SCIENCE CHINA
Information Sciences

February 2020, Vol. 63 120114:1–120114:3
https://doi.org/10.1007/s11432-019-2747-y

• LETTER •
Special Focus on Deep Learning for Computer Vision

Leveraging 3D blendshape for facial expression recognition using CNN

Sa WANG[1], Zhengxin CHENG[1], Xiaoming DENG[2], Liang CHANG[1*],
Fuqing DUAN[1*] & Ke LU[3]

[1]*College of Artificial Intelligence, Beijing Normal University, Beijing* 100875*, China;*
[2]*Bejing Key Laboratory of Human-Computer Interactions, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 1 October 2019/Revised 8 December 2019/Accepted 13 December 2019/Published online 16 January 2020

Citation Wang S, Cheng Z X, Deng X M, et al. Leveraging 3D blendshape for facial expression recognition using CNN. Sci China Inf Sci, 2020, 63(2): 120114, https://doi.org/10.1007/s11432-019-2747-y

Dear editor,

Recently, facial expression recognition (FER) has drawn increasing attention owing to its widespread application in many fields such as human-machine interaction, assistive healthcare, psychology, and online education. Although intensive research and significant progress have been made over the past two decades, FER remains a challenging task owing to the substantial variations caused by illumination, pose, occlusion, as well as an individual's age, gender, race, and other factors. There have been several methods to solve the FER problem [1]. Previous methods mainly represent facial changes using hand-crafted features [2]. However, hand-crafted features cannot effectively capture the large variability in morphological factors and facial movements. Therefore, they achieve limited performance in expression detection. Recently, convolutional neural networks (CNN)-based approaches [3–6] have been proposed. The CNN-based FER methods have a strong capacity for feature learning and representation, and it can overcome the weakness of the hand-crafted features, thus making it achieve state-of-the-art performance.

We propose a CNN-based FER method that leverages 3D blendshape (Figure 1). Originally, blendshape is a computer animation technique based on blending the neutral face and different expressions that support effective facial action representation. Therefore, we use the blendshape coefficients to enhance the performance of FER. BlendshapeFERNet, a two-stream network, is proposed to analyze both image expression information and 3D facial muscle actions. The first stream extracts expression-aware features from images (image-based FER network) and the second stream extracts features aware of facial muscle movements (blendshape regression network). Our method can exploit both the image information and the dynamic information implied in the blendshape. To get the blendshape coefficients, we construct a blendshape dataset (BlendshapeExp) that includes 49 subjects and about 100 sequences using the Faceshift Studio software. The details of BlendshapeExp can be found in Appendix A. Based on this database, we train the blendshape network to regress blendshape coefficients on input images. Experiments show that our method can consistently achieve state-of-the-art performance on three typical datasets CK+, Oulu-CASIA, and MMI. Moreover, it also has evident generalization ability on some in-the-wild datasets like RAF-DB.

Face blendshape representation. Face blendshapes are a set of 3D-models of a face where each model has a particular expression. The blendshapes only change the vertex positions because the topologies of all face models are identical. Any

[*] Corresponding author (email: changliang@bnu.edu.cn, fqduan@bnu.edu.cn)

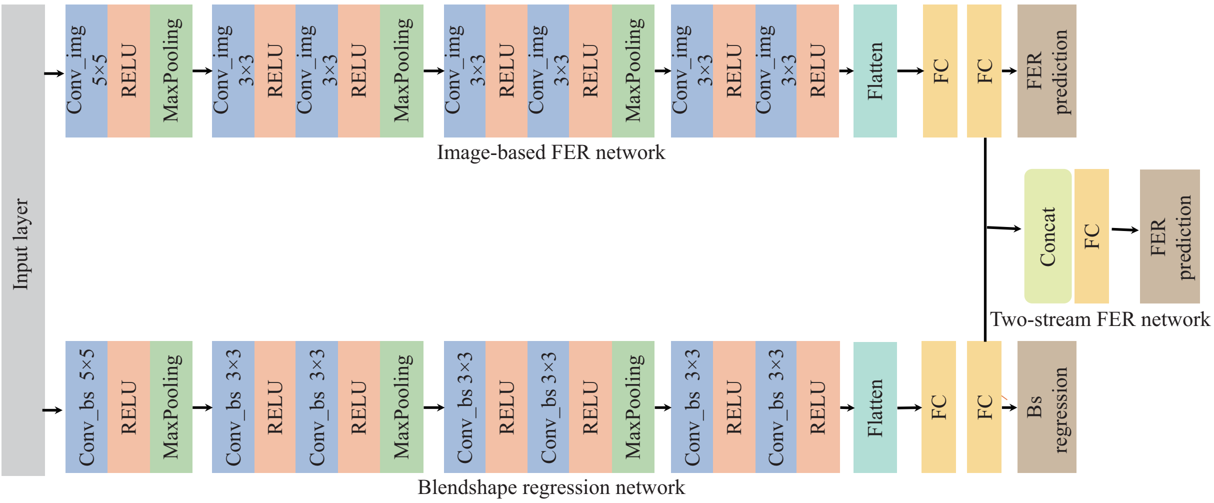

Figure 1 (Color online) Overview of our two-stream FER network, BlendshapeFERNet.

complex expression can be synthesized by blending the neutral face and particular expressions. Each expression $\boldsymbol{T}$ can be generated as

$$\boldsymbol{T} = \boldsymbol{b}_0^e + \sum_{i=1}^{l} \delta_i(\boldsymbol{b}_i^e - \boldsymbol{b}_0^e) = \boldsymbol{b}_0^e + \sum_{i=1}^{l} \delta_i \boldsymbol{d}_i^e, \quad (1)$$

where $\boldsymbol{b}_0^e$ is the neutral face, $\boldsymbol{b}_i^e$ is the i-th blendshape, $\delta_i = \delta_1, \ldots, \delta_l$ is the blending weight of expression $\boldsymbol{T}$, the range of each blendshape weight is $[0, 1]$, and $\boldsymbol{d}_i^e$ is the per-vertex 3D displacement with respect to the neutral face $\boldsymbol{b}_0^e$. Face blendshape representation provides a compact representation of the facial expression space.

Image-based FER network. The image-based FER network analyzes expression from a single image. For the image FER network, we use seven convolution layers and two fully-connected layers followed by a softmax layer as a loss layer for classification as shown in Figure 1. The network is pre-trained using a large face recognition dataset (Webface) and is further fine-tuned on benchmark FER databases. The image FER network can be formulated as follows:

$$\boldsymbol{O}_{\rm img} = h_{\rm img2}(h_{\rm img1}(c_{\rm img}(\boldsymbol{I}))), \quad (2)$$

where $\boldsymbol{I}$ represents the input image, $h_{\rm img1}$ and $h_{\rm img2}$ represent the first and second fully-connected layers of our FER network, respectively, and $c_{\rm img}$ corresponds to the convolution layers. $\boldsymbol{O}_{\rm img}$ is the predicted facial expression label of the input face image. We use a simple cross-entropy loss for the FER network and denote the groundtruth expression labels as D^*. The labels of the training sample are represented as a seven-dimensional vector denoting the target seven facial expression categories. Herein, only the component corresponding to the sample class has a value of 1, and all other components in the vector are 0. The loss function is as follows:

$$L_{\rm image}(\boldsymbol{D}, \boldsymbol{D}^*) = -\frac{1}{n}\sum_{i}\sum_{j} \boldsymbol{D}_{ij}^* \log(\boldsymbol{D}_{ij}), \quad (3)$$

where $\boldsymbol{D}_{ij} = e^{z_j^i} / \sum_{j=0}^{l} e^{z_j^i}$ is the probability prediction that the training sample i belongs to class j, which is computed by the output z of the final fully-connected layer in the image-based FER network. z_j^i denotes the j-th component of the output for the training sample i, and $\boldsymbol{D}_{ij}^*$ is the groundtruth label of the training sample i. n is the number of training examples in a batch.

Blendshape regression network. The blendshape network estimates the blendshape coefficients α_i in (1) using a face image as input. As shown in Figure 1, our blendshape net also contains seven convolutional layers and two fully-connected layers. Our blendshape network can be formulated as follows:

$$\boldsymbol{O}_{\rm bs} = h_{\rm bs2}(h_{\rm bs1}(c_{\rm bs}(\boldsymbol{I}))), \quad (4)$$

where $h_{\rm bs1}$ and $h_{\rm bs2}$ represent the first and second fully-connected layers of our blendshape network, respectively, and $c_{\rm bs}$ corresponds to the convolutional layers. 'bs' denotes blendshape, $\boldsymbol{I}$ is the input face image, and $\boldsymbol{O}_{\rm bs}$ is the input of the loss function. $\boldsymbol{o}_i$ denotes the predicted coefficients for the i-th image in a batch. The loss function is evaluated as follows:

$$\boldsymbol{L}_{\rm bs} = \frac{1}{2n}\sum_{i=1}^{n} \|\boldsymbol{o}_i - \boldsymbol{g}_i\|_2^2, \quad (5)$$

where $\boldsymbol{g}_i$ represents the groundtruth blendshape labels for the i-th image in a batch.

Two-stream FER network. We concatenate the features of the blendshape network and FER network before the last fully-connected layer and then feed the resulting tensor into a sub-network comprising a fully-connected layer and softmax layer for expression classification. By definition, the implemented channel concatenation is $\boldsymbol{F}_{\text{con}} = \{\boldsymbol{F}_{\text{img}}, \boldsymbol{F}_{\text{bs}}\}$. Herein, $\boldsymbol{F}_{\text{img}}$ and $\boldsymbol{F}_{\text{bs}}$ are the outputs of the first fully-connected layer of the FER network and the blendshape network, respectively. The feature maps $\boldsymbol{F}_{\text{img}}$ and $\boldsymbol{F}_{\text{bs}}$ can be formulated as follows:

$$\begin{aligned} \boldsymbol{F}_{\text{img}} &= h_{\text{img1}}(c_{\text{img}}(\boldsymbol{I})),\\ \boldsymbol{F}_{\text{bs}} &= h_{\text{bs}}(c_{\text{bs}}(\boldsymbol{I})). \end{aligned} \tag{6}$$

Finally, our two-stream FER network can be formulated as follows:

$$\boldsymbol{O}_{\text{con}} = h_{\text{con}}(\boldsymbol{F}_{\text{con}}), \tag{7}$$

where $\boldsymbol{O}_{\text{con}}$ is the input of the softmax function for expression recognition. We use a cross-entropy loss for the two-stream FER network that is defined as follows:

$$\boldsymbol{L}_{\text{twostream}}(\hat{D}, \boldsymbol{D}^*) = -\frac{1}{n}\sum_i \sum_j \boldsymbol{D}^*_{ij} \log(\hat{D}_{ij}), \tag{8}$$

where $\hat{D}_{ij}$ is the probability prediction that the training sample i belongs to class j which is computed by the output $\hat{z}$ of the final fully-connected layer in the two-stream FER network. $\boldsymbol{D}^*_{ij} \in \{0.0, 1.0\}$ is the groundtruth probability that the training sample i belongs to class j and n is the training sample number in a batch. The implementation details can be found in Appendix B.

Experiments. For the three datasets CK+, Oulu-CASIA, and MMI, we use the standard tenfold cross-validation for evaluation in all the experiments. For the RAF-DB dataset, we evaluate the performance of the method only on the holdout test set. Our BlendshapeFERNet method obtains the best accuracy on all three datasets and exceeds the previous best result of DeRL [4] on Oulu-CASIA by 0.18%. Moreover, our method exceeds state-of-the-art methods FN2EN [3] on CK+ and the MSCNN [5] on MMI by 0.7% and 2.2%, respectively. Experiments also show that leveraging the 3D blendshape can indeed boost the baseline image-based FER network by about 0.5%. The experimental details can be found in Appendix C.

Conclusion. We proposed the use of a 3D blendshape to enhance the performance of FER. Experiments using three publicly available datasets, CK+, Oulu-CASIA, and MMI, validate that leveraging blendshape enhances the performance of the image-based FER method, and our two-stream BlendshapeFERNet method outperforms the state-of-the-art methods. Although a significant performance boost was achieved, several issues remain open to be addressed: the recognition of facial expressions of low-intensity, exploiting blendshape to further improve the performance of real-world or video-based FER, and multi-task learning for FER and blendshape regression. In the future, we will expand our dataset and extend our framework for such application scenarios.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61572078, 61473276) and Beijing Natural Science Foundation (Grant No. L182052).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Li S, Deng W. Deep facial expression recognition: a survey. 2018. ArXiv: 180408348

2 Rahulamathavan Y, Phan R C W, Chambers J A, et al. Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. IEEE Trans Affective Comput, 2013, 4: 83–92

3 Ding H, Zhou S K, Chellappa R. Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition, 2017. 118–126

4 Yang H, Ciftci U, Yin L. Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 2168–2177

5 Zhang K, Huang Y, Du Y, et al. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process, 2017, 26: 4193–4203

6 Acharya D, Huang Z, Paudel D P, et al. Covariance pooling for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 367–374