Journal of Electronic Imaging

JElectronicImaging.org

Adaptive metric learning with deep neural networks for video-based facial expression recognition

Xiaofeng Liu
Yubin Ge
Chao Yang
Ping Jia

Xiaofeng Liu, Yubin Ge, Chao Yang, Ping Jia, "Adaptive metric learning with deep neural networks for video-based facial expression recognition," *J. Electron. Imaging* **27**(1), 013022 (2018), doi: 10.1117/1.JEI.27.1.013022.

Journal of Electronic Imaging 27(1), 013022 (Jan/Feb 2018)

Adaptive metric learning with deep neural networks for video-based facial expression recognition

Xiaofeng Liu,[a,b,c] Yubin Ge,[c,d,*] Chao Yang,[e] and Ping Jia[a,b]
[a]Chinese Academy of Sciences, Changchun Institute of Optics, Fine Mechanics and Physics, Changchun, China
[b]University of Chinese Academy of Sciences, Shijingshan District, Beijing, China
[c]Carnegie Mellon University, Department of Electrical and Computer Engineering, Pittsburgh, Pennsylvania, United States
[d]University of Pittsburgh, School of Computing and Information, Pittsburgh, Pennsylvania, United States
[e]University of South California, Institute for Creative Technologies, Los Angeles, California, United States

Abstract. Video-based facial expression recognition has become increasingly important for plenty of applications in the real world. Despite that numerous efforts have been made for the single sequence, how to balance the complex distribution of intra- and interclass variations well between sequences has remained a great difficulty in this area. We propose the adaptive $(N + M)$-tuplet clusters loss function and optimize it with the softmax loss simultaneously in the training phrase. The variations introduced by personal attributes are alleviated using the similarity measurements of multiple samples in the feature space with many fewer comparison times as conventional deep metric learning approaches, which enables the metric calculations for large data applications (e.g., videos). Both the spatial and temporal relations are well explored by a unified framework that consists of an Inception-ResNet network with long short term memory and the two fully connected layer branches structure. Our proposed method has been evaluated with three well-known databases, and the experimental results show that our method outperforms many state-of-the-art approaches. *© 2018 SPIE and IS&T* [DOI: 10.1117/1.JEI.27.1.013022]

Keywords: metric learning; video-based; facial expression recognition; deep learning.

Paper 170663 received Aug. 8, 2017; accepted for publication Jan. 23, 2018; published online Feb. 19, 2018.

1 Introduction

As the most expressive nonverbal channels of internal emotions, facial expression recognition (FER) plays a vital role in human–machine interaction systems. For the wide-spreading employment of video-based digital entertainment and health care etc., video-based FER, which is to classify an expressive human face image sequence to one of the six expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise), also has been a hot topic for decades.

Compared to static face images, videos usually provide us more information, such as multiview and temporal information. However, the FER problem becomes significantly challenging when it comes to videos and presents undesirable performances. What's more, the quality of video frames seems to be unstable, and the faces in videos always present much richer variations since the video acquisition has fewer constraints.[1] In addition, the common dynamic pattern in facial expression also exerts influences in video-based FER. Such a pattern can be divided into three phrases: onset, peak, and offset, where the onset is the beginning of the expression, the peak (aka apex) represents the maximum intensity of the expression, and the offset describes the moment when the expression vanishes. In most cases, the change of a facial expression from the onset to the offset tends to be very fast, which makes the process of video-based FER pretty challenging.[2] Also, the subjects in videos are often mobile, and this definitely brings serious motion blur and out-of-focus blur.

Despite the great efforts that have been made, FER, even for still images, remains a challenge for illumination and pose variations as well as intersubject variations (i.e., identity-specific attributes).[3] Since expressions may only involve subtle facial muscle movements, the extracted expression-related information from different classes can be dominated by the sharp-contrast identity-specific geometric or appearance features, which are not useful for FER. As shown in Fig. 1, example x_1 and x_2 are happy faces whereas x_3 is a sad face. We set $f(x_i)$ as the image representations given the corresponding extracted features. Specifically, we expect that every two faces labeled as different expressions are farther away from each other, while every two faces that share the same expression label should be closer to each other in the feature space, i.e., the distance D_2 is larger than D_1, which is the distance between examples x_1 and x_2, just as shown in Fig. 1(b). Unfortunately, the learned results of facial expression representations usually contain erroneous identity information as Fig. 1(a). Because the interidentity variations tend to be large in real cases, D_1 usually has a larger value than D_2.

In recent years, deep metric learning appeals to many research interests in the regime of image recognition due to its great learning ability to the general concept of distance metrics and its outstanding compatibility with effectively infering neighbors in the learned metric space. However, it normally suffers the problem of slow convergence and is easy to be overfitting. The softmax is computational efficient for classification, but it does not encourage the large margin separation of representations. A potential solution to the problems mentioned above is to combine the deep metric learning scheme and the softmax loss within a convolutional neural network (CNN) framework. For the deep metric

*Address all correspondence to: Yubin Ge, E-mail: yub37@pitt.edu

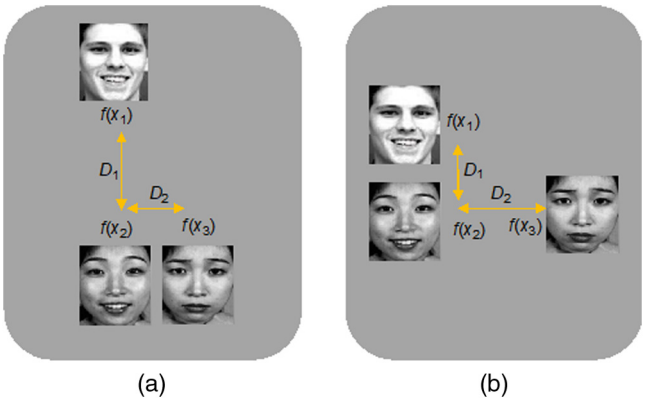

Fig. 1 Illustration of representations in feature space learned by (a) existing methods, and (b) the proposed method.

learning methods, the basic idea behind the well-known triplet loss function[4] is to force one negative example farther to the anchor example than one positive example to the anchor example with a fixed gap τ. Therefore, the triplet loss neglects the negative examples from the rest of the classes in one certain iteration. What's more, there are some special situations in which the triplet loss function is very likely to judge mistakenly when choosing an inappropriate anchor, as shown in Fig. 2(a). Inspired by the ideas from the $(N+1)$-tuplet loss[5] and coupled clusters loss (CCL),[6] we then design a $(N+M)$-tuplet clusters loss function, which incorporates a negative set that consists of N examples and a positive set with M examples in a mini-batch. A reference distance T is introduced to force the negative examples to be away from the center of positive examples and make the positive examples to simultaneously group into a small cluster around their center c^+. The circles of radius $(T+\frac{\tau}{2})$ and $(T-\frac{\tau}{2})$ centered at the c^+ form the boundary of the negative set and positive set, respectively, as shown in Fig. 2(d).

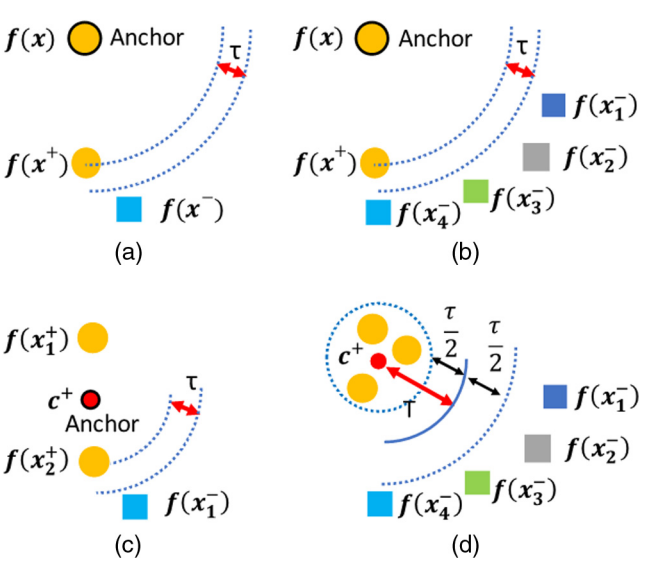

Fig. 2 Failed case of (a) triplet loss, (b) $(N+1)$-tuplet loss, and (c) CCL. The proposed $(N+M)$-tuplet clusters loss is shown in (d).The corresponding loss function of (a), (b), and (c) will not punish this distribution. However, the $f(x^+)$ in (a), (b) and $f(x_2^+)$ in (c) are still closer to some of negative sample than their anchor, in which situation the classifier will still be hard to correctly classifier the label.

In this way, our method is able to address the complex distributions of inter- and intraclass variations and solve the anchor selection problem in traditional deep metric learning methods. In addition, the margin τ and the reference distance T can be learned adaptively through the propagation in the CNN instead of being set manually as hyperparameters. An efficient and simple mini-batch construction scheme is proposed, which chooses different facial expression images from one same identity as the negative set to get rid of the difficult and expensive negative example searching, while mining the positive set online. Thus, our $(N+M)$-tuplet clusters loss guarantees that all the discriminating negative samples can be used efficiently per update, so as to reach an identity-invariant FER.

We design an Inception-ResNet network with long short-term memory to extract not only the image features but also the temporal information in videos. Considering the aim to jointly optimize the softmax loss and $(N+M)$-tuplet clusters loss to explore the potential of both the expression labels and identity labels information, we also design two branches of fully connected (FC) layers and another connecting layer to balance them. The features extracted by the expression classification branch can be fed to the following metric learning processing. This enables each branch to focus better on its own task without being disturbed by the other. In our model, we design two facial expression image sets as inputs: one negative set (images of other expressions with the same identity of the query example) and one positive set (images of the same expression from different subjects).

This paper is an extension of our previous conference work,[7] which introduces an adaptive deep metric learning methods for static FER. Compared to our previous work, the main and unique contribution of this paper is that we adapt our model for videos using a network structure, aiming to extract more discriminative expression representations with both spatial and temporal information of video-based FER. The three main contributions of this paper can be summarized as the following: (1) we propose a $(N+M)$-tuplet clusters loss function for metric learning and build an effective model for video-based FER based on that. (2) We use the online positive mining and identity-aware negative mining scheme to learn distance metrics with less calculations and input passes. Meanwhile, such an approach can maintain the good performance of video-based FER. (3) We design an Inception-ResNet network with long short term memory (LSTM) to extract not only image features but also temporal information in videos and optimize the softmax loss and $(N+M)$-tuplet clusters loss jointly in a unified two-branch FC layer metric learning framework. In our experiments, we show that our proposed approach achieves outstanding results and outperforms several state-of-the-art methods in posed facial expression dataset (e.g., CK+, MMI, and FERA).

2 Related Work

In this section, we briefly introduce the related work and we review the literature in three parts: (1) video-based FER, (2) deep learning methods for FER, and (3) conventional metric learning methods.

2.1 *Video-Based Facial Expression Recognition*

A lot of recent attempts have been made to obtain a better performance on video-based FER. A number of studies focus

on making use of redundant information contained among video frames, which include image set-based approaches, dictionary-based methods, and sequence-based methods.[8] The image set-based approaches build a model on the distribution of video frames using different techniques, e.g., linear subspace,[9] affine/convex hull,[10,11] and manifold methods.[12,13] Afterward, they can measure the similarity between each distribution to match two image sets. One obvious disadvantage of image set approaches, however, is that it is sensitive to the variable volume of video frames and complicated facial variations, which are pervasive in our real-world situations.[14] The dictionary-based approaches create many redundant dictionaries based on video frames and adopt sparse representation-based classifiers for classification.[15,16] In lots of cases, the data size of input video and the number of inputs can be pretty large, and the created dictionary, therefore, becomes a mass volume, which then often leads the dictionary-based approaches to be inefficient. The sequence-based approaches try to extract person-specific facial dynamics from continuous video frames.[17,18] This finally makes them have to rely on robust face trackers, which sometimes turn out to be difficult to realize.

2.2 *Deep Learning Methods for Facial Expression Recognition*

The developments in deep learning, especially the success of CNNs, have made high-accuracy image classification possible in recent years. It has also been shown that carefully designed neural network architectures perform well in FER,[19] and researchers have been making thousands of contributions in this field. Krizhevsky et al.[20] designed AlexNet based on the original CNN layered architecture that adds max-pooling layers and rectified linear units after several convolution layers. Szegedy et al.[21] introduced GoogLeNet that consists of multiple Inception layers. Inception means applying several convolutions on the feature map in different scales. Mollahosseini et al.[19,22] then began to use the Inception layer for the task of FER and their experiments showed that such an architecture achieved state-of-the-art results. Following the success of Inception layers, researchers have delved into this field and several versions of them have been proposed.[23,24] What's more, He et al.[25] combined the Inception layer with residual units and it shows that this new neural network architecture speeds up the training of Inception networks significantly and achieves better results.[26] By contrast, there are only limited studies on video-based FER using CNN, such as Ding and Tao[27] proposed a trunk-branch ensemble CNN, which extracts complementary information from holistic face images and patches cropped around facial components. There are several reasons for it. First of all, existing CNN-based models, which originally focus on images, cannot deal with the specialties of video frames really well. In addition, current well-known facial expression video databases are rather small, which makes it hard to train from real-world video data.

With the great success that recurrent neural networks (RNNs) have achieved in the field of natural language processing, more researchers began to explore the possibility that RNNs can also help to resolve the problems of computer vision. Traditional RNNs can learn the temporal dynamics by mapping input sequences to a sequence of hidden states and then mapping the hidden states to outputs.[28] Using RNNs, we are able to extract the temporal information in a video and escalate the accuracy of video-based FER. Even though RNNs have shown outstanding performances on different tasks, it is hard for them to learn long-term sequences because of the vanishing/exploding gradients problem.[4] LSTMs can solve this problem and memorize the context for a long period. Specifically, an LSTM has three gates: (1) the input gate (i), (2) the forget gate (f), and (3) the output gate (o), to protect and control the cell state at the timestep t. It updates for the timestep t given the inputs x_t, h_{t-1}, and C_{t-1} as the following:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \\
h_t &= o_t \cdot \tanh(C_t),
\end{aligned} \tag{1}
$$

where $\sigma(x) = [1 + \exp(-x)]^{-1}$ is the sigmoid function and x, h, C, W, and b are the input, output, cell state, parameter matrix, and parameter vector, respectively.

Recently, we have seen several works using LSTMs for the task of sequence labeling with prominent performances. Byeon et al.[5] proposed a neural network applying LSTMs in four direction sliding windows and obtained outstanding results. Donahue et al. proposed a long-term recurrent convolutional network. This network combines CNNs and LSTMs so that it is spatially and temporally deep and has the flexibility to be used in various tasks involving sequential inputs and outputs.[28] Fan et al.[29] added LSTMs after 2D-CNNs and combined the resulting feature map with 3D-CNNs for FER, which also proved the superiority of LSTMs in FER.

2.3 *Metric Learning Methods*

Even though the deep learning methods have achieved great success and popularity, the current softmax loss-based network does not present satisfactory results on intraclass compactness and interclass separation. However, researchers explored deep metric learning methods and adopted it for vehicle reidentification and person recognition problems with large intraclass variations, which gives us an inspiration that deep metric learning may offer potential solutions for FER. The initial work is to train a Siamese network with contrastive loss function.[30] It predicts where the examples from given the pairwise examples fed into two symmetric subnetworks. Because it does not involve the interactions of positive pairs and negative pairs, the Siamese network then fails to learn effective metrics with respect to large intra- and interclass variations. One improvement that cannot be ignored is the triplet loss,[31] which achieved outstanding performance in both face recognition and reidentification tasks. This method uses triplets as the inputs are, and each triplet consists of a query, a positive example and a negative example. Recently, some of its variations with faster and stable convergence have been developed, and the $(N+1)$-tuplet loss, as one of them, is the most similar method of our proposed method.[32] We use x^+ and x^- to denote the positive and negative examples of a query example x, which means that

x^+ is from the same class of x, while x^- is not. Considering $(N+1)$-tuplet that includes x, x^+ and $N-1$ negative examples $\{x_j^-\}_{j=1}^{N-1}$, the loss is

$$L(x,x^+,\{x_j^-\}_{j=1}^{N-1};f)=\log(1+\sum_{j=1}^{N-1}\exp(D(f,f^+)+\tau-D(f,f_j^-))), \quad (2)$$

where $f(\cdot)$ is an embedding kernel obtained from the CNN, which takes x and generates an embedding vector $f(x)$. We write it as f for simplicity, with f inheriting all superscripts and subscripts. $D(\cdot,\cdot)$ is defined as the Mahalanobis or Euclidean distance according to different implementations.

Despite their great popularity, the above frameworks still suffer from the costly example mining aiming to provide triplets or nontrivial pairs and poor local optima. In fact, obtaining all possible triplets or pairs would bring a quadratic or even a cubic computation complexity, respectively. Additionally, the online or offline conventional mini-batch sample selection is a huge burden, which increases the complexity further. What's more, from Fig. 2(a)–2(c), we find that the selection of the anchor point has a huge influence on them, especially when the intra- and interclass variations are large. In that case, the triplet loss, $(N+1)$-tuplet loss, and CCL become 0, since the distances from the anchor to positive examples are smaller than the distances between the anchor and negative examples for a margin τ. Therefore, the loss function would ignore these mentioned cases during the back propagation, and it also causes many more requirements of input passes with properly selected anchors to get it right. Fortunately, a recent study presented objective comparisons between the softmax loss and deep metric learning loss, and its results showed that they could be complementary to each other.[33] Inspired by this, we build a unified Inception-ResNet framework to learn this two loss function simultaneously in a more reasonable way.

3 (*N* + *M*)-Tuplet Clusters Loss

In this section, we first describe our intuition of introducing a reference distance T to control the relative boundary $(T+\frac{\tau}{2})$ and $(T-\frac{\tau}{2})$ for the positive and negative examples, respectively, as shown in Fig. 2(d). We rewrite the $(N+1)$-tuplet loss function in Eq. (3) as follows:

$$\begin{aligned}L(x,x^+,\{x_j^-\}_{j=1}^{N-1};f)&=\log\left(1+\sum_{j=1}^{N-1}\exp\left(D(f,f^+)+\left(-T+\frac{\tau}{2}+T+\frac{\tau}{2}\right)-D(f,f_j^-)\right)\right)\\&=\log\left(1+\sum_{j=1}^{N-1}\exp\left(D(f,f^+)-T+\frac{\tau}{2}\right)*\exp\left(T+\frac{\tau}{2}-D(f,f_j^-)\right)\right). \quad (3)\end{aligned}$$

In fact, the term $\exp(D(f,f^+)-T+\frac{\tau}{2})$ in the above equation is aiming to make the positive examples get closer and the term $\exp(T-\frac{\tau}{2}+D(f,f_j^-))$ that is set to force the negative examples away have an "OR" relationship. However, the loss function is very likely to neglect relatively large positive distances because of the long negative distances. One potential way that tries to deal with large intraclass variations is to construct an "AND" function for these two terms.

We further extend the triplet loss and make it for N negative examples and M positive examples. As for a practical multiclassification problem, CCL and the triplet loss only compare the query example with one certain negative example, which only pushes the embedding of the query example to be away from the selected negative class instead of each class. Thus, for these methods, we expect that the final distance metrics will be balanced after plenty of iterations. However, in the late stage of the training stage, the loss become unstable and slow convergence. Because of lacking discriminative negative examples tend to cause a single iteration may show zero errors.

The identity labels in the FER database greatly promote the hard-negative mining to mitigate the problem of the intersubject variations. In real cases, when receiving a query, we combine its negative set with all the different facial expression pictures of the same person. What's more, the traditional deep metric approaches adopt a way that randomly selects one or a group of positive examples, but some extremely hard positive examples may distort the manifold and cause the model to be overfitting. And in the case of spontaneous FER, the expression label can be falsely assigned owing to the subjectivity or varied expertise of the annotators.[2] Therefore, an efficient online mining for M randomly chosen positive examples is what we expect for some large intraclass variation datasets. Specifically, we choose the nearest

Algorithm 1 Online positive mining.

Input

query example and its randomly selected

positive set $\{x_i^+\}_{i=1}^{M}$, and negative set $\{x_j^-\}_{j=1}^{N}$

1. map examples to feature plane with CNN to get:

$$\{f_i^+\}_{i=1}^{M} \text{ and } \{f_j^-\}_{j=1}^{N}$$

2. calculate the positive cluster center $c^+=\frac{1}{M}\sum_{i=1}^{M} f_i^+$

3. calculate the distance from c^+ to each

positive and negative example $D(f_i^+,c^+)$, $D(f_j^-,c^+)$

4. search for the nearest negative distance:

$$D(x_{\text{nearst}}^-,c^+)$$

5. ignore those positive examples satisfying:

$$D(f_i^+,c^+)>D(x_{\text{nearst}}^-,c^+)$$

6. update $c^+=\frac{1}{M^*}\sum_{i=1}^{M^*} f_i^+$

Output

Online mined M^* positive examples and updated c^+

negative example and ignore those positive examples, which have longer distances. We show the details in Algorithm 1.

In summary, the proposed loss function can be expressed as follows:

$$L(\{x_i^+\}_{i=1}^{M},\{x_j^-\}_{j=1}^{N};)=\frac{1}{M*}\sum_{i=1}^{M*}\max\left(0,D(f^+,c^+)-T+\frac{\tau}{2}\right)+\frac{1}{N}\sum_{j=1}^{N}\max\left(0,T+\frac{\tau}{2}-D(f_j^-,c^+)\right). \tag{4}$$

Figure 2(d) gives a simplified geometric interpretation. From the figure and the equation, we know that only when the distances from online mined positive examples to the updated c^+ smaller than $(T-\frac{\tau}{2})$ and the negative examples have larger distances to the updated c^+ than $(T+\frac{\tau}{2})$ can the loss become zero. This is pretty consistent with the principle adopted by lots of discriminative data analysis and data cluster approaches. It can be concluded that the traditional triplet loss and its variations can be regarded as the special situations of the proposed $(N+M)$-tuplet clusters loss.

For a batch consisting of X queries, the total number of distance calculations is $2(N+M)*X$, and the required input passes to evaluate the embedding feature vectors in this proposed model are X. Normally, N and M are much smaller than X. By contrast, triplet loss requires about $2C_X^3$ calculations and C_X^3 passes, while $(N+1)$-tuplet loss requires about $(X+1)*X^2$ times calculations and $(X+1)*X$ passes. Even for a dataset with a moderate size, it is almost impossible and costly for loading all possible triplets into the limited memory during the training phase.

We then define a flexible learning task with adjustable difficulty for the network through assigning different values to T and τ. However, there are two hyperparameters that need manual tuning and validation. Inspired by the idea of adaptive metric learning for SVM,[23] we use a function $T(\cdot,\cdot)$ which is related with each example instead of a constant to formulate the reference distance. Considering the Mahalanobis distance matrix $\mathbf{M}$ in Eq. (5) is quadratic, and it is known that it can be calculated automatically through a linear FC layer as shown in Ref. 34, we assume $T(f_1,f_2)$ as a simple quadratic form, i.e., $T(f_1,f_2)=\frac{1}{2}z^t\mathbf{Q}z+\omega^t z+b$, where $z^t=[f_1^t f_2^t]\in\mathbb{R}^{2d}$, $\mathbf{Q}=\begin{bmatrix}\mathbf{Q}_{f_1f_1} & \mathbf{Q}_{f_1f_2}\\ \mathbf{Q}_{f_2f_1} & \mathbf{Q}_{f_2f_2}\end{bmatrix}\in\mathbb{R}^{2d\times 2d}$, $\omega^t=\left[\omega_{f_1}^t\,\omega_{f_2}^t\right]\in\mathbb{R}^{2d}$, $b\in\mathbb{R}$, f_1 and $f_2\in\mathbb{R}^{2d}$ are the representations of two images in the feature space

$$D(f_1,f_2)=f_1-f_{2M}^2=(f_1-f_2)^T\mathbf{M}(f_1-f_2). \tag{5}$$

Because f_1 and f_2 are symmetric, we can rewrite $T(f_1,f_2)$

$$T(f_1,f_2)=\frac{1}{2}f_1^t\tilde{\mathbf{A}}f_1+\frac{1}{2}f_2^t\tilde{\mathbf{A}}n+f_1^t\tilde{\mathbf{B}}f_2+c^t(f_1+f_2)+b, \tag{6}$$

where $\tilde{\mathbf{A}}=\mathbf{Q}_{f_1f_1}=\mathbf{Q}_{f_2f_2}$ and $\tilde{\mathbf{B}}=\mathbf{Q}_{f_1f_2}=\mathbf{Q}_{f_2f_1}$ are both the $d\times d$ real symmetric matrices [not necessarily positive semidefinite (PSD)], $c=\omega_{f_1}=\omega_{f_2}$ is a d-dimensional vector, and b is the bias term.

Then, we can obtain a new quadratic formula $H(f_1,f_2)=T(f_1,f_2)-D(f_1,f_2)$ by combining the distance metric function with the reference distance function. Substituting Eqs. (5) and (6) to $H(f_1,f_2)$, we get

$$\begin{aligned}H(f_1,f_2)&=\frac{1}{2}f_1^t(\tilde{\mathbf{A}}-2\mathbf{M})f_1+\frac{1}{2}f_2^t(\tilde{\mathbf{A}}-2\mathbf{M})f_2\\&\quad+f_1^t(\tilde{\mathbf{B}}+2\mathbf{M})f_2+c^t(f_1+f_2)+b,\\H(f_1,f_2)&=\frac{1}{2}f_1^t\mathbf{A}f_1+\frac{1}{2}f_2^t\mathbf{A}f_2+f_1^t\mathbf{B}f_2\\&\quad+c^t(f_1+f_2)+b,\end{aligned} \tag{7}$$

$$H(m,n)=\frac{1}{2}m^t\mathbf{A}m+\frac{1}{2}n^t\mathbf{A}n+m^t\mathbf{B}n+c^t(m+n)+b, \tag{8}$$

where $\mathbf{A}=(\tilde{\mathbf{A}}-2\mathbf{M})$ and $\mathbf{B}=(\tilde{\mathbf{B}}+2\mathbf{M})$. Suppose $\mathbf{A}$ is PSD and $\mathbf{B}$ is negative semidefinite, then $\mathbf{A}$ and $\mathbf{B}$ can be factorized as $\mathbf{L}_\mathbf{A}^T\mathbf{L}_\mathbf{A}$ and $\mathbf{L}_\mathbf{B}^T\mathbf{L}_\mathbf{B}$. Therefore, $H(f_1,f_2)$ can be rewritten as follows:

$$\begin{aligned}H(f_1,f_2)&=\frac{1}{2}f_1^t\mathbf{L}_\mathbf{A}^T\mathbf{L}_\mathbf{A}f_1+\frac{1}{2}n^t\mathbf{L}_\mathbf{A}^T\mathbf{L}_\mathbf{A}f_2+f_1^t\mathbf{L}_\mathbf{B}^T\mathbf{L}_\mathbf{B}f_2\\&\quad+c^t(f_1+f_2)+b=\frac{1}{2}(\mathbf{L}_\mathbf{A}f_1)^t(\mathbf{L}_\mathbf{A}f_1)\\&\quad+\frac{1}{2}(\mathbf{L}_\mathbf{A}f_2)^t(\mathbf{L}_\mathbf{A}f_2)+(\mathbf{L}_\mathbf{B}f_1)^t(\mathbf{L}_\mathbf{B}f_2)\\&\quad+c^tf_1+c^tf_2+b.\end{aligned} \tag{9}$$

Based on the above, we propose a general and computationally feasible loss function. Following the previous notations and denote $(\mathbf{L}_\mathbf{A},\mathbf{L}_\mathbf{B},c)^T$ as W, we have

$$L(W,\{x_i^+\}_{i=1}^{M},\{x_j^-\}_{j=1}^{N};f)=\frac{1}{M*}\sum_{i=1}^{M*}\max\left(0,-\mathrm{H}(f_i^+,c^+)+\frac{\tau}{2}\right)+\frac{1}{N}\sum_{j=1}^{N}\max\left(0,H(f_j^-,c^+)+\frac{\tau}{2}\right), \tag{10}$$

where $l(\cdot)$ is the label function for the mined $N+M^*$ training examples in a mini-batch. If the example x_k is from the negative set, $l(x_k)=1$, otherwise $l(x_k)=-1$. In addition, we simplify the term $\frac{\tau}{2}$ in the above equation to be the constant 1, and when we multiply the matrices to corresponding factors, it can be changed to other positive values. Our hinge-loss-like function is

$$L(W,\{x_i^+\}_{i=1}^{M},\{x_j^-\}_{j=1}^{N};f)=\frac{1}{N+M*}\sum_{k=1}^{N+M*}\max(0,l(x_k)*H(f_k,c^+)+1). \tag{11}$$

We optimize Eq. (11) using the standard stochastic gradient descent with momentum. The desired partial derivatives of each example are computed as

$$\frac{\partial L}{\partial W^l} = \frac{1}{N+M*} \sum_{k=1}^{N+M*} \frac{\partial L}{\partial X_k^l} \frac{\partial X_k^l}{\partial W^l}, \tag{12}$$

$$\frac{\partial L}{\partial X_k^l} = \frac{\partial L}{\partial X_k^{l+1}} \frac{\partial X_k^{l+1}}{\partial X_k^l}, \tag{13}$$

where X_k^l is the feature map of the example x_k at the l'th layer. Equation (12) represents the overall gradient, which is the sum of the example-based gradients, while Eq. (13) represents that the partial derivative of each example with respect to the feature maps can be calculated recursively. So, the back propagation algorithm can help to compute the gradients of network parameters.

In fact, the proposed $(N+M)$-tuplet clusters loss can be seen as a straightforward generalization from traditional deep metric learning methods, and it can be easily adopted as a drop-in replacement for the triplet loss and its variations, as well as used in tandem with other performance-boosting methods and models, including pooling functions, modified network architectures, activation functions, or data augmentations.

4 Network Architecture

The proposed two-branch FC layer joint metric learning architecture with softmax loss and $(N+M)$-tuplet clusters loss, denoted as $2\mathbf{B}(N+M)$ softmax is shown in Fig. 3, and the detailed network architecture is shown in Fig. 4. In this section, we introduce the neural network parts in our model. We use an Inception-ResNet network here because of the remarkable performances of Inception and ResNet in various tasks, especially the combination of them. The Inception deep CNN architecture was first introduced as GoogLeNet, which is also known as Inception-v1.[21] It highly utilizes the computing resources inside the network by increasing the width and depth of the neural network. In practice, it applies multiple convolutions on the same feature map with different scales trying to extract different features. The cool thing is this network can keep the computational budget constant, which makes it quite popular in real cases. As for the residual connection, it was introduced in 2015.[25] It mainly focuses on the degradation problem: when a neural network has more and more layers, its accuracy gets saturated and then degrades rapidly. The authors gave convincing theoretical and practical evidence to prove that reformulating the layers as learning residual functions with reference to the layer inputs instead of learning unreferenced functions can address this problem. Its results reached state-of-the-art and gained much popularity. In addition, considering the lack of temporal relationship that these models are able to extract, we add an LSTM unit that takes the output of the Inception-ResNet network as an input and extracts the temporal relations from it. Then, we add a fully connected layer with a softmax activation function, which outputs the resulting feature map.

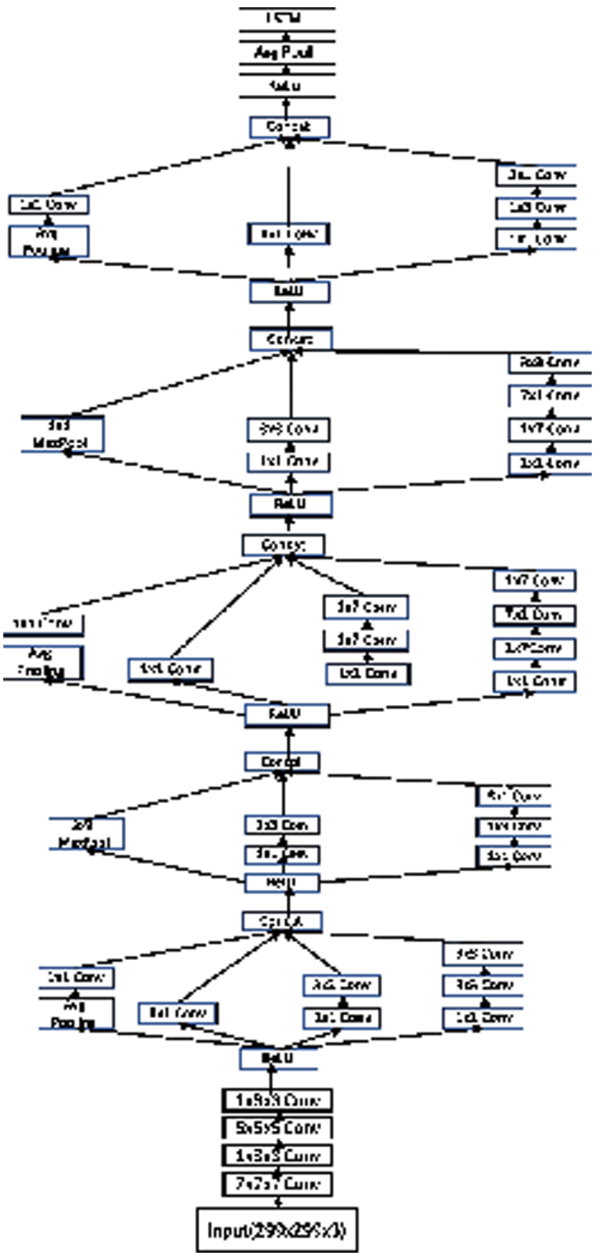

Fig. 4 The detailed network architecture.

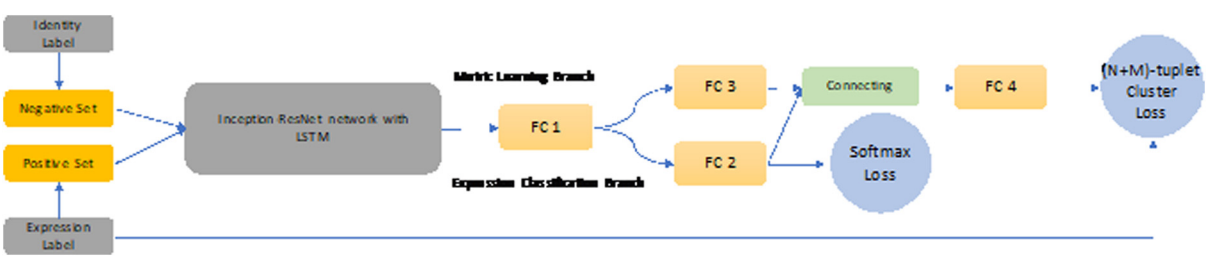

Fig. 3 The proposed network structure. In the testing phase, only the convolutional groups and expression classification branch with softmax are used to recognize a single facial expression image.

Specifically, our Inception-ResNet takes an input video with the size of $299 \times 299 \times 3$ (299×299 frame size and three color channels), and it is followed by the stem layer. After that, there are Inception-ResNet-A, Reduction-A, Inception-ResNet-B, Reduction-B, Inception-ResNet-C, Average Pooling, Dropout, an LSTM unit, and a fully connected layer, respectively. Here, each reduction block is for reducing the grid size. In addition, using an LSTM unit makes the whole model suitable for video-based tasks, due to the output from the previous Inception-ResNet block containing the time sequences within the input video. Based on that, vectorizing the resulting feature map of the Inception-ResNet block on its sequence dimension can provide the necessary sequenced input for the LSTM unit. We also find that the LSTM unit with a size of 200 hidden units has the best performance for the video-based FER task.

5 Experiments and Results

In this section, we not only introduce the face databases we used in our experiments but also report our experiment results and compare them with the state-of-the-art.

5.1 *Implementation Details*

In our experiment, we resize the faces to 299×299 pixels. The main reason why we choose such a large image size as input is the consideration that we are able to use deeper neural networks and extract more features from the input. Ten frames of each sequence are extracted for video-based recognition. Our model was implemented using Tensorflow and TFlean toolboxes on NVIDIA Titan X GPUs. In the testing stage, we take about 1.25 s for processing each sequence. All the networks in the experiment have the same settings and are trained from scratch for each database we select. In the training phrase, we used asynchronous stochastic gradient descent with the momentum of 0.9, the weight decay of 0.0001, and the learning rate of 0.015. In addition, we chose categorical cross entropy for our loss function and used accuracy as the evaluation metric.

5.2 *Experimental Evaluations*

To evaluate the effectiveness of our proposed model, we have conducted extensive experiments on three well-known and publicly available facial expression databases: $CK+$, MMI, and FERA. For the fair comparison, we follow the protocol used by previous works.[22,35] Three baseline methods are employed to demonstrate the superiority of the metric learning loss and two-branch FC layer network, respectively, i.e., adding the $(N+M)$-tuplet clusters loss or $(N+1)$-tuplet loss with softmax loss after the EC branch, denoted as $1\mathbf{B}(N+1)$ softmax or $1\mathbf{B}(N+M)$ softmax, and combining the $(N+1)$-tuplet loss with softmax loss via the two-branch FC layer structure, as $2\mathbf{B}(N+1)$ softmax. We do not compare with the triplet loss here, because the number of triplets grows cubically with the number of images, which makes it impractical and inefficient. With randomly selected triplets, the loss failed to converge during the training phase.

CK+: The extended Cohn–Kanade database (CK+)[36] contains 593 videos from 123 subjects, while only 327 sequences from 118 subjects contain facial expression labels that ranging from 7 different expressions (i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise). The label is only provided for the last frame (peak frame) of each sequence. We split the CK+ database to eight subsets in a strict subject-independent manner, and an 8-fold cross validation is employed. Data from six subsets are used for training and the others are used for validation and testing. The confusions matrix of the proposed method evaluated on the CK+ dataset is reported in Table 1. It can be observed that the disgust and happiness expressions are perfectly recognized while the contempt expression is relatively harder for the network because of the limited training examples and subtle muscular movements. As shown in Table 1, our proposed $2\mathbf{B}(N+M)$ softmax outperforms most of the state-of-the-art methods. Not surprisingly, it also beats the baseline methods obviously benefiting from the combination of deep metric learning loss and two-branch architecture (Fig. 5).

MMI: The MMI database[34] includes 31 subjects with frontal-view faces which contain a full temporal pattern of expressions, i.e., from neutral to one of six basic expressions

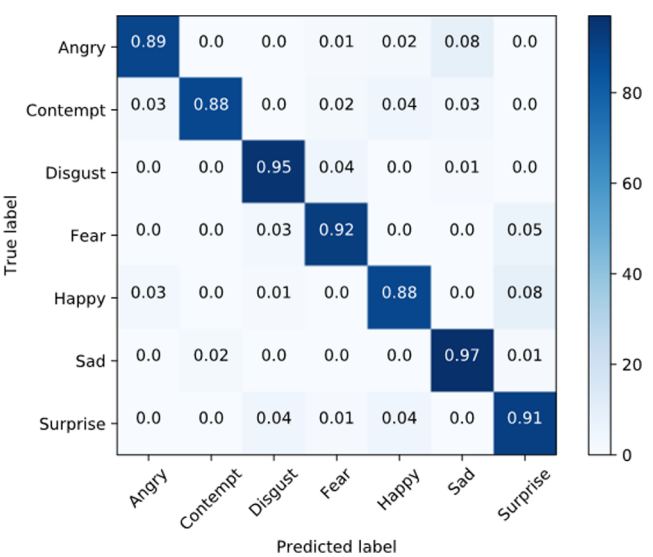

Fig. 5 Average confusion matrix obtained from the proposed method on CK+.

Table 1 Recognition accuracy comparison.

	State-of-the-art methods	$1B(N+1)$ softmax	$2B(N+1)$ softmax	$1B(N+M)$ softmax	$2B(N+M)$ softmax
CK+	84.1,[37] 88.5,[38] 92.0,[39] 93.2,[22] 93.6[40]	88.32	91.79	92.91	93.90
MMI	63.4,[41] 74.7,[39] 79.8,[42] 86.7[43]	77.13	78.93	77.72	79.03
FERA	56.1,[41] 55.6,[38] 76.7[22]	64.36	73.75	75.27	79.94

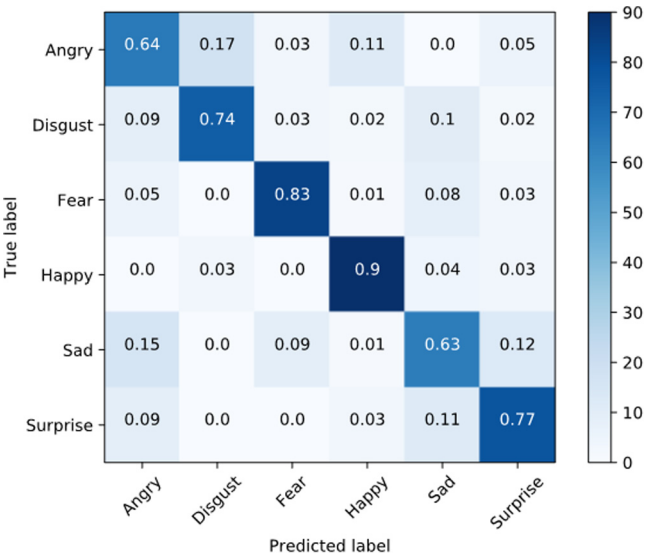

Fig. 6 Average confusion matrix obtained from the proposed method on MMI.

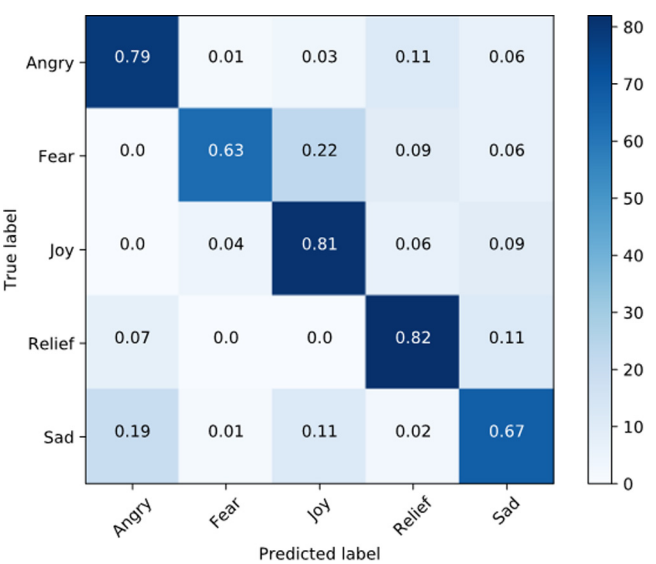

Fig. 7 Average confusion matrix obtained from the proposed method on FERA.

as time goes on, and then released. These subjects range in age from 19 to 62, which provides a great diversity. Because of the above features, it is especially favored by the video-based methods to exploit temporal information. We extracted static frames from each sequence, which resulted in 11,500 images. Afterward, we divided MMI dataset into 10 subsets for person-independent 10-fold cross validation and also divided videos into sequences of 10 frames to shape the input tensor for our network. We reported the confusion matrix of our proposed model on the MMI database in Fig. 6. As shown in Table 1, the performance improvements in this small database without causing overfitting are impressive. We can also conclude from the results that the proposed method outperforms many other works.

FERA: The GEMEP-FERA database[44] is a subset of the GEMEP corpus, which is used as database for the FERA 2011 challenge.[38] GEMEP is developed by the Geneva Emotion Research Group at the University of Geneva. This database has 87 image sequences of 7 subjects.

Each subject presents the facial expressions of the following five emotion categories: anger, fear, joy, relief, and sadness. The head poses in this database are mainly frontal with relatively fast and different movements. In addition, every video is annotated with action units and holistic expressions. Same as the above databases, we extracted static frames from the sequences, and finally obtained around 7000 images. Here, we employed a sevenfold cross validation and reported the confusion matrix of our model in Fig 7. With the augmentation of deep metric learning and two-branch FC layer network, we achieve, to our knowledge, state-of-the-art.

6 Conclusion

We propose a $(N + M)$-tuplet clusters loss and combine it with softmax loss. Using an Inception-ResNet network with LSTM and a unified two-branch FC layer joint metric learning architecture, we aim to get rid of the attribute variations due to different identities in FER and escalate the accuracy of video-based FER. We also adopt an effective online positive-mining and identity-aware negative-mining scheme, which can reduce the number of input passes and computations. The experimental results on three well-known databases, which are CK+, MMI, and FERA, show that our proposed method outperforms many of the state-of-the-art approaches. More appealing, the $(N + M)$-tuplet clusters loss function has clear intuition and geometric interpretation for generic applications. In future work, we intend to focus on this point and explore some possible uses of it in other fields.

References

1. G. B. Huang et al., "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Vol. **1**. No. 2. Technical Report 07-49, University of Massachusetts, Amherst (2007).
2. Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing lower face action units for facial expression analysis," in *Proc. Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, IEEE (2000).
3. M. F. Valstar et al., "Meta-analysis of the first facial expression recognition challenge," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **42**(4), 966–979 (2012).
4. S. Hochreiterand and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
5. W. Byeon et al., "Scene labeling with LSTM recurrent neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2015).
6. H. Liu et al., "Deep relative distance learning: tell the difference between similar vehicles," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
7. X. Liu et al., "Adaptive deep metric learning for facial expression recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops* (2017).
8. J. R. Barr et al., "Face recognition from video: a review," *Int. J. Pattern Recogn. Artif. Intell.* **26**(05), 1266002 (2012).
9. Z. Huang et al., "A benchmark and comparative study of video-based face recognition on COX face database," *IEEE Trans. Image Process.* **24**(12), 5967–5981 (2015).
10. Y. Hu, A. S. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(10), 1992–2004 (2012).
11. P. Zhu et al., "Image set-based collaborative representation for face recognition," *IEEE Trans. Inf. Forensics Secur.* **9**(7), 1120–1132 (2014).
12. M. T. Harandi et al., "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2011).
13. Z. Cui et al., "Image sets alignment for video-based face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE (2012).

14. M. Shao et al., "A comparative study of video-based object recognition from an egocentric viewpoint," *Neurocomputing* **171**, 982–990 (2016).
15. Y.-C. Chen et al., "Dictionary-based face recognition from video," in *European Conf. on Computer Vision*, Springer, Berlin, Heidelberg (2012).
16. L. Liu et al., "Toward large-population face identification in unconstrained videos," *IEEE Trans. Circuits Syst. Video Technol.* **24**(11), 1874–1884 (2014).
17. M. Bicego, E. Grosso, and M. Tistarelli, "Person authentication from video of faces: a behavioral and physiological approach using pseudo hierarchical hidden Markov models," *Lect. Notes Comput. Sci.* **3832**, 113–120 (2016).
18. A. Hadid and M. Pietikainen, "Combining appearance and motion for face and gender recognition from videos," *Pattern Recogn.* **42**(11), 2818–2827 (2009).
19. A. Mollahosseini et al., "Facial expression recognition from world wild web," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops* (2016).
20. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (2012).
21. C. Szegedy et al., "Going deeper with convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2015).
22. A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, IEEE (2016).
23. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning* (2015).
24. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
25. K. He et al., "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2016).
26. C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *AAAI* (2017).
27. C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
28. J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* (2015).
29. Y. Fan et al., "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. of the 18th ACM Int. Conf. on Multimodal Interaction*, ACM (2016).
30. S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR 2005*, Vol. **1**, IEEE (2005).
31. S. Ding et al., "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recogn.* **48**(10), 2993–3003 (2015).
32. K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems* (2016).
33. Y. Em et al., "Incorporating intra-class variance to fine-grained visual recognition," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, pp. 1452–1457, IEEE (2017).
34. M. Pantic et al., "Web-based database for facial expression analysis," in *IEEE Int. Conf. on Multimedia and Expo, ICME 2005*, IEEE (2005).
35. Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. of the 2015 ACM on Int. Conf. on Multimodal Interaction*, ACM (2015).
36. P. Lucey et al., "The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE (2010).
37. C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern Recogn. Image Anal.* **24**(1), 124–132 (2014).
38. M. F. Valstar et al., "The first facial expression recognition and analysis challenge," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, IEEE (2011).
39. M. Liu et al., "Au-aware deep networks for facial expression recognition," in *10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE (2013).
40. X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using {l}_{p}-norm MKL multiclass-SVM," *Mach. Vis. Appl.* **26**(4), 467–483 (2015).
41. M. Liu et al., "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conf. on Computer Vision*, Springer, Cham (2014).
42. S. Taheri, Q. Qiu, and R. Chellappa, "Structure-preserving sparse decomposition for facial expression analysis," *IEEE Trans. Image Process.* **23**(8), 3590–3603 (2014).
43. C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image Vis. Comput.* **27**(6), 803–816 (2009).
44. T. Bänziger and K. R. Scherer, "Introducing the Geneva multimodal emotion portrayal (gemep) corpus," Chapter 6.1 in *Blueprint for Affective Computing: A Sourcebook*, K. R. Scherer, T. Bänziger, and E. Roesch, Eds., pp. 271–294 (2010).

Xiaofeng Liu received his BEng degree in automation and his BA degree in communication from the University of Science and Technology of China, Hefei, China, in 2014. He is currently pursuing his PhD at the University of Chinese Academy of Sciences, Beijing, and is a research associate in the Department of Electrical and Computer Engineering, Carnegie Mellon University. His research interests include image processing, computer vision, and pattern recognition.

Yubin Ge is a master's student at the University of Pittsburgh, major in information science, and an research assistant at Carnegie Mellon University. He received his BS degree in information management from Zhengzhou University, Henan, China, in 2012. His current research interests include natural language processing, computer vision, and deep learning.

Chao Yang received his bachelor's degree in mathematics from the University of Science and Technology of China, Hefei, China. He is currently pursuing his PhD in computer science at the University of Southern California, USA. He worked in the Microsoft Research Asia, Toyota Technical Institute of Chicago and Adobe. His research interest includes computer vision, machine learning, and deep learning.

Ping Jia obtained his MSc degree in computer science from the University of Science and Technology of China and his PhD from the Graduate University, Chinese Academy of Sciences (CAS). He is currently the president of Changchun Institute of Optics, Fine Mechanics and Physics, CAS. His current research interests include image processing, computer vision, and optical engineering.