DISENTANGLED FEATURE BASED ADVERSARIAL LEARNING FOR FACIAL EXPRESSION RECOGNITION

Mengchao Bai[1,2], *Weicheng Xie*[1,2] *and Linlin Shen*[1,2] ✉

[1]School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, P.R. China
[2]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, P.R. China
baimengchao2017@email.szu.edu.cn, {wcxie, llshen}@szu.edu.cn

ABSTRACT

A facial expression image can be considered as an addition of expressive component to a neutral expression face. With this in mind, in this paper, we propose a novel end-to-end adversarial disentangled feature learning (ADFL) framework for facial expression recognition. The ADFL framework is mainly composed of three branches: expression disentangling branch ADFL-d, neutral expression branch ADFL-n and residual expression branch ADFL-r. The ADFL-d and ADFL-n aim to extract the expressive component and neutral component, respectively. The ADFL-r extracts the residual expression by calculating the difference between feature maps of ADFL-d and ADFL-n, and uses the residual expression feature for expression classification. Experimental results on several benchmark databases (CK+, MMI and Oulu-CASIA) show that the proposed method has remarkable performance compared to state-of-the-art methods.

Index Terms— Disentangled feature, adversarial learning, expression disentangling, residual expression

1. INTRODUCTION

Facial expression recognition (FER) is one of the most widely studied topics in computer vision due to its wide applications in human-computer interaction, medical treatment and driver fatigue surveillance, etc. Existing FER methods in the literature can be grouped into two categories in the light of their feature extraction methods: hand-crafted features based and deep learning based methods. The hand-crafted features based methods usually extract representative expression features, such as 3D SIFT [1], LBP-TOP [2] and Gabor [3], etc. The extracted features are then used to classify facial expressions by Support Vector Machine (SVM) [4] or Nearest Neighbor classifier. Since the extraction of hand-crafted features is separated from the training of classifier, some useful facial expression information may be lost. So these hand-crafted features based methods achieve limited performance.

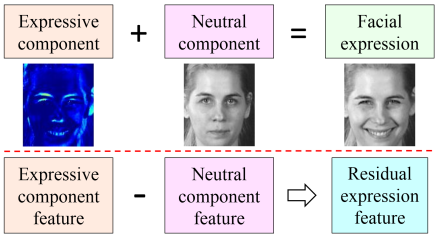

Fig. 1. Illustration of our method.

However, with the increasing computing power and emergence of large-scale database like FER2013 [5], deep learning based methods (e.g. 3DCNN [6], IACNN [7] and DTAGN [8]) are now widely adopted by scholars in the FER field. These algorithms automatically learn the expression features and train the classifier simultaneously in an end-to-end way. These methods performed better than hand-crafted features based methods. However, their capacities are still limited because of the similarities among the expressions of different categories, which may affect the performance of expression classification.

To tackle this problem, some scholars have considered that when an individual reveals a facial expression, a human may have an experience to compare their expression with other expressions observed in past to find out the expression differences [9]. It's widely believed that facial expression features can be extracted by comparing the differences between a given image and the reference image (such as neutral face image). In [10], a De-expression Residue Learning (DeRL) method was proposed, which employed cGAN [11] to disentangle the facial expression feature from a given image. Firstly, it uses a generative model to generate the corresponding neutral expression face image for any input face image, in the meantime, the expressive information is recorded in the intermediate layers. Then the expressive information is extracted from each intermediate layer and concatenated for facial expression recognition with the softmax classifier. Since

The work is supported by National Natural Science Foundation of China (Grant No. 61672357, U1713214 and 61602315), the Science and Technology Project of Guangdong Province (Grant No. 2018A050501014) and the Science and Technology Innovation Commission of Shenzhen (Grant No. JCYJ20170302153827712).

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 09:03:45 UTC from IEEE Xplore. Restrictions apply.

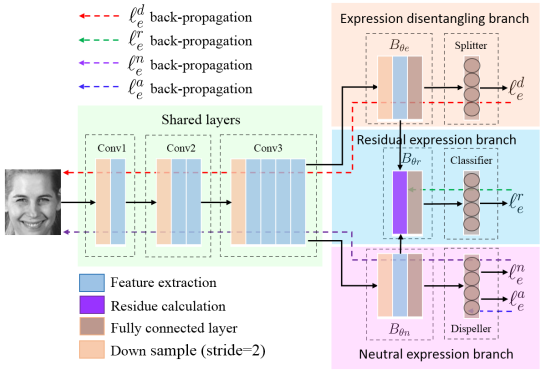

Fig. 2. The framework of adversarial disentangled feature learning.

the DeRL method contains two stages, the performance of the generative model in the first stage has a great impact on that of the FER system in the second stage. Liu *et al.* [12] proposed a distilling and dispelling auto-encoder (D^2AE) framework to perform face editing, which used encoder to extract the identity information and the complementary facial information to represent the whole facial information by adversarial leaning.

In this paper, inspired by the success of the DeRL [10] and the D^2AE [12], we proposed an end-to-end adversarial disentangled feature learning (ADFL) framework for facial expression recognition. As shown in Fig. 1, a facial expression image can be disentangled to the combination of expressive component and neutral component. The residual expression feature can be calculated by the subtraction of these features. The ADFL framework is mainly composed of three branches: expression disentangling branch, neutral expression branch and residual expression branch. They are expected to extract the expressive component, neutral component and residual expression, respectively. The proposed ADFL framework aims to disentangle the divergence in expression relative to neutral expression. The paper makes the following contributions.

(1) We proposed a novel end-to-end ADFL framework to disentangle residual expression feature by adversarial learning.

(2) The adversarial learning of expression disentangling branch and neutral expression branch ensures the effective decomposition of expressive component and neutral component.

(3) The residual expression branch automatically extract discriminate residual expression feature, which achieves competitive performance in several benchmark databases.

2. PROPOSED METHOD

In this section, we introduce the proposed ADFL framework. As shown in Fig. 2, the whole framework is composed of four parts, the shared layers S_θ, residual expression branch ADFL-r and two parallel branches: expression disentangling branch ADFL-d and neutral expression branch ADFL-n. Given a face image x, abundant face information $S_\theta(x)$ is extracted by the shared layers S_θ. Then, $S_\theta(x)$ is fed into expression disentangling branch ADFL-d and neutral expression branch ADFL-n to further disentangle expressive component and neutral component , respectively. Finally, the expressive component feature $\boldsymbol{f}_e \in R^{N_e}$ and neutral component feature $\boldsymbol{f}_n \in R^{N_n}$ are fed into residual expression branch to extract residual expression feature for facial expression recognition.

Table 1. Architectures of the proposed ADFL framework.

Components	Layers	Configurations
Shared layers	Conv1	$[3\times3, 32]\times1$, S-2 $[3\times3, 32; 3\times3, 32]\times1$
	Conv2	$[3\times3, 64]\times1$, S-2 $[3\times3, 64; 3\times3, 64]\times2$
	Conv3	$[3\times3, 128]\times1$, S-2 $[3\times3, 128; 3\times3, 128]\times4$
Expression disentangling branch	$B_{\theta e}$	$[3\times3, 256]\times1$, S-2 $[3\times3, 256; 3\times3, 256]\times1$ FC-256
	Splitter	#Expression Category - 1
Neutral expression branch	$B_{\theta n}$	$[3\times3, 256]\times1$, S-2 $[3\times3, 256; 3\times3, 256]\times1$ FC-256
	Dispeller	#Expression Category
Residual expression branch	$B_{\theta r}$	Residue calculation FC-256
	Classifier	#Expression Category - 1

2.1. Main framework

Adapted from SpherefaceNet-20 [13], the architecture of our framework is illustrated in Table 1. Conv1, Conv2 and Conv3 denote convolutional blocks that contain multiple convolutional layers and residual units are shown in double-column brackets. For example, $[3\times3, 64]\times2$ denotes two cascaded convolution layers with 64 filters of size 3×3, and S-2 denotes stride 2 in the down sample layer. Each convolutional layer is followed by a batch normalization layer and a PReLU [14] layer. FC-256 denotes a fully connected layer with 256 neurons. Residue calculation compute the channel-wise difference between the feature maps from expression disentangling branch and the feature maps from neutral expression branch.

2.2. Expression disentangling branch

As revealed in Fig. 2, the expression disentangling branch extract expressive component information $\boldsymbol{f}_e$ by the subnet $B_{\theta e}$.

$$\boldsymbol{f}_e = B_{\theta e}(S_\theta(x)) \tag{1}$$

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 09:03:45 UTC from IEEE Xplore. Restrictions apply.

Then, $\boldsymbol{f}_e$ is non-linearly mapped by softmax function defined as $\boldsymbol{y}_e = softmax(\boldsymbol{W}_e\boldsymbol{f}_e + \boldsymbol{b}_e)$, where $\boldsymbol{y}_e \in R^{N_e}$ is an N_e-dimensional vector and represents the probabilities of belonging to the corresponding class. The loss ℓ_e^d is computed by the probability vector $\boldsymbol{y}_e \in R^{N_e}$.

$$\ell_e^d = \begin{cases} 0 & i = c \\ -\log \boldsymbol{y_e^i} & i \neq c \end{cases} \tag{2}$$

Where i denotes the ground truth index and c denotes the neutral ground truth index. Only the losses of non-neutral expression categories will be back-propagated in expression disentangling branch, which ensures the effective extraction of expressive component feature. The back-propagation route of optimization over ℓ_e^d including the expression disentangling branch and the shared layers is indicated with the red dotted arrow in Fig. 2.

2.3. Neutral expression branch

Similar to the ADFL-d, the structure of neutral expression branch ADFL-n is composed of a subnet $B_{\theta n}$ and an expressive component dispeller. The ADFL-n suppresses expressive component feature and extract the neutral component feature $\boldsymbol{f}_n = B_{\theta n}(S_\theta(x))$ by the subnet $B_{\theta n}$ following the shared layers. To enable the extraction of neutral component feature, an adversarial supervised training method composed of two different loss functions ℓ_e^a and ℓ_e^n is employed.

The cross entropy loss $\ell_e^a = -\log \boldsymbol{y}_n^i$ is used to supervise the training of the expressive component dispeller based on $\boldsymbol{y}_n$, which is computed by

$$\boldsymbol{y}_n = softmax(\boldsymbol{W}_n\boldsymbol{f}_n + \boldsymbol{b}_n) \tag{3}$$

Different from ℓ_e^d, the gradient of ℓ_e^a is only back-propagated to the expressive component dispeller and the previous layers are not updated.

We use ℓ_e^n to fool the training of expressive component dispeller $\boldsymbol{y}_n$. In order to achieve this, ℓ_e^n is required to be constant over all expressions and equal to $\frac{1}{N}$. Thus, the optimization goal is equivalent to minimize the negative entropy of the predicted expression distributions,

$$\ell_e^n = \begin{cases} 0 & i = c \\ -\frac{1}{N}\sum\limits_{i}^{N} \log \boldsymbol{y}_p^i & i \neq c \end{cases} \tag{4}$$

where N denotes the number of expression categories, i denotes the ground truth index and c denotes the neutral ground truth index. The optimization over ℓ_e^n updates the neutral expression branch and the shared layers.

The total loss of the ADFL-n is the summation of ℓ_e^a and ℓ_e^n, which ensure the effective extraction of neutral component feature by adversarial learning.

2.4. Residual expression branch

The residual expression branch contains a subset $B_{\theta r}$ and an expression classifier, and aims to extract residual expression feature for expression classification. As illustrated in Fig. 2, the subnet $B_{\theta r}$ is composed of a residue calculation module (purple box) and a fully connected layer. The residue feature $\boldsymbol{f}_r$ is the subtraction of the feature maps in front of the fully connected layer in ADFL-d and ADFL-n, as calculated in equation (5).

$$\boldsymbol{f}_r = \boldsymbol{f}_e - \boldsymbol{f}_n \tag{5}$$

The residue feature $\boldsymbol{f}_r$ is further mapped by a non-linear function *Additive Margin Softmax* (AMS) [15], which is defined as equation (6):

$$\boldsymbol{y}_r = \frac{e^{s(\boldsymbol{W}_{y_r}^T\boldsymbol{f}_r - m)}}{e^{s(\boldsymbol{W}_{y_r}^T\boldsymbol{f}_r - m)} + \sum_{j=1, j\neq y_r}^{c} e^{s\boldsymbol{W}_j^T\boldsymbol{f}_r}} \tag{6}$$

where m and s are two hyper-parameters of the additive margin softmax which denote the margin among categories and scaling factor respectively, $\boldsymbol{y}_r \in R^{N_r}$ is an N_r-dimensional vector, which denotes the predicted probabilities of belonging to the corresponding class. The probability vector $\boldsymbol{y}_r$ is further employed to calculate the classification loss ℓ_e^r, where i denotes the ground truth index.

$$\ell_e^r = \begin{cases} 0 & i = c \\ -\log \boldsymbol{y_r^i} & i \neq c \end{cases} \tag{7}$$

Since the residue calculation module does not contain hyperparameters, the optimization over ℓ_e^r only updates the classifier and the fully connected layer in ADFL-r, as depicted with the green dotted arrow in Fig. 2. Similar to ℓ_e^d, only the losses of non-neutral expression categories are back-propagated in ADFL-r.

2.5. Objective function

The ADFL-d and ADFL-n are used to extract expressive component and neutral component, respectively, and are not used to predict the expression categories. Only the residual expression feature extracted by the ADFL-r is used for expression classification. In order to achieve this, the ADFL framework is jointly optimized by four loss functions ℓ_e^d, ℓ_e^a, ℓ_e^n and ℓ_e^r. The total loss function $\mathcal{L}$ is the weighted sum of ℓ_e^d, ℓ_e^a, ℓ_e^n and ℓ_e^r, as formulated in equation (8).

$$\mathcal{L} = \lambda_d\ell_e^d + \lambda_n(\ell_e^a + \ell_e^n) + \lambda_r\ell_e^r \tag{8}$$

3. EXPERIMENTAL RESULTS

In this section, we firstly describe the experimental settings and three publicly available expression databases (CK+ [16], MMI [17] and Oulu-CASIA [18]). Then we compare the performance of the proposed method with the state-of-the-art methods.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 09:03:45 UTC from IEEE Xplore. Restrictions apply.

3.1. Implementation details

Data preprocessing. For each database, we detect five facial landmark points by MTCNN [19] and align the face images to the size of 128×110 according to their facial landmarks. Then, ten faces with size of 112×96 are cropped from four corners and center of each aligned image and its horizontal flipping mirror.

Hyperparameter settings. The proposed ADFL framework is optimized using Adam optimizer [20] with betas of 0.9 and 0.999, ϵ of $1e-8$ and weight decay of 0.0005. The optimization is performed about 100 epochs with a batch size of 64 and an initial learning rate of $1e-4$. For objective function, we set $m = 0.35$, $s = 30$, $\lambda_d = 1$, $\lambda_n = 10$ and $\lambda_r = 1$.

Baseline. In order to prove the effectiveness of the proposed ADFL framework, we use a baseline network for comparison. The structure of the baseline network is similar to the cascade of the shared layers and the expression disentangling branch, as depicted in Fig. 2. But the splitter in expression disentangling branch is replaced by the classifier in residual expression branch. Note that, compared to the ADFL framework, the baseline network only updates parameters with the supervision of AMS [15] loss function rather than adversarial learning.

3.2. Databases and protocols

The Extended Cohn-Kanade database (CK+) [16] is a representative laboratory-controlled database for facial expression recognition. It contains 593 video sequences from 123 subjects. Following the 10-fold cross validation protocol in [10], the last three frames with provided label and the first frame with neutral label of each labeled sequence are selected and all subjects are divided into ten groups by their ID in an ascending order.

The MMI database [17] consists of 236 sequences from 32 subjects with six basic expressions. We selected 209 sequences captured in front view. We selected three frames in the middle of each sequence with provided label and the first frame as neutral expression image, and employed a 10-fold cross validation.

The Oulu-CASIA database [18] In our experiments, the Oulu-CASIA VIS database under strong illumination condition is used, which includes 480 image sequences from 80 subjects labeled with six basic expressions (anger, disgust, fear, happiness, sadness and surprise). Similar to the CK+ database, the last three frames and the first frame of each sequence are selected and a 10-fold cross validation is adopted.

3.3. Results

Table 2 lists the results of the proposed method, baseline and other approaches in literature on CK+, MMI and Oulu-CASIA databases.

Table 2. Overall accuracy on the CK+, MMI and Oulu-CASIA databases. "-" denotes that there is no corresponding results.

Method	Accuracy (%)		
	CK+	MMI	Oulu-CASIA
LBP-TOP [2]	88.99	59.51	68.13
3DCNN [6]	85.90	53.20	-
STM-Explet [21]	94.19	75.12	74.59
Zeng *et al.* [22]	97.35	-	-
IACNN [7]	95.37	71.55	-
DTAGN-Joint [8]	97.25	70.24	81.46
DeRL [10]	97.30	73.23	**88.00**
Baseline	94.19	62.68	83.96
ADFL(Ours)	**98.17**	**77.51**	**87.50**

CK+. The proposed ADFL framework improves the accuracy of 3.98% over the baseline, which demonstrates the effectiveness of the adversarial disentangled feature learning method for expression recognition. The previous top accuracy achieved by DeRL [10] was 97.30%. Our ADFL improved the accuracy to 98.17%, which is so far the best performance reported in literature.

MMI. The results of the proposed ADFL framework and baseline suggest that the superiority of the adversarial disentangled feature learning framework. The accuracy of our method, 77.51%, is significantly higher than that of baseline (62.68%), and the best result in literature (75.12%).

Oulu-CASIA. The accuracy of our method (87.50%) outperforms that of the baseline (83.96%) with a 3.54% gap. Our method performs better than most of the approaches in literature, and is a little bit lower than that of the DeRL, 88.00%. However, the amount of augmented training images in the second stage of DeRL is about 10 times larger than that of our approach.

4. CONCLUSIONS

In this paper, we have investigated the facial expression recognition problem by developing a novel end-to-end adversarial disentangled feature learning framework. The facial expression image can be regarded as the combination of expressive component and neutral component. The expressive component feature and neutral component feature are disentangled by adversarial learning and are further used to calculate the residual expression feature for expression classification. The evaluation results on several benchmark databases, i.e. the CK+, MMI and Oulu-CASIA, demonstrate the effectiveness of the proposed ADFL framework.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 09:03:45 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 357–360.

[2] Guoying Zhao and Matti Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[3] Weicheng Xie, Linlin Shen, Meng Yang, and Zhihui Lai, "Active au based patch weighting for facial expression recognition," *Sensors*, vol. 17, no. 2, pp. 275, 2017.

[4] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," 2001. software available at http://www. csie. ntu. edu. tw/˜ cjlin/libsvm," 2012.

[5] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al., "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[6] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[7] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 558–565.

[8] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2983–2991.

[9] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[10] Huiyuan Yang, Umur Ciftci, and Lijun Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

[11] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[12] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2080–2089.

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, p. 1.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[15] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

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

[17] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, p. 5.

[18] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749–1756.

[22] Guohang Zeng, Jiancan Zhou, Xi Jia, Weicheng Xie, and Linlin Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 423–430.

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on August 13,2021 at 09:03:45 UTC from IEEE Xplore. Restrictions apply.