Separate Loss for Basic and Compound Facial Expression Recognition in the Wild

Yingjian Li HIT_LYJ@126.COM
Harbin Institute of Technology Shenzhen, Shenzhen, China.
Yao Lu YAOLU_1992@126.COM
Harbin Institute of Technology Shenzhen, Shenzhen, China.
Jinxing Li LIJINXING158@GMAIL.COM
The Chinese University of Hong Kong Shenzhen, Shenzhen, China.
University of Science and Technology of China, Hefei, China.
Guangming Lu[*] LUGUANGM@HIT.EDU.CN
Harbin Institute of Technology Shenzhen, Shenzhen, China.

Abstract

In the past few years, facial expression recognition has made great progress because of the development of convolutional neural networks. However, the features learned only using the softmax loss are not discriminative enough for highly accurate facial expression recognition in the wild, especially for the compound facial expression recognition. To enhance the discriminative power of the learned features, we propose the separate loss for both basic and compound facial expression recognition in the wild in this paper. Such loss maximizes intra-class similarity while minimizing the similarity between different classes. The qualitative and quantitative analysis shows that the features learned using such loss function are characterized by intra-class compactness and inter-class separation. Experiments are performed on two databases in the wild and the proposed method achieves state-of-the-art results on both basic and compound expressions. Furthermore, another two databases are used to perform cross database experiments to show the generalization ability of our method.

Keywords: Facial expression recognition, convolutional neural networks, loss functions

1. Introduction

Facial expression is an important way to convey emotions and plays an significant role in human's daily communication. According to Mehrabian and Ferris (1967), facial expression is more effective in emotion transmission than verbal and vocal signals. Because of the effectiveness and convenience, facial expression recognition has become an important method to detect emotions and has been widely used in human-computer interaction, health-care, driver safety and so on.

The first work (Suwa, 1978) on automatic facial expression recognition was published in 1978. Since then, many researchers have been working on efficient and accurate facial expression recognition methods (Lyons et al., 1998; Gu et al., 2012; Happy and Routray, 2015; Martinez et al., 2017). There are two kinds of environments of facial expression recognition.

[*] Corresponding author

The first one is in the lab and the other one is in the wild. In the first few years, researchers focused on facial expression recognition in the lab. In the lab-control environment, the facial images are usually posed and with fixed perspectives. Some databases have been proposed for facial expression recognition in the lab, such as CK+ (Kanade et al., 2000; Lucey et al., 2010), MMI (Pantic et al., 2005), and JAFFE (Lyons et al., 1998). Facial expression recognition in the wild is more challenging because the facial images are in arbitrary illuminations, poses and perspectives. From a practical point of view, it makes more sense to study facial expression recognition in the wild. As a result, more and more researchers devoted themselves to facial expression recognition in the wild. Some databases in the wild have also been collected, such as RAF-DB (Li et al., 2017) and AffectNet (Mollahosseini et al., 2019).

Most of the past works (Lyons et al., 1998; Pantic et al., 2005) focused on the six basic emotions, i.e. happiness, surprise, anger, sadness, fear and disgust, which were proposed by Ekman (1993). Such basic emotions are usually used together with the neutral emotion. However, Some facial expressions are complex and can not be simply classified as one basic emotion. As a result, Du et al. (2014) proposed compound emotions, which were combined by more than one basic emotion. Some samples of the 6 basic expressions with neutral expression and the 11 compound expressions are shown in Figure 1. Recognizing compound facial expressions in the wild is more challenging than basic expressions due to the subtle changes of emotions.

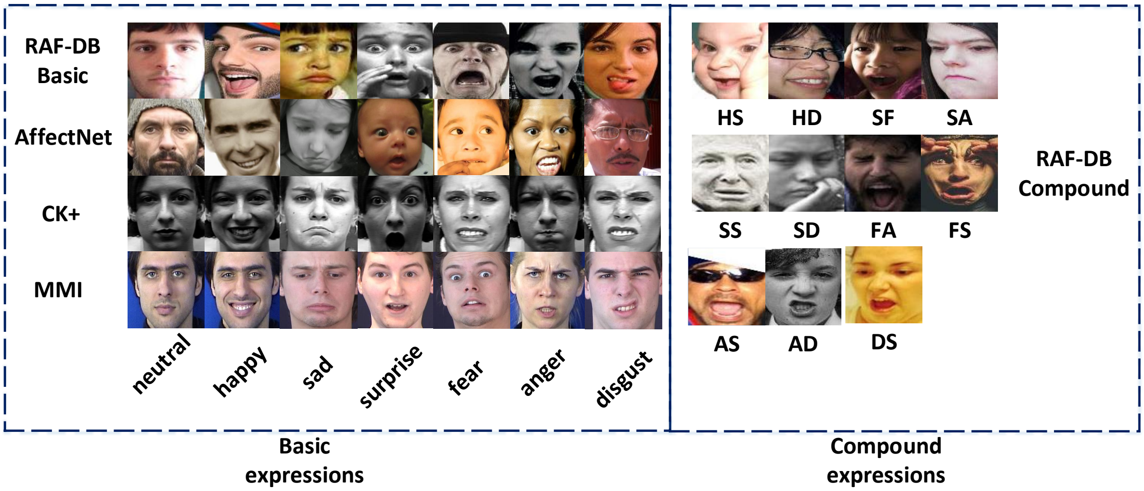


Figure 1: Basic and compound facial expression samples from different databases. HS, HD, SF, SA, SS, SD, FA, FS, AS, AD, DS are abbreviations of Happily Surprised, Happily Disgusted, Sadly Fearful, Sadly Angry, Sadly Surprised, Sadly Disgusted, Fearfully Angry, Fearfully Surprised, Angrily Surprised, Angrily Disgusted, Disgustedly Surprised, respectively.

Extracting discriminative features is important for efficient facial expression recognition algorithms. In the early years, researchers usually use manually designed features, such as

Local Binary Pattern (LBP) (Zhao and Pietikainen, 2007) and Local Directional Number (LDN) features (Rivera et al., 2013), for facial expression recognition. Such features are effective for lab-controlled facial images because the number of them is small and the image quality is good. However, due to the increase in the amount and complexity of facial images in the wild, manually designed features can not meet the requirements. Thanks to the development of deep learning techniques, this problem is partly solved. Deep learning techniques, especially the Convolutional Neural Network (CNN), automatically learn features from the training data other than use the manually designed features. The CNN is optimized by loss functions. Accordingly, the design of the loss function is crucial for the CNN to learn discriminative features. The softmax loss function is the most commonly used loss, which is very effective for uncomplicated classification tasks. However, the facial images in the wild are complex. The features learned only using the softmax loss are not discriminative enough for highly accurate facial expression recognition in the wild, especially for the compound facial expression recognition. To enhance the discriminative power of the learned features, we propose the separate loss function for facial expression recognition in the wild in this paper. Such separate loss is used together with the softmax loss to learn discriminative features. In addition to the wildly researched basic facial expressions, the rarely studied compound facial expressions are also considered. The main contributions of this paper are as follows.

(1) To learn discriminative features, the separate loss is proposed for both widely studied basic facial expressions and rarely studied compound facial expressions.

(2) The separate loss consists of the inter-class term and the intra-class term, which are used to maximize intra-class similarity and minimize the similarity between different classes, respectively. The values of such two terms are commensurate so that they can be added together without the need of any balance hyper parameters.

(3) The results of the qualitative and quantitative analysis indicate the effectiveness of the proposed loss function. We also perform both independent database and cross database experiments. The experiment results show that the proposed method achieves state-of-the-art performance and possesses good generalization ability.

The rest of the paper is organized as follows. Section 2 presents a review of related works. In Section 3, the loss function is introduced, including the overview of the recognition method, the learning process and the qualitative and quantitative analysis. Section 4 describes the experiments and result discussions. Section 5 concludes the paper.

2. Related Works

2.1. Facial expression recognition in the lab

Since only a part of facial patches are active for facial expressions (Zhong et al., 2012), some researchers tried to use some active patches for facial expression recognition. Zhong et al. (2012) defined some common and specific patches. The common patches were used to recognize all expressions while the specific patches were used for only a particular expression. Such patches were located by the proposed two-stage multi-task sparse learning framework. The size of patches was fixed in Zhong et al. (2012). Zhong et al. (2015) promoted this work using multi-scale active patches instead of the ones with fixed size. Inspired by Zhong et al. (2012), Happy and Routray (2015) selected 19 patches around eyes, nose and mouth for facial expression recognition. For each patch, LBP features were extracted to train Support Vector Machines (SVMs). For each pair of expressions, a classifier was trained and the top-4 patches were obtained to distinguish such pair of expressions. Sun et al. (2018a) chose three patches around facial organs and explored the proper size of these patches. Decision level fusion was used to obtain the final recognition result from these three patches.

Some researchers used the whole face instead of facial patches. Lopes et al. (2017) applied some pre-processing techniques such as rotation correction and intensity normalization to extract expression specific features from facial images and explored the presentation order of the samples during training. Zhang et al. (2017) proposed a deep evolutional spatial-temporal network for facial expression sequences recognition. Specially, a part-based hierarchical bidirectional recurrent neural network and a multi-signal convolutional neural network were proposed to learn the temporal information and spatial information. The fused spatial-temporal network achieved state-of-the-art performance for facial expression sequence recognition.

2.2. Facial expression recognition in the wild

Facial expression recognition in the wild is much challenging than that in the lab. Most of the methods are based on the deep learning technique, especially the CNN. Feature fusion and attention mechanisms are used by researchers to improve the performance.

Pons and Masip (2017) suggested that fusing some CNN for facial expression recognition can outperform individual CNN classifiers. In Pons and Masip (2017), two kinds of baseline architectures, i.e. CNN-4L and VGG-16 (Simonyan and Zisserman, 2015), were used. 72 and 64 individual models were trained based on different configurations using CNN-4L and VGG-16, resectively. A CNN model was proposed to fuse all the individual models. Experiments showed that the accuracy of the proposed method was significantly higher than other methods. Ji et al. (2019) proposed a network that consists of an intra-category common feature representation channel and an inter-category distinction feature representation channel. The two channels were weighted fused by a fully connected network. This fusion method achieved state-of-the-art performance.

Attention mechanisms help the CNN models to focus on the important features. Li et al. (2018) proposed a patch gated CNN for facial expression recognition. To learn the local features, 24 patches were selected according to facial landmarks. A patch gated unit with an attention net is designed to extract the features of each patch. The features of all patches are weighted fused for facial expression recognition. As a promotion, global-local-based CNN (Li et al., 2019) was proposed by adding the whole face area to patch gated CNN. Sun et al. (2018b) proposed a CNN model with visual attention to learn the region of interest for facial expression recognition. The experimental results showed the effectiveness of the attention model.

2.3. Loss functions

The loss function is very important for deep models. Wen et al. (2016) proposed the center loss for face recognition. Such loss can learn the center of deep features for each class and minimize the distances between the deep features and their corresponding class centers. Some researchers promoted the center loss in facial expression recognition since its effectiveness. Li et al. (2017) proposed the Locality Preserving (LP) loss. Such loss, together with the softmax loss, can preserve the locality closeness while maximizing the inter-class scatters. The center for LP loss is calculated by k nearest neighbors instead of all samples of one expression. In Li and Deng (2019), the LP loss was averaged by the number of training data. Cai et al. (2018) proposed island loss by adding the inter-class factor to the center loss. The island loss can reduce the intra-class variations and enlarge the inter-class differences. The values of the center loss and new added factor are not commensurate. As a result, a hyper parameter is added to balance them. Such hyper parameter is usually set empirically and needs lots of experiments to verify the value of it.

3. Facial Expression Recognition Using Separate Loss

3.1. Overview of the recognition method

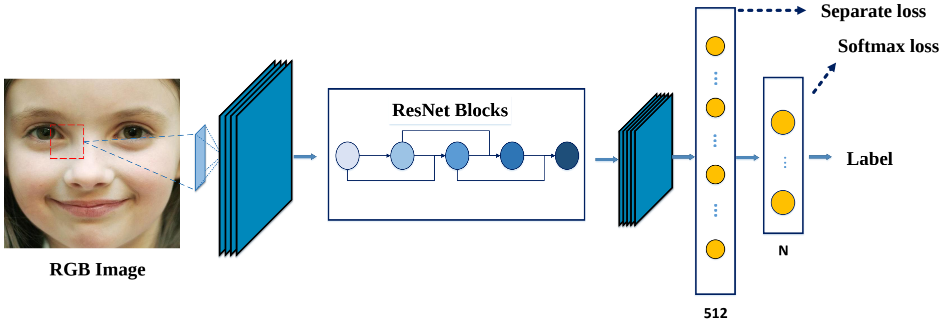


Figure 2: The structure of Resnet18 for facial expression recognition. N is the number of classes of expressions. The softmax loss and separate loss are calculated from the output layer and the layer before the last layer, respectively.

We use the ResNet18 (He et al., 2016) to extract features for facial expression recognition, the structure of which is shown in Figure 2. The CNN takes RGB images as input data and outputs the predicted label of facial expression. There are 7 units and 11 units in the output layer for the basic expressions and the compound expressions, separately. In this paper, both the proposed separate loss and the softmax loss are used to optimize the CNN model. The softmax loss is used to minimize the error between output labels and target labels, whcih is calculated from the last layer. The separate loss is used to guide the features and is calculated from the last pooling layer, i.e. the layer before the last layer.

3.2. Separate loss

Some notations are used in this paper. $\mathcal{R}^d$ denotes the set of d-dimension column vectors. $\boldsymbol{x}_i \in \mathcal{R}^d$ is the feature of the ith sample and y_i is the label. $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ is the set of deep features and $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ is the set of labels. Let $\boldsymbol{c}_i$ be the center of features with label y_i. $\boldsymbol{x}_i^{\mathrm{T}}$ is the transform of $\boldsymbol{x}_i$. L_s and L_{sep} are the averaged softmax loss and separate loss, respectively. $\|\boldsymbol{x}_i\|_2$ is the L_2 norm of $\boldsymbol{x}_i$.

The separate loss consists of two parts, i.e. the intra-class loss and the inter-class loss. Both of the them are based on normalized cosine similarity described by Equation (1).

$$Sim(\boldsymbol{a}_1, \boldsymbol{a}_2) = \frac{1}{2}\left(\frac{\boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_2}{\|\boldsymbol{a}_1\|_2 \|\boldsymbol{a}_2\|_2} + 1\right), \tag{1}$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathcal{R}^d$ are two nonzero vectors and $Sim(\boldsymbol{a}_1, \boldsymbol{a}_2) \in [0, 1]$.

The normalized intra-class loss and inter-class loss are written as Equations (2) and (3), respectively.

$$L_{intra} = 1 - \frac{1}{2m}\sum_{i=1}^{m}\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}_{y_i}}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{c}_{y_i}\|_2} + 1\right), \tag{2}$$

$$L_{inter} = \frac{1}{2|S|(|S|-1)}\sum_{k\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S},\\ j\neq k}}\left(\frac{\boldsymbol{c}_k^{\mathrm{T}}\boldsymbol{c}_j}{\|\boldsymbol{c}_k\|_2 \|\boldsymbol{c}_j\|_2} + 1\right), \tag{3}$$

where m is the batch size, $\mathcal{S} = \{1, 2, ..., s\}$ is the set of classes and $|\mathcal{S}|$ is the size of it. In the training stage, L_{intra} and L_{intra} are minimized so that the features in the same class become more similar and those in different classes become dissimilar.

From Equations (1), (2) and (3), we obtain that $L_{intra} \in [0, 1]$ and $L_{inter} \in [0, 1]$, which means that the two loss functions are commensurate. As a result, we do not need a hyper parameter to balance the two loss functions when they are added together. The proposed separate loss is written as

$$L_{sep} = L_{intra} + L_{inter} \tag{4}$$

The back propagation and the center update are two key steps in the training stage using separate loss. In the back propagation process, the gradient of the separate loss with respect to $\boldsymbol{x}_i$ is as follow.

$$\frac{\partial L_{sep}}{\partial \boldsymbol{x}_i} = -\frac{1}{2m}\left(\frac{1}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{c}_{y_i}\|_2}\boldsymbol{c}_{y_i} - \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}_{y_i}}{\|\boldsymbol{x}_i\|_2^3 \|\boldsymbol{c}_{y_i}\|_2}\boldsymbol{x}_i\right) \tag{5}$$

The centers are updated according to both L_{intra} and L_{inter}. Let $\Delta\boldsymbol{c}_{j,intra}$ and $\Delta\boldsymbol{c}_{j,inter}$ be the increments of $\boldsymbol{c}_j$ caused by L_{intra} and L_{inter}, respectively. $\Delta\boldsymbol{c}_{j,intra}$ and $\Delta\boldsymbol{c}_{j,inter}$ are computed using Equations (6) and (7), respectively.

$$\Delta\boldsymbol{c}_{j,intra} = -\frac{1}{2m\left(1+\sum_{i=1}^{m}\delta(y_i, j)\right)}\sum_{i=1}^{m}\delta(y_i, j)\left(\frac{1}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{c}_j\|_2}\boldsymbol{x}_i - \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}_j}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{c}_j\|_2^3}\boldsymbol{c}_j\right) \tag{6}$$

$$\Delta\boldsymbol{c}_{j,inter} = \frac{1}{2|S|(|S|-1)}\sum_{k\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S},\\ j\neq k}}\left(\frac{1}{\|\boldsymbol{c}_k\|_2 \|\boldsymbol{c}_j\|_2}\boldsymbol{c}_k - \frac{\boldsymbol{c}_k^{\mathrm{T}}\boldsymbol{c}_j}{\|\boldsymbol{c}_k\|_2 \|\boldsymbol{c}_j\|_2^3}\boldsymbol{c}_j\right) \tag{7}$$

where $\delta(y_i, j) = 1$ if $y_i = j$ and $\delta(y_i, j) = 0$ if $y_i \neq j$. Then, we can use Equation (8) to update the centers.

$$\boldsymbol{c}_j^{t+1} = \boldsymbol{c}_j^t - \alpha(\Delta\boldsymbol{c}_{j,intra}^t + \Delta\boldsymbol{c}_{j,inter}^t) \tag{8}$$

where α is a learning rate of the centers.

In this paper, the separate loss is used together with the averaged softmax loss. A weight λ is applied to balance such two loss functions. Equation (9) shows the overall loss function. The training steps are introduced in Algorithm 1 in detail.

$$L = L_s + \lambda L_{sep} \tag{9}$$

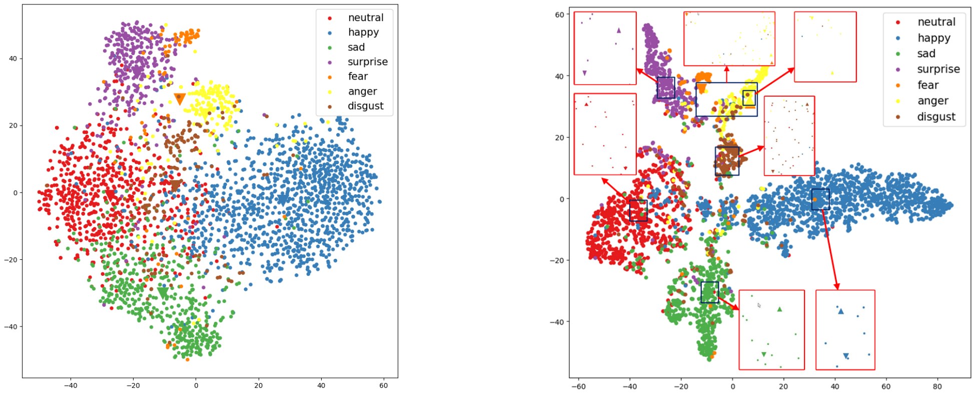
Algorithm 1: Training steps using the separate loss

```
Input: The training set $\mathcal{X}$, the target set $\mathcal{Y}$, the learning rate $\eta$ and $\alpha$ for CNN and
       centers, the batch size $m$, maximum iteration steps $T$, the hyper parameter $\lambda$.
Output: parameters $W$ of the network, parameters $\theta$ of the softmax layer.
for $t$=1 to $T$:
    calculate the overall loss using Equation (9): $L = L_s + \lambda L_{sep}$,
    update the parameters in the softmax layer: $\theta^{t+1} = \theta^t - \eta \frac{\partial L_s^t}{\partial \theta^t}$,
    update the centers using Equation (8): $\boldsymbol{c}_j^{t+1} = \boldsymbol{c}_j^t - \alpha(\Delta \boldsymbol{c}_{j,intra}^t + \Delta \boldsymbol{c}_{j,inter}^t)$,
    calculate the back propagation error: $\frac{\partial L^t}{\partial \boldsymbol{x}_i^t} = \frac{\partial L_s^t}{\partial \boldsymbol{x}_i^t} + \lambda \frac{\partial L_{sep}^t}{\partial \boldsymbol{x}_i^t}$,
    based on the back propagation error, update the parameters of the network
    according to the chain rule: $W^t + 1$=$W^t$-$\eta \frac{\partial L^t}{\partial W^t}$=$W^t$-$\eta \frac{\partial L^t}{\partial \boldsymbol{x}_i^t} \frac{\partial \boldsymbol{x}_i^t}{\partial W^t}$.
end for
```

3.3. Analysis and discussion

Qualitative analysis. To evaluate the function of the separate loss, we visualize the features learned on RAF-DB using the t-sne (Der Maaten and Hinton, 2008) algorithm, as shown in Figure 3. Figures 3(*a*) and 3(*b*) show the features learned using the softmax loss and the separate loss, respectively. Compared with the features in Figure 3(*a*), those in Figure 3(*b*) are more discriminative and the features in the same class are more close.

Quantitative analysis. To show the effectiveness of the proposed loss accurately, we calculate the inter-class and intra-class similarity, which is shown in Table 1. In Table 1, NE, HA, SA, SU, FE, AN and DI are short for neutral, happy, sad, surprise, fear, anger and disgust, respectively. The intra-class similarity of expression y_i is the average value of the cosine similarity between the features labeled y_i and their centers $\boldsymbol{c}_i$. The inter-class similarity is the average value of cosine similarity between the centers of different classes. Compared with the intra-class similarity of each expression using the softmax loss, the intra-class similarity using the separate loss is bigger in most cases, as well as the average intra-class similarity. These results show that the features of the same class are more concentrated using the separate loss. The average inter-class similarity using the separate loss is much lower than that using the softmax loss, which indicates that the separate loss is helpful to distinguish the features from different classes. We also compute the true centers by averaging the representations of each class and output the learned centers. The similarities between the true and learned centers are computed in each epoch. In the first epoch, the similarities for the seven classes are 0.72, 0.78, 0.63, 0.60, 0.48, 0.51 and 0.57, respectively, which enjoy an increase with the rise of iterations. Specially, the similarities



(a) softmax loss

(b) softmax loss and separate loss

Figure 3: Visualization of the features learned on the RAF-DB database. The △ and ▽ with different colors are the learned centers and true centers, respectively.

Table 1: Inter-class and intra-class similarity

Loss	NE	HA	SA	SU	FE	AN	DI	average intra-classs	average inter-classs
L_s	0.92	0.93	0.92	0.93	0.92	0.93	0.91	0.92	0.89
L_s+L_{sep}	0.96	0.97	0.95	0.95	0.89	0.92	0.91	0.94	0.75

reach 0.99, 1.00, 0.99, 0.96, 0.66, 0.91 and 0.86, respectively, after arriving at the maximized epoch. The result shows that the learned centers can automatically approximate the true ones, as shown in Figure 3(b).

Discussion. Although the separate loss shares the main idea with Linear Discriminant Analysis (LDA), i.e., intra-class compactness and inter-class separation, they are different. LDA is usually used to classify samples or reduction the dimensionality of features. However, the separate loss is used as a object function to optimize deep models. Moreover, LDA optimizes a function defined by intra-class and inter-class scatter matrices, while our loss optimizes intra-class and inter-class similarities directly.

4. Experiments and Results

4.1. Databases

We use four databases to evaluate the performance of the proposed method. Some samples from these databases are shown in Figure 1. The RAF-DB (Li et al., 2017) database contains about 30,000 facial expression images with basic emotions from the real word. In this paper, 12,271 images are used for training and 3,068 images are used for testing. In addition to the basic expression images, there are 3162 training images and 792 testing images with compound emotions in the RAF-DB database. These images are also used in this paper.

The AffectNet (Mollahosseini et al., 2019) database is the largest facial expression database in the wild, which contains more than 1,000,000 facial images collected from the internet. We use the manually labeled images with basic expressions in this paper, including 283,901 training images and 3,500 testing images.

There are 327 labeled facial expression sequences in the CK+ (Kanade et al., 2000) database. In this paper, we eliminate 18 sequences which are labeled as contempt because such expression is not widely used. For each of the rest 309 sequences, we use the last frame as the target expression and select the first frame of each sequence as the neutral expression.

The MMI (Pantic et al., 2005) database consists of 205 labeled sequences with frontal faces. In each sequence, the middle frame is with peak expression while the first and last frames are with a neutral expression. We extract the first frame as the neutral expression and the middle frame as the labeled expression.

4.2. Experiment settings

In this paper, the pre-trained weights on ImageNet are used to initialize the ResNet18. We use the stochastic gradient descent algorithm to train the model. The batch size and initial learning rate are set as 100 and 0.01 respectively. Because the numbers of training images are different in the RAF-DB and AffectNet databases, some hyper parameters are set differently for them. For AffectNet, the max iteration epoch is set as 20 and the learning rate reduces to one tenth every 5 epoches. For RAF-DB, we set the max iteration epoch as 60 and the learning rate reduces to one tenth every 20 epoches. We conduct experiments by setting lambda to values ranging from 0.1 to 1.1 with a step of 0.1. The accuracies are 85.79%, 85.82%, 86.05%, 85.92%, 85.98%, 85.85%, 86.21%, 86.15%, 86.38%, 86.18% and 86.08%, respectively, as shown in Figure 4. It shows that lambda locating in [0.7,1] can ensure to obtain a satisfactory performance and the best result is obtained by $\lambda = 0.9$. Therefore, $\lambda = 0.9$ is used for the rest experiments. All the experiments are performed on a server with two TITAN Xp GPUs. It takes about 31 minutes to obtain the optimized model on RAF-DB and 4.3 hours on AffectNet.

4.3. Independent database experiments

We use the RAF-DB (basic and compound) database and the AffectNet database to evaluate the proposed loss function and the experiment results are shown in Table 2. For the basic expressions, the proposed separate loss achieves an accuracy of 86.38% on the RAF-DB database and 58.89% on the AffectNet database. For the compound expressions, the accuracy is 58.84% using the separate loss.

Table 2 also shows the comparison between the existing methods and our method. Li and Deng (2019) proposed the locality preserving loss to make the features more close in the same class. Li et al. (2018, 2019) used both local and global feature to improve the accuracy. Compared with these methods, our method achieves better performances for both basic and compound expressions. This is because that both the inter-class and intra-class similarities are considered in the separate loss. Therefore, the learned features are characterized by

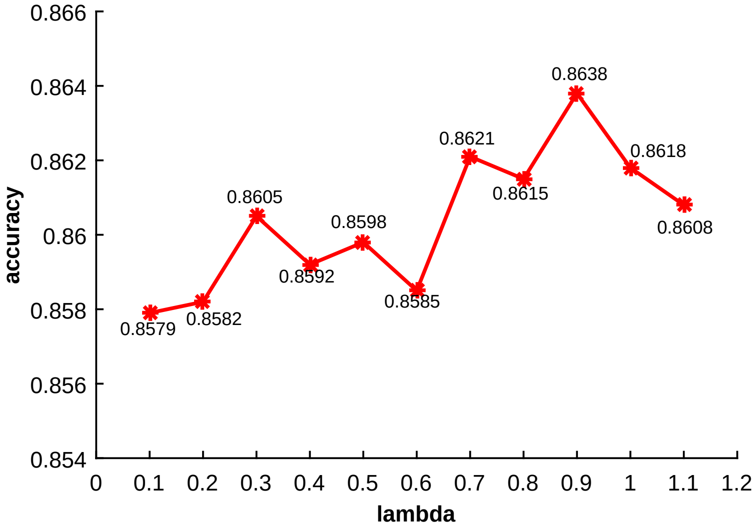


Figure 4: The accuracies on RAF-DB using different values of λ.

intra-class compactness and inter-class separation, which is conducive to facial expression recognition.

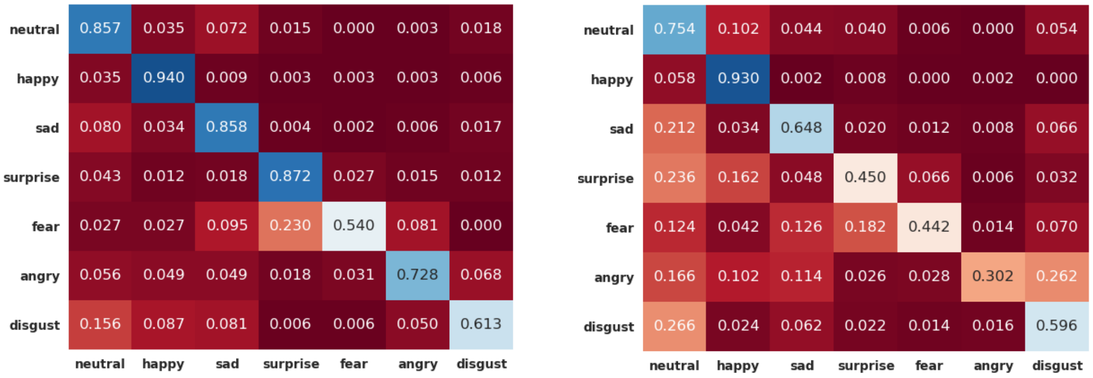
Compared with the results using the softmax loss, those using the separate loss are better. For the basic expressions, the accuracies increase by 0.72% and 0.52% on RAF-DB and AffectNet. The accuracy increases by 1.9% on the compound expressions compared with that of the baseline. These results also indicate the effectiveness of the proposed separate loss.

Table 2: Accuracy on independent databases (%)

method	**RAF-DB compound**	**RAF-DB basic**	**Affectnet**
DLP-CNN (Li and Deng, 2019)	57.95	84.13	/
PG-CNN (Li et al., 2018)	/	83.27	55.33
gACNN (Li et al., 2019)	/	85.07	58.78
ResNet18+L_s(baseline)	56.94	85.66	58.37
ResNet18+L_s+L_{sep}	**58.84**	**86.38**	**58.89**

The confusion matrix is used to show the performance on each expression. The confusion matrices for basic and compound expressions are shown in Figures 5 and 6, respectively. The images in RAF-DB and AffectNet are imbalanced for different facial expressions. From Figures 5(a) and 5(b), we find the expressions with more training images have higher accuracy. For example, in the RAF-DB database, the number of training images of happy expression is the largest and that of fear expression is the smallest. Correspondingly, the happy expression has the highest accuracy and the fear expression has the lowest accuracy. This rule is also right for the AffectNet database. The compound expressions are more difficult to be recognized than the basic ones due to the complexity. Therefore, the accuracies are low for some compound expressions, as shown in Figure 6. The disgustedly surprised, sadly fearful and sadly surprised expressions are the three ones that have the lowest accuracies. Meanwhile, the numbers of training images of such three expressions are the smallest among all expressions. In addition to the complexity of the expressions, the lack of training data also affects the accuracies of such three expressions.

To explore the influence of the number of elements of the deep features, we add a fully connected layer at the end of ResNet18, so that the number of feature elements can be adaptively changed. We set the numbers to 64, 128, 256, 512, 768 and 1024, respectively. The accuracies on RAF-DB are 85.89%, 86.08%, 86.21%, 86.41%, 86.25% and 86.31%, respectively, demonstrating the robustness of the proposed method on the feature element number.

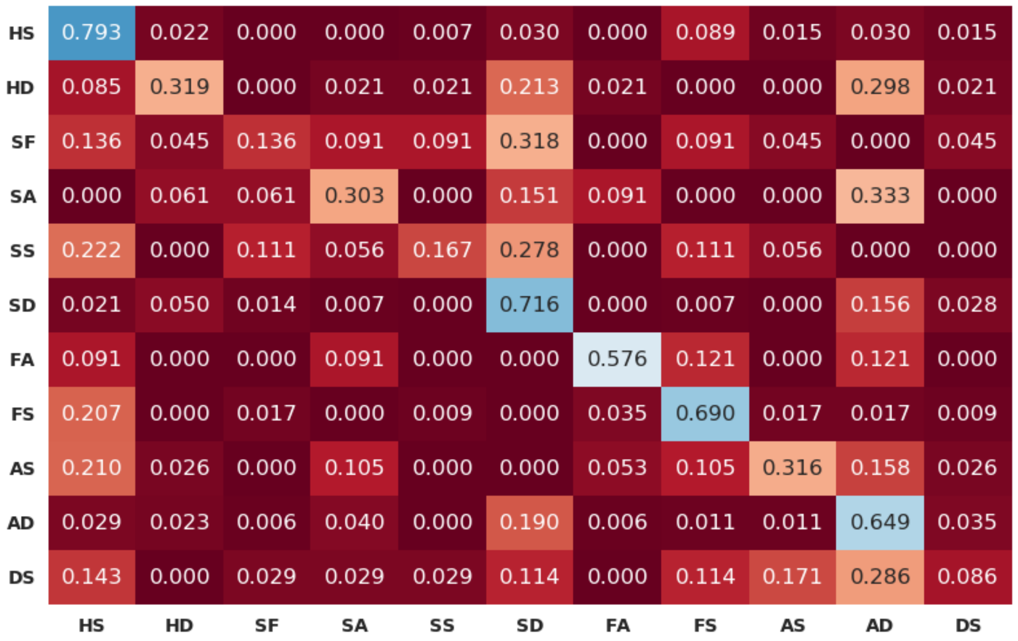


(a) RAF-DB basic (b) AffectNet

Figure 5: Confusion matrices on different databases.

4.4. Cross database experiments

Cross database experiments on the CK+ and MMI databases are carried out to evaluate the generalization ability of the proposed method, the result of which is shown in Table 3. In the cross database experiments, the models are trained on AffectNet or RAF-DB and tested on CK+ and MMI without fine-tuning. Table 3 shows that the models trained on AffectNet achieve better performance than those trained on RAF-DB, which is caused by the number of training images. Compared with the gACNN (Li et al., 2019), our method achieves higher accuracy in most cases. This result indicates the good generalization ability of our method. The accuracy of the model which is trained on AffectNet and tested on CK+ is much lower than that of Li et al. (2019). This may be caused by the usage of different data. Li et al. (2019) did not state which part of the CK+ database was used. The data of CK+ used in this paper is the same as Li and Deng (2019).



	HS	HD	SF	SA	SS	SD	FA	FS	AS	AD	DS
HS	0.793	0.022	0.000	0.000	0.007	0.030	0.000	0.089	0.015	0.030	0.015
HD	0.085	0.319	0.000	0.021	0.021	0.213	0.021	0.000	0.000	0.298	0.021
SF	0.136	0.045	0.136	0.091	0.091	0.318	0.000	0.091	0.045	0.000	0.045
SA	0.000	0.061	0.061	0.303	0.000	0.151	0.091	0.000	0.000	0.333	0.000
SS	0.222	0.000	0.111	0.056	0.167	0.278	0.000	0.111	0.056	0.000	0.000
SD	0.021	0.050	0.014	0.007	0.000	0.716	0.000	0.007	0.000	0.156	0.028
FA	0.091	0.000	0.000	0.091	0.000	0.000	0.576	0.121	0.000	0.121	0.000
FS	0.207	0.000	0.017	0.000	0.009	0.000	0.035	0.690	0.017	0.017	0.009
AS	0.210	0.026	0.000	0.105	0.000	0.000	0.053	0.105	0.316	0.158	0.026
AD	0.029	0.023	0.006	0.040	0.000	0.190	0.006	0.011	0.011	0.649	0.035
DS	0.143	0.000	0.029	0.029	0.029	0.114	0.000	0.114	0.171	0.286	0.086

Figure 6: Confusion matrix on the compound database. HS, HD, SF, SA, SS, SD, FA, FS, AS, AD, DS are abbreviations of Happily Surprised, Happily Disgusted, Sadly Fearful, Sadly Angry, Sadly Surprised, Sadly Disgusted, Fearfully Angry, Fearfully Surprised, Angrily Surprised, Angrily Disgusted, Disgustedly Surprised, respectively.

Table 3: Accuracy of cross database experiments (%)

method	**database (train)**	**CK+ (test)**	**MMI(test)**
gACNN (Li et al., 2019)	AffectNet	**91.64**	70.37
gACNN (Li et al., 2019)	RAF-DB	81.07	59.51
ResNet18+L_s+L_{sep}	AffectNet	82.04	**71.39**
ResNet18+L_s+L_{sep}	RAF-DB	**81.38**	**69.71**

5. Conclusions

In this paper, the separate loss is proposed to guide the CNN to learn discriminative features for both basic and compound facial expression recognition in the wild. The separate loss consists of two terms. The first term is used to maximizes intra-class similarity and the second one is used to minimize the inter-class similarity. These two terms are commensurate so that we do not need an additional hyper parameter to balance them. We perform both qualitative and quantitative analysis on the learned features using the separate loss to show the effectiveness of such loss. The result shows that the features learned using such loss function are characterized by intra-class compactness and inter-class separation. We carry out experiments on two databases in the wild, i.e. RAF-DB and AffectNet, and the proposed loss function achieves state-of-the-art results. The results of cross database experiments on CK+ and MMI show the good generalization ability of our method.

Acknowledgments

This work is supported by the Shenzhen Fundamental Research fund (JCYJ20180306172023 949, JCYJ20170811155442454), Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China. This work is also supported by China Postdoctoral Science Foundation (2019TQ0316), NSFC fund (61906162), Shenzhen Research Institute of Big Data, and Shenzhen Institute of Artificial Intelligence and Robotics.

References

Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 302–309. IEEE, 2018.

Laurens Van Der Maaten and Geoffrey E Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (15):201322355, 2014.

Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993.

Wenfei Gu, Cheng Xiang, YV Venkatesh, Dong Huang, and Hai Lin. Facial expression recognition using radial encoding of local gabor features and classifier synthesis. *Pattern Recognition*, 45(1):80–91, 2012.

S L Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, 6(1):1–12, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing*, 333:231–239, 2019.

Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE, 2000.

Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2852–2861, 2017.

Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Patch-gated cnn for occlusion-aware facial expression recognition. In *International Conference on Pattern Recognition*, pages 2209–2214, 2018.

Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2019.

André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE, 1998.

Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: a survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2017.

Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3):248–252, 1967.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, pages 5–pp. IEEE, 2005.

Gerard Pons and David Masip. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, 9(3): 343–350, 2017.

Adin Ramirez Rivera, Jorge A Rojas Castillo, and Oksam Chae. Local directional number pattern for face analysis: Face and expression recognition. *IEEE Transactions on Image Processing*, 22(5):1740–1752, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *international conference on learning representations*, 2015.

Ai Sun, Yingjian Li, Yuehmin Huang, Qiong Li, and Guangming Lu. Facial expression recognition using optimized active regions. *Human-centric Computing and Information Sciences*, 8(1):33, 2018a.

Wenyun Sun, Haitao Zhao, and Zhong Jin. A visual attention based roi detection method for facial expression recognition. *Neurocomputing*, 296:12–22, 2018b.

Motoi Suwa. A preliminary note on pattern recognition of human emotional expression. In *Proceedings of International Joint Conference on Pattern Recognition*, pages 408–410, 1978.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, 2017.

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 915–928, 2007.

Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE, 2012.

Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics*, 45(8):1499–1510, 2015.