 11 AUs case.

AU	Train samples N:P	MSR accuracy (%)	Ad-MSR accuracy (%)	AU	Train samples N:P	MSR accuracy (%)	Ad-MSR accuracy (%)
AU1	30.00	85.88	**88.18**	AU12	1.22	81.68	**83.45**
AU2	57.36	71.23	**73.01**	AU17	191.70	70.68	**74.73**
AU4	81.06	68.57	**73.91**	AU20	220.85	64.28	**66.60**
AU5	105.94	63.80	**67.72**	AU25	2.37	87.23	**89.63**
AU6	27.03	83.28	**84.44**	AU26	2.79	84.17	**86.19**
AU9	152.51	68.76	**74.37**				

3.05% by observing AU1, AU2, and AU4. In order to better the phenomenon, we optimized MSR and proposed Ad-MSR. When recognizing an AU, Ad-MSR can learn some knowledge from recognizing other AUs to improve the network's anti-noise and generalization ability. For single AU, the performance of Ad-MSR is more stable than that of MSR, with increasing the number of AUs. The accuracy was reduced by 0.18%, 0.28% and 1.39% by observing AU1, AU2, AU4. Compared with MSR, it is more stable. Besides, we also found that for the same AU, the accuracy of Ad-MSR is significantly improved compared with that of MSR. The experimental results show that for multi-label learning, the loss gain β can improve the accuracy for a single label by correcting its deviation from all labels, and use the relevance of other labels to improve network's generalization ability.

AUs is distributed in different parts of a whole face, and the changes between different AUs will affect each other. This can be effectively processed by deep CNN, which shows a powerful ability to hierarchically capture spatial structure information [36].

Table 3
The accuracies of MSR and Ad-MSR on EmotioNet database at 3, 5, 8, and 11 AUs cases.

AU	3 AUs		5 AUs		8 AUs		11 AUs	
	MSR	Ad-MSR	MSR	Ad-MSR	MSR	Ad-MSR	MSR	Ad-MSR
1	87.24	**88.36**	86.89	**88.33**	86.57	**88.20**	85.88	**88.18**
2	71.33	**73.28**	72.94	**73.04**	71.94	**72.99**	71.53	**73.00**
4	72.38	**74.78**	71.78	**74.14**	71.41	**74.25**	69.33	**73.39**
5			**67.69**	67.64	65.21	**68.42**	63.85	**67.72**
6			83.76	**84.15**	83.78	**84.30**	83.44	**84.44**
9					71.45	**75.47**	69.14	**74.73**
12					82.36	**83.96**	81.88	**83.45**
17					**73.53**	72.42	70.89	**74.04**
20							64.60	**66.60**
25							87.52	**89.62**
26							84.67	**86.19**
Mean	76.98	**78.81**	76.61	**77.46**	75.78	**77.50**	75.70	**78.31**

Table 4
The results come from the EmotioNet database. [**Bold**] denotes the best performance. **Bold** denotes the second best performance.

AU	*F*1-score (%)						AUC (%)						Train samples N:P	Test samples N:P	Validation samples N:P
	DRML	ResNet18	AlexNet	DSCMR	AlexNet +Ad-MSR	L-DSCMR	DRML	ResNet18	AlexNet	DSCMR	AlexNet +Ad-MSR	L-DSCMR			
1	2.10	21.67	2.02	**67.68**	60.71	**[89.84]**	18.04	38.94	19.98	**48.77**	47.21	**[49.77]**	30.00	1.00	1.00
2	3.40	3.93	3.49	**[52.62]**	50.34	**51.42**	17.51	36.29	18.94	**49.75**	49.46	**[49.83]**	57.26	5.34	5.28
4	4.02	7.00	4.81	**[37.37]**	25.54	**33.65**	12.04	17.81	8.31	**49.92**	49.16	**[53.78]**	81.06	14.62	13.43
5	5.19	7.62	5.26	**[34.36]**	27.31	**28.36**	6.55	0.99	4.82	49.77	**[50.05]**	**49.93**	105.94	14.44	14.32
6	4.78	32.46	5.00	**[47.51]**	41.76	**45.70**	28.43	38.62	27.79	49.27	**49.33**	**[50.46]**	27.03	14.14	13.43
9	1.33	10.09	5.01	**[27.15]**	**22.72**	22.18	9.39	13.99	6.69	**[51.45]**	49.80	**50.47**	152.51	26.75	26.39
12	**80.91**	72.92	66.30	76.21	76.42	**[83.09]**	**50.33**	49.23	50.28	**[51.13]**	50.64	50.02	1.22	1.06	1.00
17	9.58	17.24	2.01	**18.21**	6.93	**[19.22]**	4.42	50.08	16.97	**55.84**	52.76	**[63.37]**	191.70	74.45	101.53
20	0.99	10.00	3.19	**[27.53]**	24.67	**25.00**	6.04	11.24	5.32	**[50.10]**	49.75	**50.03**	220.85	18.50	19.29
25	**80.44**	70.28	54.72	72.05	66.42	**[82.52]**	47.80	49.50	48.28	**[52.48]**	**51.01**	50.31	2.37	2.68	2.40
26	**83.67**	81.20	58.61	79.62	72.44	**[88.32]**	43.46	48.66	44.10	**50.33**	49.86	**[50.55]**	2.79	0.78	0.76

CNN can be used to extract features of a certain AU, and more importantly, deep CNN can capture the spatial dependence between multi-areas. We designed a deep CNN so that a convolutional layer naturally captures the correlation between neighboring AUs, and a stack of convolutional layers can capture the correlation of even all AUs. In order to eliminate the negative impact of the deep network, the residual unit is used in the proposed DSCMR network structure.

In order to evaluate the performances of DSCMR and L-DSCMR, we compared it with the recent related work DRML on EmotioNet and CK+ databases. The same experimental setting was done on AlexNet and ResNet18. Ad-MSR is also used as the loss function in AlexNet denoted by AlexNet+Ad-MSR. And then we add to the AlexNet loss calculation.

Table 4 lists the *F*1-scores and AUCs of the above methods for 11 AUs. AU1 is used as a benchmark with the proportion of negative and positive samples in the training set its training set is 30 times, and the proportion in the testing set is 1. That is intended to evaluate the robustness of DSCMR with the increasing proportion of negative and positive samples in the training set. As is shown in Table 4, the performances of DMRL, ResNet-18, and AlexNet are not very good. DSCMR is more effective when facing the proportion of negative and positive samples is larger. Compared with AlexNet+Ad-MSR to DSCMR, it can be seen that the stacking of convolutional layers and the use of residual units are effective. The fact that the performance of ResNet-18 is better than that of AlexNet also illustrates that. The idea of the local convolution is applied to the DSCMR network. A local convolution component can capture local appearance changes, forcing the weights in each local patch to be updated independently. An identity addition component is introduced along with a “skip connection” from the input patch. That helps avoid vanishing gradient issues during training the network. The final result shows that the performance of L-DSCMR is good. The experimental result shows that some methods can extract valid information and have excellent performance in the verification set. However, the performance of L-DSCMR is the most significant. In all AUs, the proportion of negative and positive samples of AU20 is the largest. Comparing with methods without Ad-MSR, DSCMR performance is good, because Ad-MSR corrects not only this defect but it also the single AU recognition deviation.

Finally, we performed the same experimental setting on the CK+ database. Comparing with the EmotioNet database, the number of samples is smaller on the database. From Table 5, we can see that the proportion of negative and positive samples of each AU is not very large during the data acquisition process of CK+ relative to EmotionNet database. And each method performs well on the final recognition of AU. It shows that the equalization of the positive and negative samples has a significant influence in the recognition. So the proposed DSCMR have better performance on the multi-label sample imbalance problem. The final results also show that the performance of AlexNet+Ad-MSR is indeed better than that of AlexNet because the noise is removed and the generalization ability is improved. The deep CNN is used to extract high-level features and spatial information, and the mechanism of residual units is used to eliminate the problem of gradient disappearance. So on most AU recognition, the performances of ResNet-18 are better than those of AlexNet.However, the AUs of Tables 4 and 5 have some differences in the experimental results between the two databases. For example, for AU4, the sample rate in Emotion-Net is 81.06 and the CK+ is 1.93. All indicators of *F*1-score on CK+ are larger than Emotion-Net, which is caused by the imbalance of samples. When Ad-MSR add in network’s loss functions, the accuracy of recognition has been relatively improved. For AUC, the accuracy of the same network CK+ is higher which not have Ad-MSR. For networks joining Ad-MSR, the accuracy of the

Table 5
The results come from the CK+ database. [**Bold**] denotes the best performance. **Bold** denotes the second best performance.

AU	*F*1-score (%)						AUC (%)						Train Samples N:P	Test Samples N:P	Validation Samples N:P
	DRML	ResNet18	AlexNet	DSCMR	AlexNet +Ad-MSR	L-DSCMR	DRML	ResNet18	AlexNet	DSCMR	AlexNet +Ad-MSR	L-DSCMR			
1	50.00	64.15	34.09	**54.23**	45.56	**[64.25]**	49.99	49.85	49.48	49.87	**50.00**	**[50.05]**	2.16	2.93	2.54
2	40.00	**[70.96]**	43.47	**64.70**	47.05	58.57	**50.00**	49.48	48.63	49.96	**50.00**	**[50.53]**	2.27	5.94	6.31
4	**[72.28]**	42.85	52.63	61.11	54.32	**69.13**	**49.99**	45.67	45.58	**[50.00]**	48.94	49.98	1.93	2.57	2.16
5	**52.63**	60.00	35.08	42.24	47.36	**[66.67]**	**49.97**	31.16	30.50	**49.99**	49.85	47.22	4.54	8.07	3.87
6	57.78	60.00	54.54	**[68.18]**	57.14	**62.22**	**[55.43]**	27.52	45.80	53.16	49.80	**55.04**	2.98	5.94	5.88
7	24.39	26.67	35.48	**35.89**	27.02	**[47.05]**	**49.93**	41.45	49.57	**[50.00]**	49.90	49.08	3.22	5.21	6.31
9	**[54.54]**	36.36	21.27	**53.84**	35.29	51.53	31.37	28.80	20.87	**49.98**	49.94	**[62.26]**	5.12	12.11	15.71
12	**75.56**	51.28	60.00	**[80.00]**	63.41	**75.56**	**49.57**	5.22	35.43	**40.10**	25.58	9.32	3.61	3.21	3.5
15	57.14	24.00	53.52	**[82.92]**	61.53	**71.42**	**[50.10]**	30.61	27.43	50.02	49.92	**[51.03]**	6.10	4.61	4.08
17	81.92	60.24	68.81	**[89.74]**	67.39	**82.35**	49.62	49.58	48.40	**50.60**	36.80	**[54.56]**	1.88	2.27	1.85
20	58.06	21.05	48.80	**83.87**	66.67	**[84.84]**	**[49.97]**	11.80	7.29	28.18	**49.50**	30.44	6.10	5.94	9.63
23	41.67	15.00	31.11	**[75.00]**	50.09	**73.26**	**[50.01]**	32.94	42.27	**49.87**	49.71	43.06	9.44	10.8	6.315
24	**30.76**	20.00	17.00	**[36.36]**	22.22	26.67	27.01	28.77	31.12	**49.83**	**[50.03]**	44.62	6.88	15.85	18.5
25	76.42	79.33	78.94	**85.49**	77.06	**[89.06]**	49.71	**[51.09]**	**51.06**	50.00	41.23	49.98	1.29	1.80	2.06

two databases is closed, which also indicates that the sample imbalance is for identification. The accuracy rate has a great impact. For AU25, the positive and negative sample rete is around to 1, so the indicators are closed.

4. Conclusions

AU recognition is a samples imbalance problem and also is a multi-label learning problem. For the problem, we proposed a novel Multi-label Slope Rate (MSR) loss function and an Advanced-MSR (Ad-MSR) loss function in deep network architecture. In the architecture, a local convolution and residual units are used. The proposed two loss functions are evaluated on two expression databases.

Declarations of interest

None.

Acknowledgments

This paper is supported in part by grants from the National Natural Science Foundation of China (61772511) and in part by the Director Fund of National Engineering Laboratory for Public Security Risk Perception and Control by Big Data (18112403).

References

[1] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, J. Tao, Micro–expression recognition using color spaces, IEEE Trans. Image Process. 24 (12) (2015) 6034–6047.
[2] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System (The Manual onCD Rom), Network Information Research Corporation, Salt Lake City, 2002.
[3] M.S. Bartlett, G. Littlewort, C. Lainscsek, I.R. Fasel, J.R. Movellan, Machine learning methods for fully automatic recognition of facial expressions and facial actions, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, SMC, 2004, pp. 592–597. (1)
[4] J.F. Cohn, F. De la Torre, Automated face analysis for affective, in: The Oxford Handbook of Affective Computing, 2014, p. 131.
[5] S. Du, Y. Tao, A.M. Martinez, Compound facial expressions of emotion, Proc. Natl. Acad. Sci. 111 (15) (2014) E1454–E1462.
[6] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.
[7] L. Zhong, Q. Liu, P. Yang, J. Huang, D.N. Metaxas, Learning multiscale active facial patches for expression analysis, IEEE Trans. Cybern. 45 (8) (2015) 1499–1510.
[8] M. Valstar, M. Pantic, Fully automatic facial action unit detection and temporal analysis, in: Proceedings of the Computer Vision and Pattern Recognition Workshop, CVPRW'06, IEEE, 2006, p. 149.
[9] S. Lucey, A.B. Ashraf, J.F. Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, in: Face Recognition, InTech, 2007.
[10] Y. Zhu, F. De la Torre, J.F. Cohn, Y.-J. Zhang, Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior, IEEE Trans. Affect. Comput. 2 (2) (2011) 79–91.
[11] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I.R. Fasel, J.R. Movellan, et al., Automatic recognition of facial actions in spontaneous expressions., J. Multimed. 1 (6) (2006) 22–35.
[12] S. Eleftheriadis, O. Rudovic, M. Pantic, Multi-conditional latent variable model for joint facial action unit detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3792–3800.
[13] B. Jiang, M.F. Valstar, M. Pantic, Action unit detection using sparse appearance descriptors in space-time video volumes, in: Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), IEEE, 2011, pp. 314–321.
[14] W.-S. Chu, F. De la Torre, J.F. Cohn, Selective transfer machine for personalized facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3515–3522.
[15] X. Ding, W.-S. Chu, F. De la Torre, J.F. Cohn, Q. Wang, Facial action unit event detection by cascade of tasks, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2400–2407.
[16] Y.-I. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.
[17] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699.
[18] Z. Wang, Y. Li, S. Wang, Q. Ji, Capturing global semantic relationships for facial action unit recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3304–3311.
[19] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, Towards class-imbalance aware multi-label learning., in: Proceedings of the IJCAI, 2015, pp. 4041–4047.
[20] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[21] B. Wu, S. Lyu, B.-G. Hu, Q. Ji, Multi-label learning with missing labels for image annotation and facial action unit recognition, Pattern Recognit. 48 (7) (2015) 2279–2289.
[22] K. Zhao, W.-S. Chu, F. De la Torre, J.F. Cohn, H. Zhang, Joint patch and multi-label learning for facial action unit and holistic expression recognition, IEEE Trans. Image Process. 25 (8) (2016a) 3931–3946.
[23] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016b, pp. 3391–3399.
[24] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[25] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLSMOTE: approaching imbalanced multilabel learning through synthetic instance generation, Knowl.-Based Syst. 89 (2015) 385–397.
[26] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.
[27] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[28] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Web-scale training for face identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2746–2754.
[29] G.B. Huang, H. Lee, E. Learned-Miller, Learning hierarchical representations for face verification with convolutional deep belief networks, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2518–2525.

[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[31] N. Wang, D.-Y. Yeung, Learning a deep compact image representation for visual tracking, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 809–817.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[33] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015 arXiv preprint arXiv:1502.03167.
[34] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2010, pp. 94–101.
[35] C. Fabian Benitez-Quiroz, R. Srinivasan, A.M. Martinez, EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5562–5570.
[36] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324 .

Su-Jing Wang received the Master's degree from the Software College of Jilin University, Changchun, China, in 2007. He received the Ph.D. degree from the College of Computer Science and Technology of Jilin University in 2012. He was a postdoctoral researcher in Institute of Psychology, Chinese Academy of Sciences from 2012 to 2015. He is now an Associate Researcher in Institute of Psychology, Chinese Academy of Sciences. He has published more than 40 scientific papers. He is One of Ten Selectees of the Doctoral Consortium at International Joint Conference on Biometrics 2011. He was called as Chinese Hawkin by the Xinhua News Agency. His current research interests include pattern recognition, computer vision and machine learning. He serves as an associate editor of Neurocomputing (Elsevier).

Bo Lin is studying at the College of Software, Xi'an Jiaotong University, mainly studies data mining and image processing. He internships in the Key laboratory of behavior Sciences, Institute of Psychology, Chinese Academy of Sciences.

Yong Wang is currently pursuing the Masters degree in Software Engineering at Xi'an Jiaotong University. His research interests include machine learning, computer vision, and micro-expression.

Tong-qiang Yi is a master's degree candidate, was born in Shandong Province, China,in 1992. He is currently engaged in project research in school of electrical engineering, Xi'an Jiaotong University. Mainly engaged in power system fault diagnosis, as well as machine learning and pattern recognition related algorithm research.

Bochao Zou received Ph.D. degree from Beijing Institute of Technology in 2018, where he studied 3D display quality assessment, stereoscopic vision and visual attention. During that, he was a visiting graduate student in Visual Attention Lab at Harvard Medical School/Brigham and Women's Hospital from Sep. 2015 to Sep. 2017. He is currently a joint postdoctoral researcher between China Academy of Electronics and Information Technology and the State Key Laboratory of Virtual Reality in Beihang University, where his research interests mainly revolve around affective computing, especially facial expression recognition and remote-Photoplethysmograph (rPPG) for emotion analysis.

Xiang-wen Lyu received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics in 2015. He is a senior engineer in China Academic of Electronics and Information Technology. His interests include high performance computing, Computer Vision. He won one first Prize award of Science and Technology of Beijing Shijingshan district in 2017.