EXPRESSION-IDENTITY FUSION NETWORK FOR FACIAL EXPRESSION RECOGNITION

Haifeng Zhang[1,3] , *Wen Su*[4] , *Zengfu Wang*[1,2,3]

[1] University of Science and Technology of China, Hefei, China.
[2] National Engineering Laboratory for Speech and Language Information Processing, Hefei, China.
[3] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China.
[4] Zhejiang Sci-Tech University, Hangzhou, China

ABSTRACT

Research shows that the facial expression recognition is strongly related to the person's identity. This paper presents an expression-identity fusion network to address the great inter-subject variations in facial expression recognition. The model is designed to jointly learn identity-related features and expression-related features via two branches with the same expression image input. A bilinear module is introduced to fuse two kinds of features and learn the relationship between them. Experimental results show that identity-related features can greatly boost the performance of facial expression recognition. Our method outperforms most of the state-of-the-art. On two popular facial expression databases (CK+ and Oulu-CASIA), our method achieves 96.02% and 85.21% recognition accuracy, respectively.

Index Terms— Facial expression recognition, Expression features, Identity features, Bilinear fusion

1. INTRODUCTION

As a crucial subfeild of face recognition, facial expression recognition has drawn increasing attention from the computer vision community. It makes a wide array of applications ranging from fatigue surveillance, distance learning, human-computer interaction to medical treatment.

Although significant progress has been made for facial expression recognition [1], [2], the current main challenge comes from the great inter-subject variations introduced by age, gender, race and other person-specific characteristics [3]. Specifically, it is difficult to distinguish whether a certain appearance of face is attributed to the identity of a person or his expression. Thus, the influence of identity deserves more attention.

Over the past several years, the relationship between facial expression recognition and face recognition has caught the interests of numerous researchers. There is extensive evidence showing facial expression recognition and face recognition are interdependent. A classic view of Haxby *et al.* [4] points out that expression and identity are a distributed representation in human brain. The expression and identity representations are processed in different cortical regions of brain. Different representations initiate different subsequent neural models. However, the information is interrelated and there is overlap between the two representations. Specifically, when it is difficult to recognize facial expression relying on expression simply, the recognition system will refer to the identity.

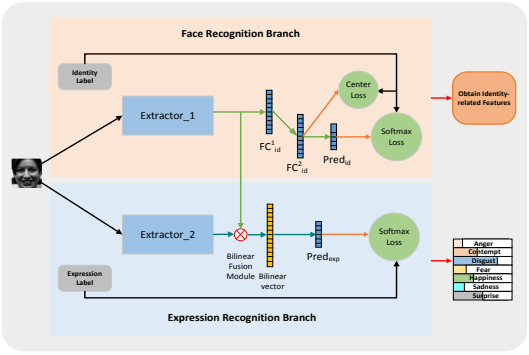

Fig. 1. An overview of our proposed expression-identity fusion network. Best viewed in color.

Recently, some deep learning based facial expression recognition methods take human identity into account. For example, Zhang *et al.* [5] proposed a multi-signal CNN (MSCNN) that takes a pair of facial expression images as input. They use both expression recognition and face verification signals as supervision to calculate different loss. Meng *et al.* [6] proposed an identity-aware CNN (IACNN) that contains two identical sub-CNNs. One stream uses expression-sensitive contrastive loss to learn expression-discriminative features, and the other stream uses identity-sensitive contrastive loss to learn identity-related features for identity-invariant expression recognition. The main idea of these methods is using identity information to increase inter-expression differences and to reduce intra-expression variations. There was scarcely a work to jointly utilize the identity-related features and expression-related features and

This work was supported by the National Natural Science Foundation of China (No.61472393).

to model the relationship between them.

Motivated by the above observations, we propose an expression-identity fusion network (EIFN). To the best of our knowledge, it is the first time the relationship between the identity-related features and expression-related features is established in facial expression recognition tasks. The architecture of EIFN is illustrated in Fig.1. It consists of two branches which are used to extract identity-related features and expression-related features from the input facial expression image respectively. A fusion module aggregates these two kinds of features and models the relationship between them. A special training strategy further enhances the performance of the model. Our method has been evaluated on two popular facial expression databases, namely CK+ [7] and Oulu-CASIA [8]. We demonstrate that our method outperforms most of the state-of-the-art.

2. PROPOSED METHOD

2.1. Face Recognition Branch

The face recognition branch contains an identity-related feature extractor $\mathcal{I}$, two fully connected layers and a softmax loss layer. We feed an image I into extractor $\mathcal{I}$ to obtain the identity-related feature maps F_{id}:

$$F_{id} = \mathcal{I}(I)\,. \tag{1}$$

In this work, we consider a partial network of VGG-16 [9] with initial parameters obtained from pre-trained model on ImageNet [10] truncated at the last convolutional layer ($\mathrm{conv5\_3}$), followed by non-linearity (ReLU) [11] and max pooling as extractor $\mathcal{I}$. The feature maps are passed through two fully connected layers. Finally, the softmax loss is employed to produce the probabilistic distribution over the target subject. We utilize the joint supervision of softmax loss and center loss to train this branch. With a special training strategy, we get the identity-related features which are supervised by identity label and expression label simultaneously. These features are fuse with the following expression recognition branch for facial expression recognition. We will detail the configuration of loss and training strategy in section 2.3.

2.2. Expression Recognition Branch

The expression recognition branch generates expression-related features and combines identity-related features to make the final facial expression classification. It contains an expression-related feature extractor $\mathcal{E}$, a bilinear fusion module and a softmax loss layer. A facial expression image I is fed into extractor $\mathcal{E}$ and then we get the expression-related feature maps F_{exp}:

$$F_{exp} = \mathcal{E}(I)\,. \tag{2}$$

Structure of extractor $\mathcal{E}$ is the same as the extractor $\mathcal{I}$. The outputs of two extractors are aggregated by a bilinear fusion module to obtain a bilinear vector. It is passed through a softmax loss layer to obtain facial expression classification. We optimize this branch by minimizing the softmax loss $\mathcal{L}_{S_{exp}}$. This will be explained in section 2.3.

In our bilinear fusion module, we compute the outer product of two feature maps to fuse them at each location. $\mathcal{I}(l, I)$ is the identity descriptor at location l from identity-related feature maps F_{id}. $\mathcal{E}(l, I)$ is the expression descriptor at location l from expression-related feature maps F_{exp}. The bilinear fusion module is mathematically given as follows:

$$bilinear(l, I, \mathcal{I}, \mathcal{E}) = \mathrm{vec}(\mathcal{I}(l, I) \otimes \mathcal{E}(l, I))\,, \tag{3}$$

$$\begin{aligned} F &= pooling_{l\in\mathcal{L}}\left\{bilinear(l, I, \mathcal{I}, \mathcal{E})\right\} \\ &= \sum_{l\in\mathcal{L}} bilinear(l, I, \mathcal{I}, \mathcal{E})\,. \end{aligned} \tag{4}$$

Here, $\otimes$ calculates the outer product of two vectors and outputs a matrix. $\mathrm{vec}(.)$ turns the matrix into a vector. We aggregate the bilinear features across all spatial locations in the image by the sum-pooling operation. F is the resulting bilinear vector for the input facial expression image. Before F is feeding into the softmax loss layer, it is passed through a signed squared root operation ($y = sign(F)\sqrt{|F|}$) and a L_2-normalization ($z = y/\|y\|_2$).

With the aforementioned operation, our bilinear fusion module can capture pairwise correlations between identity-related feature maps and expression-related feature maps. At each spatial location, every channel of identity-related feature maps interacts with that of expression-related feature maps, thus leading to a fusion of them. For example, the pixels of each channel at the top left corner of identity-related feature maps will be multiplied by the pixels of each channel at the same spatial location on expression-related feature maps. Then a fusion matrix is formed. The summation pooling operation aggregates the fusion matrices at all locations. Each element of the bilinear vector establishes a pixel-level association between identity features and expression features. The weight of each element in the bilinear vector is learned by the network. Elements with large weights have a higher impact in expression recognition tasks. It means that the relationship between identity and expression in the expression recognition task is learned autonomously by the neural network. The bilinear fusion module can also be considered as an implementation of overlapping representations mentioned in the introduction section.

2.3. Loss Function and Training Strategy

We train face recognition branch and facial expression recognition branch simultaneously. The overall loss is defined as:

$$\mathcal{L} = \lambda_2\mathcal{L}_{id} + \lambda_3\mathcal{L}_{S_{exp}} = \lambda_2(\mathcal{L}_{S_{id}} + \lambda_1\mathcal{L}_{C_{id}}) + \lambda_3\mathcal{L}_{S_{exp}}\,. \tag{5}$$

where λ_1, λ_2, λ_3 are the weight of each loss. $\mathcal{L}_{id}$ is supervised by identity label. It represents the identity classification error. $\mathcal{L}_{S_{exp}}$ is supervised by expression label. It represents the expression classification error. $\mathcal{L}_{id}$ is a joint loss which calculated as the weighted sum of the softmax loss $\mathcal{L}_{S_{id}}$ and the center loss $\mathcal{L}_{C_{id}}$. Center loss is a new supervision signal proposed by Wen *et al.* [12] to enhance the discriminative power of the deeply learned features by reducing the intra-class variations for face recognition. Because $\mathcal{L}_{id}$ is affected by identity-related features, the gradients of $\mathcal{L}_{id}$ are backpropagated to face recognition branch. However, our ultimate goal is recognize facial expression. We should focus on minimizing $\mathcal{L}_{S_{exp}}$ and obtain accurate expression classification. $\mathcal{L}_{S_{exp}}$ is affected by both identity-related features and expression-related features. A natural way is to use $\mathcal{L}_{S_{exp}}$ to supervise the extraction of identity-related features and expression-related features simultaneously. Therefore, we utilize a special training strategy. $\mathcal{L}_{S_{exp}}$ are backpropagated to both the expression recognition branch and the identity-related feature extractor $\mathcal{I}$. Note that, two fully connected layers and the softmax loss layer are not included in $\mathcal{I}$ while they are contained in face recognition branch. Such modeling naturally leads to an additional regularization for $\mathcal{I}$. It forces $\mathcal{I}$ to consider both the extraction of identity-related features and the final facial expression recognition task. The final identity-related features are a kind of expression guided identity-related features. They are more suitable for expression recognition tasks than identity-related features that are not supervised by expressions. It has been proven in ablation experiments.

3. EXPERIMENTS

3.1. Databases, Preprocessing and Implementation

We assess the performance of EIFN on two popular facial expression databases: CK+ database [7] and Oulu-CASIA database [8]. The CK+ database contains 327 sequences of 118 subjects. All sequences are annotated with six basic facial expressions (i.e. Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), and Surprise (Su)) and one non-basic expression (Contempt (Co)). The Oulu-CASIA database has 480 sequences collected from 80 subjects. All sequences are labeled with six basic facial expressions. Sequences in two databases begin with a neutral expression and end with a peak expression. As a general procedure [5], [6], [15], [18], [19], the last three frames of each sequence are used for training and testing. Each face region in two databases is aligned, cropped and re-scaled to 64×64. Besides, data augmentation which provides additional training images is employed. Each preprocessed training image is rotated by $[-15°, -10°, -5°, 0°, 5°, 10°, 15°]$ respectively. Horizontally flipping is also applied. The result training set is 14 times larger than the original one. As in the previous method [5], [14], [15], [16], [17], [18], [19], we use the 10-fold subject-independent cross validation strategy. The final results are reported as the average of the 10 runs.

The training of EIFN has two stages. First, two databases are used to mutual initialization the model. When assessing the performance of EIFN on CK+, we pre-train EIFN on Oulu-CASIA. We optimize the overall loss $\mathcal{L}$ using SGD with a batch size of 100, momentum of 0.9. The learning rate and weight decay for EIFN are 0.01, 0.0005 respectively. λ_1, λ_2, λ_3 are set to 0.0001, 1 and 6, respectively. Then, the model is further fine-tuned on the training set of CK+. In this stage, the learning rate for EIFN is starts from 0.01, and is divided by 10 after every 20 epochs. Besides, we change λ_1, λ_2, λ_3 as 0.0001, 2 and 1, respectively. Similarly, when assessing the performance of EIFN on Oulu-CASIA, we firstly pre-train EIFN on CK+. Secondly, we fine-tune the model on the training set of Oulu-CASIA. The differences that λ_1, λ_2, λ_3 in the first step are set to 0.0001, 2 and 1, respectively. In the second stage, they are change to 0.0001, 1, 6, respectively.

3.2. Ablation Experiments

We implement two kinds of ablation experiments. One is a baseline named EFN. It has the same architecture as EIFN. The training strategy is also same, excepting that the λ_2 of both stages is set to 0. It means that EFN do not consider the impact of identity. Compared with baseline, it can reflect the importance of identity information in facial expression recognition. The other is EIFN* which only backpropagate the gradients of $\mathcal{L}_{S_{exp}}$ to the expression recognition branch. This means EIFN* utilize pure identity-related features rather than expression guided identity-related features.

From Table 1 and 2, we can see the results of EFN, EIFN and EIFN* on two databases. Both EIFN and EIFN* perform better than our baseline model EFN. These indicate that the introduction of identity information promotes the performance of facial expression recognition. Besides, comparing EIFN with EIFN*, the performance has been further improved. The effectiveness of our special training strategy has been proven.

3.3. Results and Analysis

We compare our method with other state-of-the-art. The experimental results of EIFN on two databases are shown in Table 1 and 2. EIFN achieves an average recognition accuracy of 96.02% and 85.21% on CK+ and Oulu-CASIA, respectively. Our method outperforms most of the state-of-art on both databases, including image-based methods and sequence-based methods. Our results seem to be a little worse than PHRNN-MSCNN [5], DTAGN [17] and FN2EN [19]. Note that [17] utilizes dynamic geometry-appearance features, [5] utilizes partial-whole, geometry-appearance and dynamic-still features. The inputs of their model are very

Table 1. Average accuracy on CK+ [7] for seven expressions classification.

Method	Setting	Accuracy
3DCNN-DAP [13]	sequence-based	92.40%
STM-ExpLet [14]	sequence-based	94.19%
IL-CNN [15]	image-based	94.35%
LOMo [16]	sequence-based	95.10%
IACNN [6]	image-based	95.37%
MSCNN [5]	image-based	95.54%
DTAGN [17]	sequence-based	97.25%
PHRNN-MSCNN [5]	sequence-based	98.50%
EFN (baseline)	image-based	94.30%
EIFN*	image-based	95.01%
EIFN	image-based	**96.02%**

Table 2. Average accuracy on Oulu-CASIA [8] for six expressions classification.

Method	Setting	Accuracy
STM-ExpLet [14]	sequence-based	74.59%
IL-CNN [15]	image-based	77.29%
MSCNN [5]	image-based	77.69%
DTAGN [17]	sequence-based	81.86%
LOMo [16]	sequence-based	82.10%
PPDN [18]	image-based	84.59%
PHRNN-MSCNN [5]	sequence-based	86.25%
FN2EN [19]	image-based	87.71%
EFN (baseline)	image-based	81.74%
EIFN*	image-based	83.33%
EIFN	image-based	**85.21%**

Table 3. Confusion matrix of EIFN evaluated on CK+ [7].

	An	Co	Di	Fe	Ha	Sa	Su
An	**93.33**	0	0	0	0	6.67	0
Co	0	**88.89**	0	0	0	11.11	0
Di	0	0	**100**	0	0	0	0
Fe	0	4.00	0	**92.00**	0	0	4.00
Ha	0	0	0	0	**100**	0	0
Sa	10.71	0	0	3.57	0	**85.72**	0
Su	0	2.41	0	0	0	0	**97.59**

Table 4. Confusion matrix of EIFN evaluated on Oulu-CASIA [8].

	An	Di	Fe	Ha	Sa	Su
An	**82.08**	3.75	2.50	0	11.67	0
Di	12.50	**76.25**	1.25	0	10.00	0
Fe	2.50	1.25	**85.00**	3.75	1.25	6.25
Ha	0	0	6.25	**92.92**	0.83	0
Sa	11.67	3.75	0	2.08	**82.50**	0
Su	0	0	7.50	0	0	**92.50**

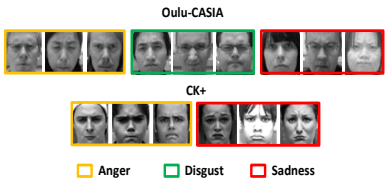

Fig. 2. Some examples on CK+ and Oulu-CASIA.

complex. While our method needs only static appearance features, which is more favorable for online applications or snapshots where per frame labels are preferred. [19] fails to achieve end-to-end optimization, while our model is an end-to-end approach. Furthermore, our method outperforms MSCNN [5] and IACNN [6], which also utilized the identity information.

The confusion matrices of EIFN evaluated on two databases are reported in Table 3 and 4. Our method performs well in terms of fear, happiness and surprise. However, the performances on contempt, sadness of CK+ and disgust, anger, sadness of Oulu-CASIA are relatively poor. The poor performance of recognizing contempt is mainly due to the lack of contempt image data. There are only 18 of 327. In particular, on both CK+ and Oulu-CASIA, sadness and anger are confused. On Oulu-CASIA, disgust, sadness and anger are confused. Fig.2 shows some examples of high degree similarity of anger, disgust and sadness of CK+ and Oulu-CASIA.

4. CONCLUSIONS

In this paper, we present a two branch expression-identity fusion network for facial expression recognition. Identity-related features and expression-related features are fused and the relationship between them is learned. A special training strategy further improved the performance of the model. Our proposed method was evaluated on two popular facial expression databases (CK+ and Oulu-CASIA).

In summary, with only static appearance information, EIFN outperforms most of the state-of-the-art, and even some methods utilize the spatio-temporal and geometric information. Furthermore, we show that the using of identity-related features greatly boosted the performance of facial expression recognition.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No.61472393).

6. REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, no. 1, pp. 39-58, 2009.

[2] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, no. 6, pp. 1113-1133, 2015.

[3] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Metaanalysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 42, no. 4, pp. 966-979, 2012.

[4] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science* vol. 293, no. 5539 pp. 2425-2430, 2001.

[5] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing,* vol. 26 no. 9, pp. 4193-4203, 2017.

[6] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, "Identity-aware convolutional neural network for facial expression recognition, in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG),* 2017, pp. 558-565.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW),* 2010, pp. 94-101.

[8] G. Zhao, X. Huang, M. Taini, S.Z. Li, M. PietikaInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing.* vol. 29, no. 9, pp. 607-619, 2011

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556, 2014.*

[10] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2009, pp. 248-255.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in Advances in neural information processing systems, 2012, pp. 1097-1105.

[12] Y. Wen, K. Zhang, Z. Li, Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV),* 2016, pp. 499-515.

[13] M. Liu, S. Li, S. Shan, R. Wang, X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proceedings of the IEEE Asian Conference on Computer Vision (ACCV),* 2014, pp. 143-157.

[14] M. Liu, S. Shan, R. Wang, X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2014, pp. 1749-1756.

[15] J. Cai, Z. Meng, A. S. Khan, Z. Li, Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG),* 2018 pp. 302-309.

[16] K. Sikka, S G. harma, M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016, pp. 5580-5589.

[17] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2015, pp. 2983-2991.

[18] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, S. Yan, Peak-piloted deep network for facial expression recognition, in: Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2016 pp. 425-442.

[19] H. Ding, S. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG),* 2017 pp. 118-126.