Facial Expression Recognition with Identity and Emotion Joint Learning

Journal:	*Transactions on Affective Computing*
Manuscript ID	TAFFC-2017-12-0277
Manuscript Type:	Brief
Keywords:	facial expression recognition, emotion recognition, face recognition, joint learning, transfer learning

Facial Expression Recognition with Identity and Emotion Joint Learning

Ming Li,*Member,IEEE*, Hao Xu, Xingchang Huang, Zhanmei Song, Xiaolin Liu, and Xin Li *Fellow, IEEE*

Abstract—Different subjects may express a specific expression in different ways due to inter-subject variabilities. In this work, besides training deep-learned facial expression feature (emotional feature), we also consider the influence of latent face identity feature such as the shape or appearance of face. We propose an identity and emotion joint learning approach with deep convolutional neural networks (CNNs) to enhance the performance of facial expression recognition (FER) tasks. First, we learn the emotion and identity features separately using two different CNNs with their corresponding training data. Second, we concatenate these two features together as a deep-learned Tandem Facial Expression (TFE) Feature and feed it to the subsequent fully connected layers to form a new model. Finally, we perform joint learning on the newly merged network using only the facial expression training data. Experimental results show that our proposed approach achieves 99.31% and 84.29% accuracy on the CK+ and the FER+ database, respectively, which outperforms the residual network baseline as well as many other state-of-the-art methods.

Index Terms—Facial expression recognition, Emotion recognition, Face recognition, Joint learning, Transfer learning

1 Introduction

Facial Expression Recognition (FER) is a well defined task, aiming to recognize facial expressions with discrete categories (e.g., neutral, sad, contempt, happy, surprise, angry, fear, disgust, etc.) or continuous levels (e.g., valance, arousal) from still images or videos. Although many recent works focus on video or image sequence based FER tasks [1], [2], still image based FER still remains as a challenging problem. First, the differences between some facial expressions might be subtle and thus difficult to classify them accurately in some cases. Second, different subjects express the same specific facial expression in different ways due to the inter-subject variability and their facial biometric shapes, etc.

These two challenging problems can be visualized in the following examples in Fig. 1. The left part of Fig. 1 contains two representative faces and both of them are labeled with ”sad” facial expression. However, their eight-category classification scores have great differences in our initial experiment, which is shown in Table 1. The prediction for the left subject is reasonable as the ”sad” emotion ranks the top. However, for the person on the right, the score of ”angry” is a little higher than ”sad”, which did not correctly predict her emotion.

Generally, the inter-speaker variability of the faces could potentially lead to errors in emotion classification because the neutral face of one subject could already be very similar to the typical faces of other emotion categories (e.g. the right subject in Fig. 1 and the ”angry” emotion). Therefore, we believe that if we add a compact description of the subject’s facial identity or biometric information as an auxiliary input to our model, the FER system can become more robust against the inter-speaker variability just as the speaker adaptation technique in speech recognition tasks [3].

Previous works show that deep neural network based methods have achieved excellent performance in face related recognition tasks [4] [5] [6] [7] [8] [9] [1]. In face recognition, Deep Convolutional Neural Networks (CNNs) outperform those traditional methods with hand-crafted features [10] [11] [12] [13], and even perform better than human beings [6] [4] [5] [14]. However, in FER tasks, the system performance still needs to be further enhanced. Lack of large scale labeled training data, inconsistent and unreliable emotion labels and inter-subject variabilities all limit the performance of CNN on the FER task. Therefore, in this work, we aim to utilize additional face recognition training data to perform identity and emotion joint learning for FER.

Related to our work, Xu et al. [15] proposed a transfer learning method from face recognition to FER using CNN directly. Also, Jung et al. [1] proposed a joint fine-tuning method that jointly learns the parameters from image sequences. However, unlike these two methods, we do not transfer the network structures and parameters from face recognition to FER directly. Instead, we extract the high-level identity feature from the face recognition network and consider it as an auxiliary input feature for our FER model. As shown in Fig. 2, we concatenate both the high-level emotion and identity features as Tandem Facial Expression (TFE) features and feed it to the subsequent fully connected layers to form a new network.

In this paper, we adopt the CNN architecture to discover latent identity and emotion features. First, we pre-train latent emotion and identity features separatively using two different CNNs (ResNet [16] for emotion and DeepID [4] for identity) with their own training data. Furthermore, we merge these two networks together by concatenating the

- *Ming Li is with the School of Electronics and Information Technology, Sun Yat-sen University and Data Science Research Center, Duke Kunshan University. Hao Xu and Xingchang Huang are with the School of Data and Computer Science, Sun Yat-sen University. Zhanmei Song, Xiaolin Liu, and Xin Li are with the School of Preschool Education, Shandong Yingcai University.*
Corresponding author: Zhanmei Song, E-mail: songzhanmei@126.com

Manuscript received; revised.

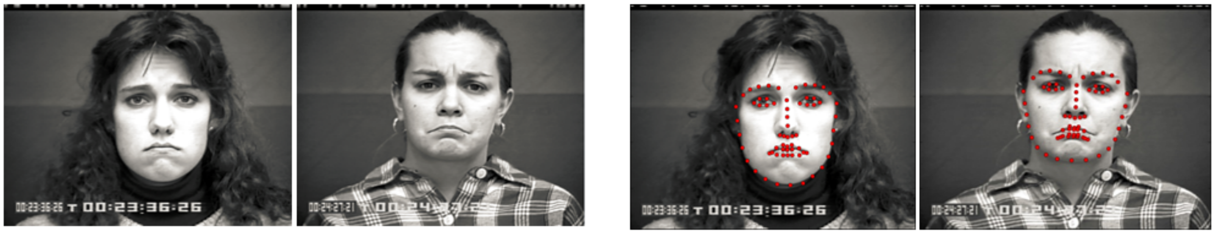


Fig. 1. Two example face images with ”sad” facial expressions. The left part is the original faces of these two subjects and the right part is the faces with detected landmarks.

TABLE 1
The predicted scores on 8 facial expression categories for the two example images in Fig. 1.

Figure	Neutral	Angry	Contempt	Disgust	Fear	Happy	Sad	Surprise
Left	**0.2536**	0.0042	0.0038	0.0004	0.0025	0.0003	**0.7332**	0.0020
Right	0.0002	**0.5110**	0.0	0.0009	0.0001	0.0001	**0.4876**	0.0001

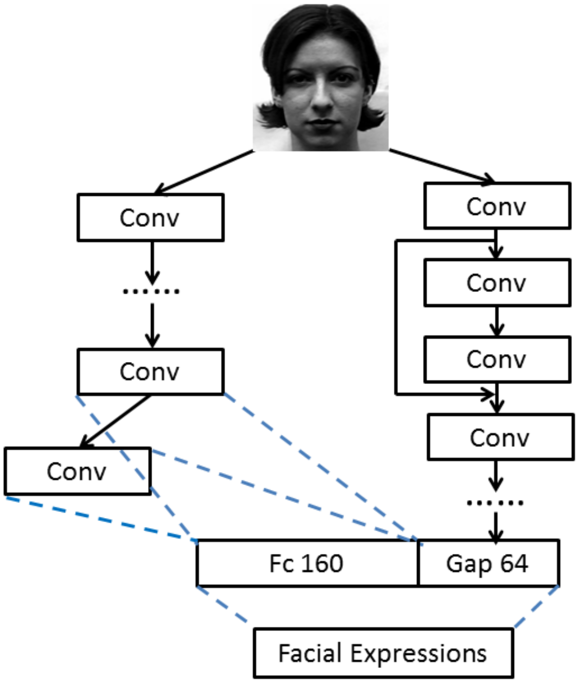


Fig. 2. Our model consists of two convolutional neural networks. The left one represents the DeepID network learning the identity features. The right deep residual network is trained with facial expression databases. After training separatively, the identity feature and the deep-learned emotion feature are concatenated as the TFE features and feed to the subsequent fully connected layers. Finally, we perform joint learning on the new merged network using only the facial expression database.

deep-learned features and feed to a new fully connected layer. Finally, we use FER training data to jointly learn the parameters of the merged new network. To the best of our knowledge, there is no previous work using auxiliary deep identity feature with deep emotion feature together for joint facial expression learning.

2 Related Work

In this section, we will introduce two main types of features used in the FER task, namely hand-crafted features and deep-learned features.

2.1 Hand-Crafted Feature Based Method

Before deep learning based approaches dominate face recognition and FER tasks, many works have been conducted based on the hand-crafted features. These approaches usually perform frontend feature extraction and backend classification separately [9]. During the stage of feature extraction, traditional features, such as Local Binary Patterns (LBP) [10] [12], Gabor wavelet coefficients [11], Scale-Invariant Feature Transform (SIFT) [13], and GaussianFace [14] are designed with prior domain knowledge. Moreover, supervised classifiers, such as SVM [17], feedforward Neural Network (NN) [18] and Extreme Learning Machine [19], [20] are adopted for the subsequent modeling.

2.2 Deep-Learned Feature Based Method

Generally, deep learning based methods outperform hand-crafted feature based approaches and achieve state-of-the-art performance on both face recognition [6] [4] [5] and FER

[21] [1] [9] [2] tasks. For example, Sun et al. [6] proposed CNN model and DeepID features for face verification. In order to boost the performance, Sun et al. [4] proposed a Siamese network to train in face pairs. Furthermore, CNN models have been widely used on the FER task. Tang et al. [7] replace the softmax layer with SVM in the CNN framework and achieved the best accuracy on FER2013 dataset [22] in the ICML 2013 Representation Learning Challenge. Emad et al. [23] then proposed a FER+ dataset with more accurate labels and produced a benchmark on this dataset with VGG13 network. Jung et al. [1] and Zhao et al. [2] both consider temporal structures on top of CNNs to model the image sequences.

Recent high performance models normally accompany with deep architectures and a large number of convolutional kernels. Alex et al. [24] has proposed AlexNet for the ImageNet challenge and achieve better performance than other methods at that time. Subsequently, Karen et al. [25] presented their deeper networks with 16 and 19 layers respectively and found that deeper networks can bring better performance. Futhermore, He et al. [16] proposed deep residual network (ResNet) and trained a CNN with 152 layers. ResNet can converge faster and perform more accurately due to its residual learning mechanism, shortcut connection [16] and batch normalization [26]. Also, Christian et al. [27] proposed GoogleNet and its inception_v4 architecture. They further combined the residual network architecture with Inception-v4 as Inception-ResNet [28] and achieve better performance on the ImageNet dataset [29]. In this work, our approach adopts deep ResNet and CNN for their high performance and less chance of overfitting on image related tasks.

A recent work proposed by Xu et al. [15] used transfer learning with CNN for FER. The difference between our work and their work is that we learn both the emotion and identity features using two separate deep convolutional neural networks and construct a deep-learned Tandem Facial Expression (TFE) feature in the merged model instead of just transferring the weights of the pre-trained face recognition networks and fine-tuning. Compared with the work by Jung et al. [1], we use feature-level concatenation of deep-learned identity and emotion features to form a new network with joint learning rather than fine-tuning two softmax layers with the landmark-based features.

3 Our Approach

In this section, we will introduce our proposed method in details.

3.1 Overview of our Network Architecture

As shown in Fig. 2, our model consists of two CNNs. The left one is the DeepID network proposed in [4], [6], containing four convolutional layers. Actually, the architecture we use is the same as the ConvNet structure proposed in [4], which learns fully-connected layer (DeepID2 feature) from both the third and fourth convolutional layers, generating a compact feature representation with 160 dimensions. The right network is constructed according to the deep ResNet [16] and we choose the ResNet18 structure and also build a shallower network called ResNet12 for different tasks based on the size of the input images and the size of datasets. In ResNet, we use shortcut connection for deep residual learning and batch normalization layers [26] for faster convergence and better accuracy. We do not use fully-connected layer (FC) to flatten the feature maps. Instead, we use global average pooling (Gap) to generate a compact representation with reduced number of parameters [16]. But for convenience, we still use fully-connected layer to model the concatenated TFE features.

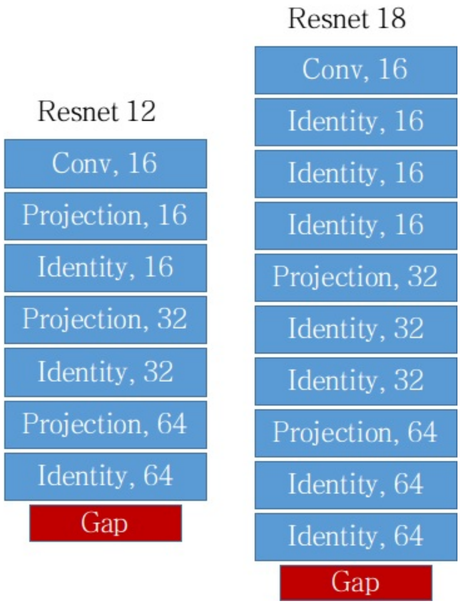


Fig. 3. Our ResNet12 and ResNet18 architectures. "project", "identity" and "gap" are projection block, identity block and global average pooling, respectively.

Actually, the left and the right networks are considered as a merged joint network in Fig. 2. During training, the features of DeepID network and ResNet are concatenated as the input for fully-connected layers and jointly learn the entire network for FER tasks. In order to guide this merged network to extract identity and emotion features, we do not train from scratch but pre-train the weights of both two subnetworks separately with the corresponding datasets except the fully connected layer.

Deep residual network (ResNet) has achieved great success in multiple challenges [16]. ResNet is based on the deep residual learning framework and it is easier to optimize the residual rather than the original mapping. Therefore, in our work we adopt ResNet for training emotion features. Specifically, the architecture of residual networks is made up of the following two blocks shown in Fig. 4. The left one, called "identity block", contains a shortcut link connecting the input x to the convolution output. The right one, called "projection block", contains a convolution operation in the shortcut connection, aiming to ensure the same output size of feature maps using a 1×1 convolution with a stride of 2 while doing the element-wise sum at the output.

In our work, the right block is used to increase the number of convolutional kernels and decrease the output size of feature maps by half. The stride for each convolution operation is 2 and there is no max-pooling layer. After the

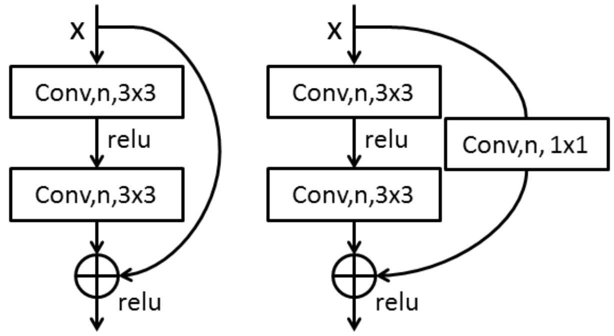


Fig. 4. The building blocks of our residual network. n denotes the number of filters in convolution.

last convolution operation, we use Global Average Pooling (Gap) layer to generate a 64-dimensional emotion feature as shown in Fig. 2. Considering the size of datasets, we use two ResNet structures. The first one is ResNet18 structure for processing the FER+ dataset [23] with over 35k 48×48 images. For the smaller dataset, CK+, we reduce the number of parameters by removing four identity building blocks when $n = 16, 16, 32, 64$ and adding one project building block when $n = 16$. We call this shallower network as ResNet12. We adopt batch normalization after each convolution and before the rectified linear unit (ReLU) activations [30] [16]. The architectures of these two ResNets are shown in Fig. 3.

3.2 Identity and Emotion Feature Concatenation

Suppose that the identity and emotion features of an arbitrary input image are represented as Z_i and Z_e, respectively. Then, we can reconstruct the new TFE representation Z_{tfe} by concatenating Z_i and Z_e together.

However, sometimes the deep-learned Z_i and Z_e features are not in the same scale because both network structure and training data are different. In our work, we firstly normalize Z_i and Z_e with batch normalization [26]. Then we concatenate these two features together to form the TFE feature Z_{tfe}.

4 Experimental Results

4.1 Datasets

In this work, we evaluate the proposed method on two popular FER datasets, namely Extended Cohn-Kanade (CK+) database [31] and FER+ database [23]. These two FER datasets are used for deep-learned emotion feature extraction and joint learning, while the identity features are learned from the CASIA-WebFace database [32].

- **CASIA-WebFace [32]:** In this dataset, the face images are collected from the Internet, containing 10,575 subjects and 494,414 images. It is usually considered as a standard large scale outside training dataset for face verification challenge on the LFW task [33].
- **LFW:** LFW (Labeled Faces in the Wild) dataset contains 13,233 face images from 5,749 identities collected on the Internet. As a benchmark for comparison, LFW suggests reporting performance with 10-fold cross validation using splits they have randomly generated (6,000 pairs) [33]. However, it is inadequate for training a network because the majority of identities in LFW have only one or two face images. Therefore, many works use outside large scaled datasets for model training and perform testing on LFW according to the standard protocols.
- **CK+:** The CK+ database includes 327 image sequences with labeled facial expressions. For each image sequence, only the last frame is provided with an expression label. In order to collect more images for training, we usually selected the last three frames of each sequence for training or validation purpose. Additionally, the first frame from each of the 327 labeled sequences would be chosen as the ”neutral” expression. As a result, this dataset can provide totally 1308 images with 8 labeled facial expressions. For testing, we follow the 10-fold cross validation testing protocol on the CK+ database.
- **FER+:** This dataset comes from the face expression recognition challenge [22] in the ICML 2013 Representation Learning Workshop. It consists 28,709 48×48 face images for training. The test set has 3,589 images and there are totally 7 discrete facial expressions (anger, disgust, fear, happiness, sadness, and surprise) for classification. However, due to its noisy labels, this dataset is labeled again using crowd-sourced services [23]. In this way, majority voting on the labels or multi-label learning will be feasible on the FER+ dataset.

4.2 Parameter Settings

For the DeepID network, we follow the same parameter setting in [6], with a 160-dimensional representation in the fully-connected layers. A dropout layer is used after the DeepID layer, with a probability of 0.4 to reduce over-fitting [34].

For the deep ResNet, we have two different settings for the number of layers. We use the ResNet18 architecture for the FER+ dataset but use a shallower network ResNet12 for the CK+ dataset, which has much less data for training and testing.

As for Stochastic Gradient Descent (SGD) method during back propagation [35], we apply different parameters in CK+ dataset and FER+ dataset. For CK+, we initialize the learning rate as 0.16 and 0.01 respectively for ResNet and DeepID network with a mini-batch size of 128 and a momentum of 0.9 [36]. The networks in our model are trained for up to 200 epochs and fine-tuned for up to 100 epochs as well. For FER+, the learning rate is initialized as 0.1 for ResNet training and 0.001 for final joint learning.

4.3 Pre-Processing

For CASIA-WebFace dataset, we need to use pre-processing pipeline, including face detection, face landmarks detection face alignments and face cropping. We use the tools from mmlab, CUHK [37] to detect face and landmarks. After these processing steps, those missed faces are removed and there are totally 435,863 faces remaining. Then we use a template (the first image) in the LFW [33] dataset to align the faces in the CASIA-WebFace. Finally, images from

TABLE 2
Average accuracy on CK+ of our models using 10-fold cross validation

Our Methods	Average Accuracy on CK+
ResNet12	97.56%
TFE-JL	**99.31%**

the datasets we use (CASIA-WebFace, CK+, LFW) will be cropped in the same way retaining the eye brow and jaw. As LFW provides a deep-funneled version and CK+ is collected containing the frontal whole faces, there is no need to do face alignment. FER+ has been pre-processed and cropped as well.

During training the DeepID network using CASIA-WebFace, we randomly select one image from each person for validation and therefore generate a validation set of 10,575 images, while the remaining images are used for training. It is worth noting that the training set of CASIA-WebFace is augmented with horizontal flipping while the training set in FER+ is augmented with horizontal flipping, shifting and rotation. In addition, all the images are pre-processed with per-pixel mean subtraction and standard deviation normalization [7].

4.4 Evaluation on CASIA-WebFace and LFW

After the pre-processing step, we train our DeepID network for extracting the auxiliary identity feature on CASIA-WebFace. As mentioned above, we select one image from each identity and therefore construct a testing set with 10,575 images and a training set with 425,288 images. After training 200 epochs for DeepID network, the accuracy on the testing set of CASIA-WebFace can achieve 69% and the accuracy on LFW for face verification can achieve 91% using cosine similarity with a 0.15 threshold.

Since our goal is to extract a reasonable good quality identity feature, we may not need to fully optimize the face verification performance on LFW. We use a single DeepID network and single patch for each face image without any ensemble method.

4.5 Evaluation on CK+

During the 10-fold cross validation testing, we train our ResNet from scratch for each rotation with 200 epochs. After the first step training, we combine these two networks, which extract identity features and emotion features respectively, to form a 224-dimensional TFE representation. Finally, we jointly learn the parameters of the merged network with the CK+ training data.

One example of the training and joint learning process on the CK+ database is provided in Fig. 5. The training accuracy (blue) increase gradually while the validation accuracy (orange) increase with fluctuation but finally converge to 97.56%. In the joint learning stage using the TFE feature, the accuracy on the validation set converges faster and better than the first training stage and can achieve up to 99.24%. As shown in Table 2, the performance of our proposed method outperforms the ResNet baseline by 1.68% absolutely.

Besides the comparisons with our own implemented baselines, we also compare our method with other state-of-the-art approaches. Our method can achieve around 2% absolute improvement compared with PPDN model [2] as shown in Table 3. Actually, the sequence-based PPDN also can achieve 99.3% accuracy on the test set but it was pre-trained on CASIA-WebFace and used the whole sequence of images in CK+ for training, which does not match with our experimental setting (only the first one and the last three frames of each CK+ sequence are used).

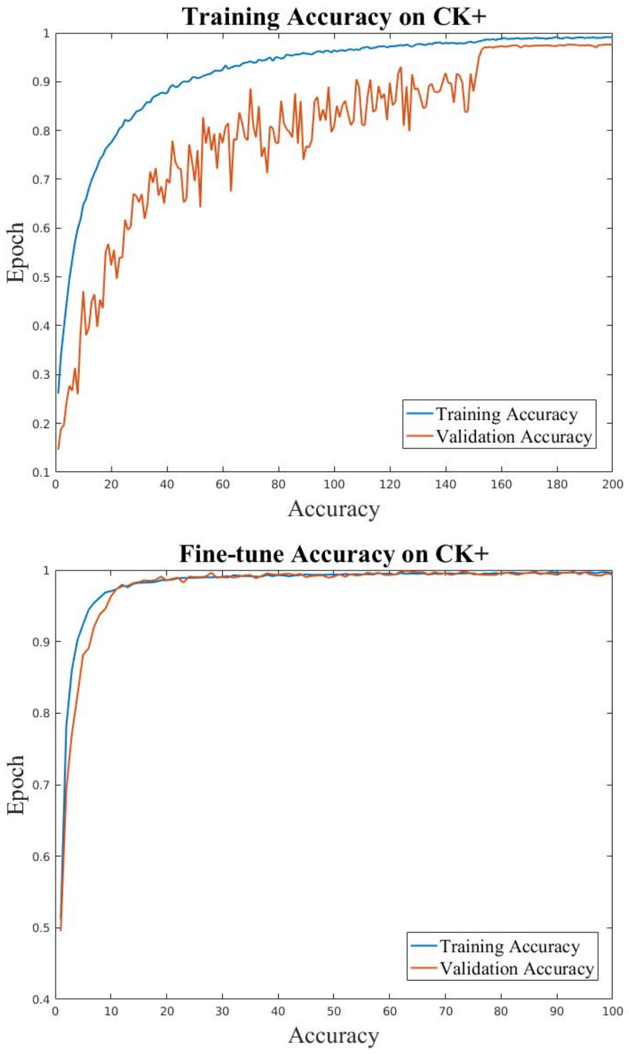


Fig. 5. Training and validation performance on the CK+ database during the training and the joint learning stage.

TABLE 3
Comparion with state-of-the-art methods

Methods	Average Accuracy on CK+
PPDN [2]	97.3%
DTAGN(Weighted Sum) [1]	96.94
DTAGN(Joint) [1]	97.25
BDBN [9]	96.7
TFE-JL	**99.31%**

4.6 Evaluation on FER+

Similarly, we show the performance of our proposed methods on the FER+ dataset in Table 4. We train these two

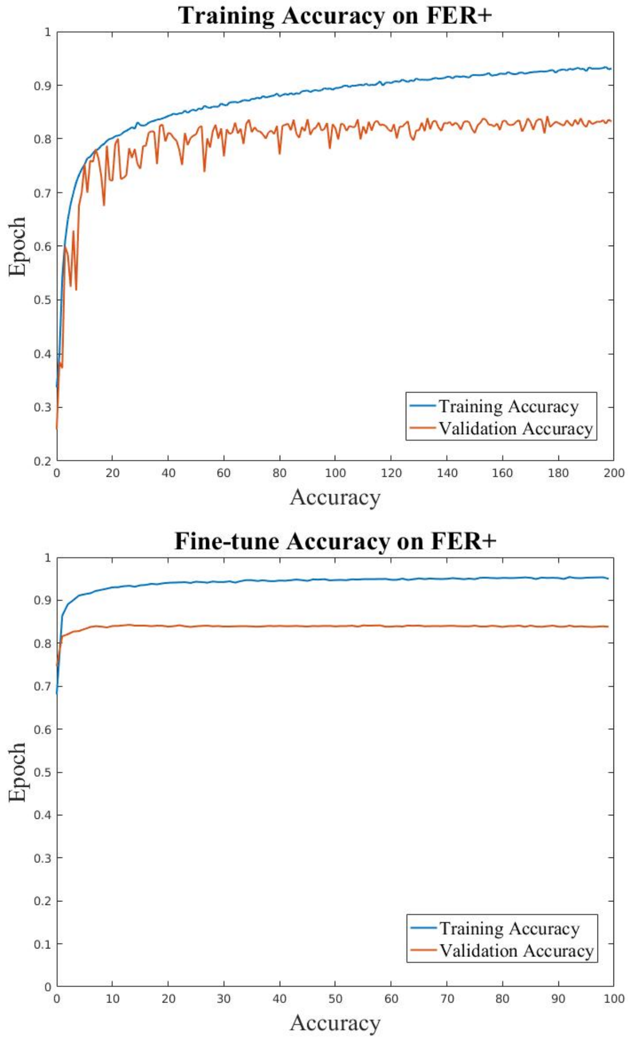


Fig. 6. Training and validation performance on the FER+ dataset during the training and the joint learning stage.

TABLE 4
Accuracy on FER+ of our proposed methods

Methods	Accuracy on FER+
ResNet18	83.09%
TFE-JL	**84.29%**

models on the training set and evaluate them on the private test set. We mainly focus on the private test set for direct comparison with VGG13(MV) proposed by [23]. Specifically, we train our ResNet18 on the training set with 200 epochs and the pre-trained accuracy is 83.09% on the private test set. Then we jointly learn the network using TFE features generated from DeepID and ResNet, getting an improvement up to 1.2% from 83.09% to 84.29%.

In Table 5, we compare our methods with the state-of-the-art approaches on FER2013 and FER+. The work DLSVM-L2 has been presented in [7] and ranks the top in the ICML2013 Representation Learning Challenge on the FER2013 dataset. The VGG13(MV) system outperforms the DLSVM-L2 baseline as it used new labels on FER+. Therefore, we just compare our proposed TFE-JL method with VGG13(MV). As shown in Table 5, the proposed TFE-JL method also achieves 0.5% accuracy gain compared with the average performance of VGG13(MV) model, which again shows the advantage and effectiveness of the proposed identity and emotion joint learning framework.

TABLE 5
Accuracy on FER2013 with old and new labels compared with the state-of-the-art methods

Labels	Methods	Accuracy on FER2013
Old	DLSVM-L2 [7]	71.2%
FER2013	Zhou et al. [22]	69.267%
	Maxim Milakov [22]	68.821%
	Radu+Marius+Cristi [22]	67.484%
FER+	Our implementation of [15]	71.1%
New	VGG13(MV) [23]	83.852%
FER+	TFE-JL	**84.29%**

We also implement the single network transfer learning method in [15] by directly using the deep-id identity feature learned from CASIA-WebFace as inputs for the subsequent SVM modeling on FER+ database. Results in Table 5 show that our proposed joint learning method also outperforms the single network transfer learning approach.

Furthermore, considering the problem that the face images in FER2013 or FER+ dataset are not aligned as well as in CASIA-WebFace used for pre-training the DeepID network, the gain of our joint learning approach on FER+ may not be as large as in the CK+ database.

In order to demonstrate how our proposed identity and emotion joint learning method improves the FER performance, we list the output scores of 8 facial expression categories on ten representative face images from the test set in Table 6. These face images are misclassified using our baseline (ResNet) model but are corrected by the proposed identity and emotion joint learning method. As shown in Fig. 7, these images were firstly misclassified, the reason might be their appearances. For example, image (5) looks ”sad” and image (7-8) look ”angry”. By adding their identity information as an auxiliary input to our FER model, the TFE joint learning approach could reduce the inter-subject variability.

5 Conclusion and Future Work

In this work, we learn both the emotion and identity features using two separate deep convolutional neural networks and construct a deep-learned Tandem Facial Expression (TFE) feature by feature level concatenation. We perform fine-tuning on the newly merged model instead of just transferring the weights of the pre-trained face recognition networks. Experimental results show that the proposed approach outperforms the residual network baseline as well as many other state-of-the-art methods on two popular FER databases, namely CK+ and FER+. Future works include investigating the effect of face image format differences or alignment mismatch between face recognition data and FER data as well as exploring other transfer learning and multi-task learning methods for the FER task.

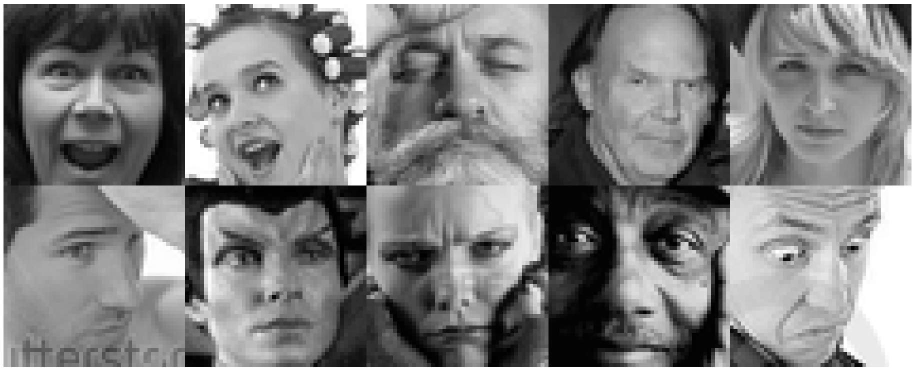


Fig. 7. Ten representative face images whose prediction is corrected by our joint learning method. These face images are indexed with number 1 to 10 from left to right, top to bottom.

TABLE 6
The predicted scores of those representative images in Fig. 7 using ResNet and our TFE joint learning method. For each face image, the first line is the output score using the ResNet baseline and the second line is the output score of our TFE joint learning method.

Image	Neutral	Happy	Surprise	Sad	Angry	Disgust	Fear	Contempt
1	0.0	0.0093	**0.9905**	0.0767	0.0	0.0001	0.0001	0.0
	0.0	**0.6180**	0.3726	0.0	0.0005	0.0012	0.00072	0.0004
2	0.0023	0.3341	**0.6633**	0.0	0.0001	0.0	0.0	0.0001
	0.0005	**0.6191**	0.3493	0.0	0.0200	0.0076	0.0032	0.0002
3	0.4771	0.0	0.0	**0.5229**	0.0	0.0	0.0	0.0
	0.7262	0.0	0.0	0.2735	0.0	0.0002	0.0	0.0
4	0.3601	**0.6396**	0.0	0.0	0.0	0.0	0.0	0.0003
	0.6632	0.3219	0.0002	0.0001	0.0058	0.0005	0.0003	0.0081
5	0.4883	0.0	0.0	**0.5097**	0.0020	0.0001	0.0	0.0
	0.7153	0.0	0.0	0.2837	0.0003	0.0005	0.0001	0.0002
6	0.4172	0.0	0.0	**0.5798**	0.0002	0.0006	0.0	0.0022
	0.7978	0.0	0.0	0.1997	0.0001	0.0013	0.0001	0.0009
7	0.4940	0.0	0.0	0.0019	**0.5037**	0.0004	0.0	0.0
	0.5742	0.0	0.0006	0.0047	0.4027	0.0097	0.0017	0.0063
8	0.0059	0.0	0.0	0.1972	**0.7967**	0.0	0.0	0.0001
	0.0095	0.0	0.0	**0.7818**	0.2000	0.0017	0.0049	0.0020
9	0.3260	0.0007	0.2076	0.0275	0.0016	0.0020	**0.4337**	0.0010
	0.7291	0.0081	0.0612	0.0347	0.0397	0.0683	0.0420	0.0168
10	0.0983	0.0	0.3654	0.0006	0.0096	0.0	**0.5260**	0.0002
	0.0560	0.0	**0.5034**	0.0046	0.0176	0.0106	0.4046	0.0031

Acknowledgments

This research was funded in part by the National Natural Science Foundation of China (61401524,61773413), Natural Science Foundation of Guangzhou City (201707010363), Science and Technology Development Foundation of Guangdong Province (2017B090901045), National Key Research and Development Program (2016YFC0103905).

References

[1] H. Jung, S. Lee, J. Yim, and S. Park, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 2983–2991.

[2] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 425–442.

[3] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding*, 2014, pp. 55–59.

[4] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1988–1996, 2014.

[5] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *Computer Science*, 2015.

[6] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[7] Y. Tang, "Deep learning using support vector machines," *ICML Workshop on Representational Learning*, 2013.

[8] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1997–2009, 2016.

[9] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[10] X. Feng, M. Pietikinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Computer Engineering and Applications*, vol. 17, no. 4, pp. 592–598, 2007.

[11] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition." *IEEE Transactions on Image Processing*, vol. 11, no. 4, p. 467, 2002.

[12] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," *Computer Science*, 2014.

[15] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional

networks,” in *International Conference on Natural Computation*, 2016, pp. 702–708.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 770–778.

[17] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *The Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[18] L. Ma and K. Khorasani, “Facial expression recognition using constructive feedforward neural networks,” *IEEE Transactions on Systems Man Cybernetics Part B Cybernetics*, vol. 34, no. 3, p. 1588, 2004.

[19] S. J. Wang, H. L. Chen, W. J. Yan, Y. H. Chen, and X. Fu, “Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine,” *Neural Processing Letters*, vol. 39, no. 1, pp. 25–43, 2014.

[20] D. Ghimire and J. Lee, “Extreme learning machine ensemble using bagging for facial expression recognition,” *Journal of Information Processing Systems*, vol. 10, no. 3, p. 443 458, 2014.

[21] Z. Yu and C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” in *ACM on International Conference on Multimodal Interaction*, 2015, pp. 435–442.

[22] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. H. Lee, “Challenges in representation learning: A report on three machine learning contests,” *Neural Networks*, vol. 64, p. 59, 2015.

[23] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” in *ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 2, p. 2012, 2012.

[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Computer Science*, 2014.

[26] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[30] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *International Conference on Machine Learning*, 2010, pp. 807–814.

[31] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, 2002, p. 46.

[32] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[33] G. B. Huang, M. A. Mattar, H. Lee, and E. Learned-Miller, “Learning to align from scratch,” *Advances in Neural Information Processing Systems*, pp. 764–772, 2012.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, “Handwritten digit recognition with a back-propagation network,” in *Advances in Neural Information Processing Systems*, 1990, pp. 396–404.

[36] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *International Conference on Machine Learning*, 2013, pp. 1139–1147.

[37] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

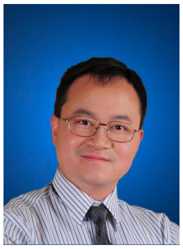
Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in May 2013. He is currently an associate professor at School of Electronics and Information Technology, Sun Yat-Sen University and an associate professor at Data Science Research Center, Duke Kunshan University. His research interests are in the areas of speech processing and multimodal behavior signal analysis with applications to human centered behavioral informatics notably in health, education and security. Works co-authored with his colleagues have won awards at Body Computing Slam Contest 2009, IEEE DCOSS 2009, Interspeech2011 Speaker State Challenge, Interspeech2012 Speaker Trait Challenge, and ISCSLP 2014 best paper award. He received the IBM faculty award at 2016.

Hao Xu is an undergraduate student at School of Data and Computer Science, Sun Yat-sen University. He is a research assistant in Speech and Multimodal Intelligent Information Processing Laboratory at School of Electronics and Information Technology. His research interests include facial expression recognition and gaze recognition.

Xingchang Huang is an undergraduate student at Sun Yat-sen University, major in computer science. He is a research assistant in Speech and Multimodal Intelligent Information Processing Laboratory at School of Electronics and Information Technology. His research interests include machine learning, computer vision and their application on face recognition and facial expression recognition.

Zhanmei Song received the Ph.D. degree in Early Childhood Education(ECE) from East China Normal University in 2012. She is currently a professor at Department of Early Childhood Education of Shandong Yingcai University. Her research interests include Compensation Education for Children Left-behind and Kindergarten Curriculum. She is the General Secretary of the National Association of ECE,Chinese Society of Education.

Xiaolin Liu received a master degree in Early Childhood Education from Northeast China Normal University in 2009,and now studying Doctor Degree of Education Management at Dhurakij Pundit University, Thailand. She is currently an Associate Professor in the School of Preschool Education at Shandong Yingcai University. Her research interests include multimodel technology in early childhood education.



Xin Li received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2005. He is currently a Professor in the School of Preschool Education at Shandong Yingcai University. His research interests include signal processing and data analytics. He is a Fellow of IEEE.