Adversarial Feature Distillation for Facial Expression Recognition

Mengchao Bai[1,2], Xi Jia[1,2], Weicheng Xie[1,2], and Linlin Shen[1,2(✉)]

[1] School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, People's Republic of China
{baimengchao2017,jiaxi}@email.szu.edu.cn, {wcxie,llshen}@szu.edu.cn
[2] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, People's Republic of China

Abstract. Human face image contains abundant information including expression, age and gender, etc. Therefore, extracting discriminative feature for certain attribute while expelling others is critical for single facial attribute analysis. In this paper, we propose an adversarial facial expression recognition system, named expression distilling and dispelling learning (ED^{2}L), to extract discriminative expression feature from a given face image. The proposed ED^{2}L framework composed of two branches, i.e. expression distilling branch ED^{2}L-t and expression dispelling branch ED^{2}L-p. The ED^{2}L-t branch aims to extract the expression-related feature, while the ED^{2}L-p branch extracts the non-related feature. The disentangled features jointly serve as a complete representation of the face. Extensive experiments on several benchmark databases, i.e. the CK+, MMI, BU-3DFE and Oulu-CASIA, demonstrate the effectiveness of the proposed ED^{2}L framework.

Keywords: Facial expression recognition · Feature distilling · Feature dispelling · Adversarial learning

1 Introduction

Facial expression is one of the most important characteristics for people to express emotion and interact with others. In the field of computer vision and machine learning, numerous studies have been conducted on the facial expression recognition (FER) due to its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems [1]. In [2], Ekman and Friesen firstly defined six basic emotions, including anger, disgust, fear, happiness, sadness and surprise. Contempt was subsequently added as one of the basic emotions [3].

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).

A. C. Nayak and A. Sharma (Eds.): PRICAI 2019, LNAI 11672, pp. 80–92, 2019.
https://doi.org/10.1007/978-3-030-29894-4_7

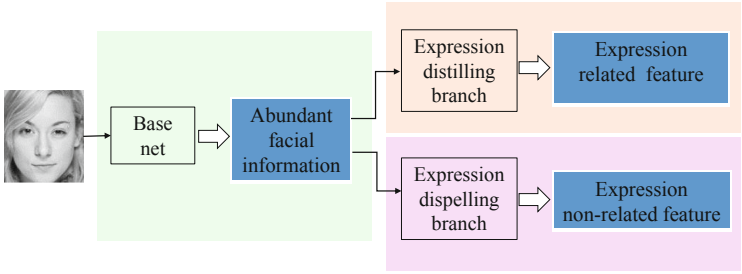


Fig. 1. Overview of our approach. A facial image contains abundant information. Our approach consists of two branches, which separate expression-related and non-related features for facial expression recognition by adversarial learning.

Current FER systems in the literature can be classified into two categories according to their feature extraction methods: hand-crafted features based methods and deep learning based methods. Majority of hand-crafted features based methods employed features such as LBP-TOP [4] and Gabor [5] to represent a given image. The extracted features are then used to classify facial expressions by Support Vector Machine (SVM) [6] or Nearest Neighbor classifier. Zhao and Pietikainen [4] proposed the LBP-TOP operator for expression recognition, which extracts co-occurrence features by computing concatenated LBP histograms from three orthogonal planes. Xie et al. [5] employed the Gabor surface feature (GSF) to represent the facial expression and SVM for classification. Since the extraction of hand-crafted features is separated from the training of classifier, these methods may lose useful facial information and achieve limited performance.

To extract sufficient and representative features, the deep learning based methods (e.g. IACNN [7] and DTAGN [8]) were adopted to facial expression analysis. Meng et al. [7] proposed an identity-aware CNN network to capture both expression-related and identity-related information, which achieved 95.37% accuracy on the CK+. Jung et al. [8] proposed the DTAGN composed of two different deep networks to extract temporal appearance feature from image sequences and temporal geometry feature from temporal facial landmark points, respectively. Although the performance of these methods are better than the hand-crafted features based methods, their capacities are still limited. Because a human face contains various attributes, e.g. age, skin color and gender, these expression features may be confused with other facial attributes related features.

With consideration to the aforementioned issues, some scholars tried to extract facial expression feature by comparing the differences between query face image and neutral face image. Yang et al. [9] proposed a De-expression Residue Learning (DeRL) method to extract expressive component (the difference between neutral expression and other expressions). The DeRL composed of two stages: First, a generator is trained using cGAN [10] to regenerate the neutral face image for a facial expression image. Then, the expression

information contained in the intimidate layers of the generative model was captured and concatenated for facial expression recognition. Since 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 in the second stage. Liu et al. [11] proposed a distilling and dispelling auto-encoder (D^2AE) framework to perform face editing. Its encoder contains two branches: identity-distilling and identity-dispelling branches, to extract the identity information and the complementary facial information, respectively. Features in the two streams represent different information of a face, which were then used by the decoder to manipulate facial attributes.

In this paper, inspired by the success of the DeRL [9] and D^2AE [11], we propose an end-to-end adversarial expression distilling and dispelling learning (ED^2L) framework for facial expression recognition, as depicted in Fig. 1. Similar to Liu et al. [11], the proposed ED^2L have two branches, i.e. the expression distilling and dispelling branches. Since the facial expression database is much smaller than those databases for face identification, the facial expression database is not large enough to train complex face identification network. We use SpherefaceNet-20 [12] instead of Inception-ResNet [13] as the backbone of our framework, which makes our network structure much lighter than D^2AE. The model parameter size of D^2AE is about 20 times larger than that of our approach, which saves computational resources and brings about a faster convergence during training our framework. In addition, Additive Margin Softmax [14] is used in our expression distilling branch as the loss function. Also, as shown with the purple dotted arrow in Fig. 2, the optimization of l_e^p in the expression dispelling branch updates Base net, $B_{\theta p}$ and dispeller simultaneously. The proposed ED^2L framework aims to separate discriminative expression feature from other face information. Our main contribution can be summarized as follows:

- A adversarial ED^2L framework is proposed to disentangle expression-related feature from a given face.
- The adversarial learning of the proposed ED^2L framework ensures the effective extraction of the expression-related and non-related features.
- The automatically learned expression-related feature achieves competitive performance in several benchmark databases.

2 Methods

In this section, we introduce the proposed ED^2L framework. As visualized in Fig. 2, the entire framework consists of three parts, the base net S_θ and two parallel branches: expression distilling branch ED^2L-t and expression dispelling branch ED^2L-p. Given a face image x, a variety of face attribute information $S_\theta(x)$ is extracted by the base net S_θ. Then, $S_\theta(x)$ is fed into expression distilling branch $B_{\theta t}$ and expression dispelling branch $B_{\theta p}$ to further extract expression-related and non-related features, respectively. The expression-related feature $\boldsymbol{f}_t \in R^{N_t}$ and non-related feature $\boldsymbol{f}_p \in R^{N_p}$ jointly serve as a complete representation of the face.

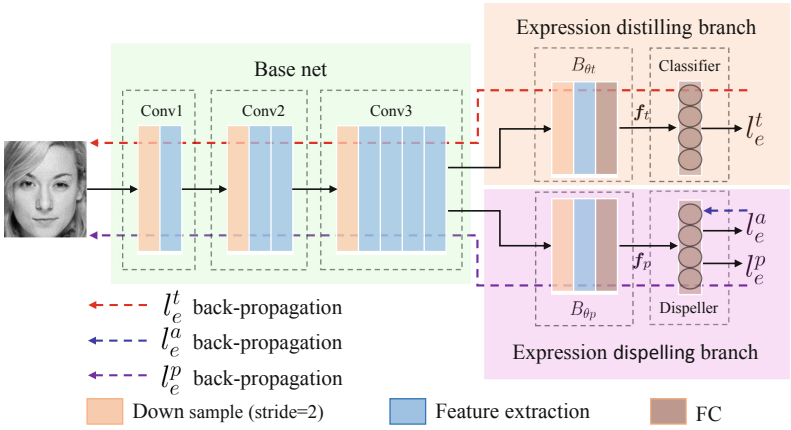


Fig. 2. The expression distilling and dispelling framework. (Color figure online)

2.1 Base Net

Adapted from SpherefaceNet-20 [12], the architecture of our framework is demonstrated in Table 1. Conv1, Conv2 and Conv3 denote convolutional blocks that contain multiple convolutional layers and residual units. $[3 \times 3, 64] \times 2$ 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 [15] layer. FC-256 denotes a fully connected layer with 256 neurons.

2.2 Expression Distilling Branch

We propose the expression distilling branch ED^{2}L-t to extract discriminative expression-related information $\boldsymbol{f}_t$. As revealed in Fig. 2, $\boldsymbol{f}_t$ is extracted using the subnet $B_{\theta t}$ after the base net.

$$\boldsymbol{f}_t = B_{\theta t}(S_\theta(x)) \tag{1}$$

Then, $\boldsymbol{f}_t$ is mapped by a non-linear function *Additive Margin Softmax* [14], defined in Eq. (2),

$$\boldsymbol{y}_t = \frac{e^{s(W_{y_t}^T \boldsymbol{f}_t - m)}}{e^{s(W_{y_t^i}^T \boldsymbol{f}_t - m)} + \sum_{j=1, j \neq y_t^i}^{c} e^{s W_j^T \boldsymbol{f}_t}} \tag{2}$$

where $\boldsymbol{y}_t \in R^{N_t}$ is an N_t-dimensional vector, which represents the probabilities of belonging to the corresponding class, m and s are two hyper-parameters of the additive margin softmax which denote the margin among categories and scaling

Table 1. Architectures of the proposed ED^{2}L framework.

Components	Layers	Configurations
Base net	Conv1	$[3 \times 3, 32] \times 1$, S-2 $[3 \times 3, 32; 3 \times 3, 32] \times 1$
	Conv2	$[3 \times 3, 64] \times 1$, S-2 $[3 \times 3, 64; 3 \times 3, 64] \times 2$
	Conv3	$[3 \times 3, 128] \times 1$, S-2 $[3 \times 3, 128; 3 \times 3, 128] \times 4$
Expression distilling branch	$B_{\theta t}$	$[3 \times 3, 256] \times 1$, S-2 $[3 \times 3, 256; 3 \times 3, 256] \times 1$ FC-256
	Classifier	#Expression Category
Expression dispelling branch	$B_{\theta p}$	$[3 \times 3, 256] \times 1$, S-2 $[3 \times 3, 256; 3 \times 3, 256] \times 1$ FC-256
	Dispeller	#Expression Category

factor, respectively. The classification loss l_e^t is computed by the probability vector $\boldsymbol{y}_t \in R^{N_t}$, where i denotes the ground truth index.

$$l_e^t = -\log \boldsymbol{y}_t^i \tag{3}$$

The back-propagation route of l_e^t optimization including the expression distilling branch $B_{\theta t}$ and base net S_θ is indicated with the red dotted arrow in Fig. 2.

2.3 Expression Dispelling Branch

Similar to the ED^{2}L-t, the structure of expression dispelling branch ED^{2}L-p composed of a subnet $B_{\theta p}$ and an expression dispeller. The ED^{2}L-p inhibits expression-related feature and extracts the non-related feature $\boldsymbol{f}_p$ by the subnet $B_{\theta p}$ following the base net.

$$\boldsymbol{f}_p = B_{\theta p}(S_\theta(x)) \tag{4}$$

In order to ensure that the ED^{2}L-p can extract expression non-related feature, an adversarial supervised training method composed of two different loss functions l_e^a and l_e^p is employed.

The cross entropy loss $l_e^a = -\log \boldsymbol{y}_p^i$ is leveraged to supervise the training of the expression dispeller based on $\boldsymbol{y}_p$, which is computed by $\boldsymbol{y}_p = softmax(W_p \boldsymbol{f}_p + b_p)$. Note that the gradient of l_e^a is only back-propagated to the expression dispeller and does not update the previous layers, which is different from l_e^t.

l_e^p is proposed to fool the training of expression dispeller $\boldsymbol{y}_p$. In other words, l_e^p 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, where N denotes the number of expression categories.

$$l_e^p = -\frac{1}{N} \sum_j^N \log \boldsymbol{y}_p^j \tag{5}$$

The optimization of l_e^p updates the expression dispelling branch B_{θ_p} and the base net S_θ.

The sum of l_e^a and l_e^p constitutes the total loss function of the expression dispelling branch. Note that $\boldsymbol{y}_p$ of the feature dispelling branch is not used to predict the expression category.

2.4 Objective Function

The ED^{2}L framework is jointly optimized by three loss functions l_e^t, l_e^a and l_e^p. The total loss function L is the weighted sum of l_e^t, l_e^a and l_e^p, as formulated in Eq. (6).

$$L = \lambda_t l_e^t + \lambda_p (l_e^a + l_e^p) \tag{6}$$

3 Experiments and Results

In this section, we evaluate the performance of the proposed approach on four benchmark databases, including CK+ [16], MMI [17], BU-3DFE [18] and Oulu-CASIA [19], and compare the results with the state-of-the-art methods.

3.1 Implementation Details

Data Preprocessing. For each database, the faces are first detected by the MTCNN [20] and aligned to the resolution of 128×110 according to their corresponding landmarks. Then, ten gray patches with the size of 112×96 are generated by cropping from four corners and center of each aligned image and the horizontal flipping mirror.

Hyperparameters. The proposed ED^{2}L framework is optimized using Adam optimizer [21] 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 $m = 0.35$, $s = 30$, $\lambda_t = 1$ and $\lambda_p = 10$.

3.2 Databases

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. Among these videos, only 327 sequences from 118 subjects are labeled with seven expressions (anger, contempt, disgust, fear, happiness, sadness and surprise). In order to compare with other methods, the 10-fold cross validation protocol in [9] is followed. The last three frames of each labeled sequence are selected and all subjects are divided into ten groups based their ID in an ascending order. Every subgroup is further selected as testing set to evaluate the model performance, and the remaining subgroups are used for training in the 10-fold cross validation.

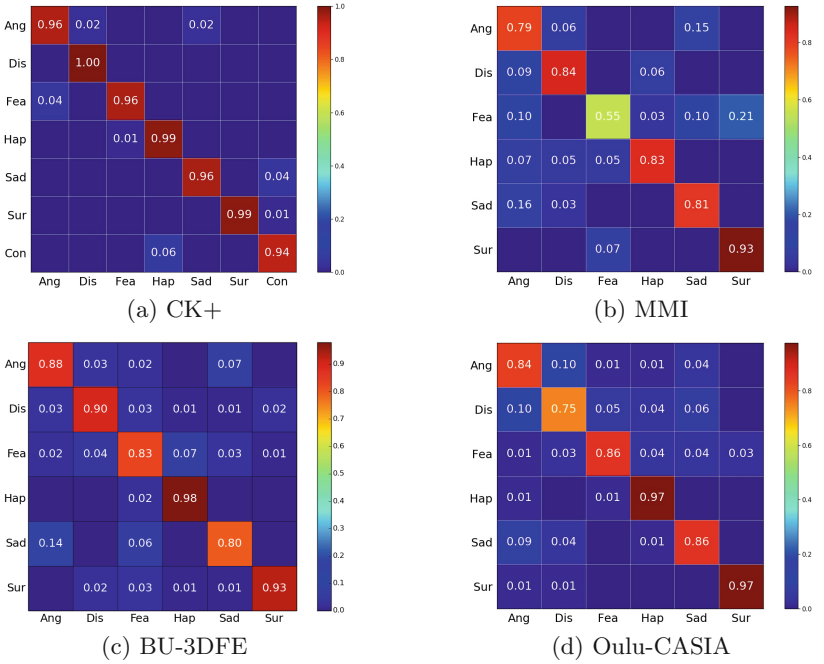


Fig. 3. Confusion matrix of the ED^{2}L framework with fine-tuning for the CK+, MMI, BU-3DFE and Oulu-CASIA databases. The labels on the vertical and horizontal axis represent ground truth and predicted expressions, respectively.

The MMI database [17] consists of 236 sequences from 32 subjects with six basic expressions. We select 209 sequences captured in front view. Since the sequences of this database begin with the neutral expression and show a peak expression near the middle of the sequences. We select three frames in the middle of each sequence and employ a 10-fold cross validation similar to that of the CK+ database.

The BU-3DFE database [18] consists of 2500 pairs of 3D face models and texture images of 100 subjects (56 female and 44 male subjects). Each subject displayed six basic facial expressions (anger, disgust, fear, happiness, sadness and surprise) with four intensity levels and a neutral expression. Following the test protocol in [9], only the texture images with high-intensity expressions (i.e. the last two levels) were selected. The selected pictures were further divided into 10 subject-independent groups.

The Oulu-CASIA database [19] contains two subsets, i.e. the Oulu-CASIA NIR database and the Oulu-CASIA VIS database, which were captured under three different illumination conditions (dark, weak and strong) using a NIR camera and a VIS camera, respectively. In our experiments, only the Oulu-CASIA VIS database under strong illumination condition is used. The Oulu-

Table 2. Overall accuracy on the CK+ database. Remark that *w.* and *w.o.* denote the use of fine-tuning, or not, respectively.

Method	Accuracy (%)
LBP-TOP [4]	88.99
3DCNN [23]	85.90
STM-Explet [24]	94.19
IACNN [7]	95.37
DTAGN-Joint [8]	97.25
DeRL [9]	97.30
Baseline(*w.o.*)	94.19
Baseline(*w.*)	94.50
Ours(*w.o.*)	96.33
Ours(*w.*)	**97.86**

CASIA VIS database 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 of each sequence are selected and a 10-fold cross validation is applied.

3.3 Experiments

Baseline. In order to prove the effectiveness of the proposed ED^{2}L framework, we employed a baseline network for comparison which has the same structure as the ED^{2}L framework without ED^{2}L-p branch.

Transfer Learning. Training of the CNN is prone to over-fitting because the number of images in the CK+, MMI, BU-3DFE and Oulu-CASIA databases are insufficient. Therefore, firstly, we trained the ED^{2}L framework on the FER2013 [22] database with the same parameter settings described in Sect. 3.1 and used the pretrained model as the base model. Then, the base model was further fine tuned using the CK+, MMI, BU-3DFE and Oulu-CASIA databases. When training the baseline model, the same procedure was adopted.

3.4 Results

CK+. The overall accuracy of 10-fold cross validation is displayed in Table 2. The proposed ED^{2}L framework outperforms the baseline with a 3.36% gap, which suggest the effectiveness of the adversarial learning between two branches. Compared to other methods, our approach achieves the best performance, i.e. 97.86% and beats all hand-crafted features based methods (LBP-TOP [4]) and CNN-based methods (3DCNN [23], STM-Explet [24], IACNN [7], DTAGN-Joint [8] and DeRL [9]). Figure 3(a) shows the confusion matrix of ED^{2}L framework for the CK+ database. Diagonal of this matrix, suggests that our method performed remarkably well in recognizing the expressions of disgust, happiness and surprise.

Table 3. Overall accuracy on the MMI database.

Method	Accuracy (%)
LBP-TOP [4]	59.51
STM-Explet [24]	75.12
DTAGN-Joint [8]	70.24
IACNN [7]	71.55
DeRL [9]	73.23
Baseline($w.o.$)	62.68
Baseline($w.$)	76.56
Ours($w.o.$)	72.73
Ours($w.$)	**80.38**

Table 4. Overall accuracy on the BU-3DFE database.

Method	Accuracy (%)
Wang et al. [25]	61.79
Berretti et al. [26]	77.54
Yang et al. [27]	84.80
Li et al. [28]	86.32
Lopes [29]	72.89
DeRL [9]	84.17
Baseline($w.o.$)	86.00
Baseline($w.$)	87.17
Ours($w.o.$)	87.83
Ours($w.$)	**88.67**

MMI. Table 3 lists the results of the proposed ED^{2}L framework, together with that of baseline and other approaches in literature. The accuracy of our approach with fine tuning, 80.38%, is significantly higher than that of baseline (76.56%), and the best results in literature (75.12%). As shown from the confusion matrix of MMI database in Fig. 3(b), the ED^{2}L framework has a remarkable recognition performance for the expression of surprise.

BU-3DFE. As it can be seen in Table 4, the accuracy of our approach, 87.83% show a better performance than that of the baseline (86.00%) and the best result in literature (86.32%). As illustrated in Fig. 3(c), our approach performed well in recognizing the expression of happiness.

Table 5. Overall accuracy on the Oulu-CASIA database.

Method	Accuracy (%)
LBP-TOP [4]	68.13
STM-Explet [24]	74.59
Atlases [30]	75.52
DTAGN-Joint [8]	81.46
PPDN [31]	84.59
DeRL [9]	**88.0**
Baseline($w.o.$)	83.96
Baseline($w.$)	84.58
Ours($w.o.$)	85.21
Ours($w.$)	**87.71**

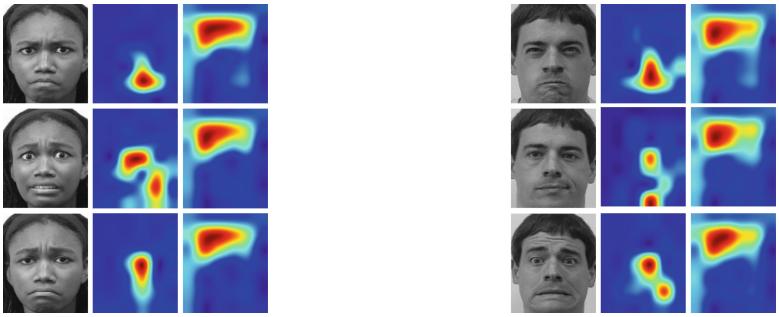
(a) ED^{2}L (b) baseline

Fig. 4. Visualization of the features extracted by the adversarial ED^{2}L framework and baseline network, using t-SNE [32]. (Color figure online)

Oulu-CASIA. The overall accuracy of 10-fold cross validation is illustrated in Table 5. Fine-tuning has also been shown to improve the accuracy of our framework from 85.21% to 87.71%, which is again higher than that of baseline, 84.58%. When the performance of our framework is better than most of the approaches in literature, our accuracy is a little bit lower than that of DeRL, 88.0%. However, the number of training images (60,600) used for pretrained model in DeRL is much bigger than that of our approach (28,709). The amount of augmented training images in the second stage of DeRL is also about 10 times larger than that of our approach.

3.5 Visualization

In order to further illustrate the effectiveness of the proposed ED^{2}L framework, we extract the image features of the CK+ database from the FC-256 layer of the ED^{2}L-t branch and baseline, respectively. We use the first validation set of the 10-fold cross validation protocol to extract these features. Note that as subject independent division is used, the subjects in the first fold only present six expressions,



(a) Female subject (b) Male subject

Fig. 5. Visualization of the feature heat-maps extracted from the down sample layers of ED^{2}L-t and ED^{2}L-p. The left column is the input image, the middle and right columns are the feature heat-maps extracted from ED^{2}L-t and ED^{2}L-p, respectively.

i.e. anger, disgust, fear, happiness, sadness and surprise. As depicted in Fig. 4, the features extracted by the ED^{2}L framework are densely clustered for each expression category and easy to distinguish. There are distinct boundaries between features of different expressions. While the features extracted by the baseline network are non-discriminative and have ambiguous boundaries, e.g. the points of fear expression (blue points) are mixed with others. The results qualitatively suggests that the proposed approach has an extraordinary ability to extract discriminative expression-related information, mainly due to the adversarial supervised learning of the expression distilling and dispelling branches.

In addition, we extract the feature maps of the CK+ database from the down sample layer of ED^{2}L-t and ED^{2}L-p, respectively. The 10th validation set of the 10-fold cross validation protocol is used to extract these feature maps composed of 256 channels. Then the sum of these feature maps is normalized to $[0, 1]$ to calculate the heat-maps. The feature heat-maps are resized to 112×96 to match the size of input image. In Fig. 5, we extract and visualize the feature maps for different expressions of two different subjects. For different expressions of the female subject shown in Fig. 5(a), the heat-maps extracted from ED^{2}L-t differ significantly with each other, while the ED^{2}L-p heat-maps are almost the same. The same conclusion can be drawn for the male subject shown in Fig. 5(b). The examples clearly suggest that ED^{2}L-t tries to look at regions sensitive to expressions like eyes, nose and mouth, while ED^{2}L-p focus on expression invariant regions like forehead.

4 Conclusions

In this paper, we present an adversarial expression distilling and dispelling learning (ED^{2}L) framework for facial expression recognition. The framework uses expression distilling (ED^{2}L-t) and dispelling (ED^{2}L-p) branches to extract expression-related and non-related features, respectively. The features learned by

two branches jointly serve as a complete representation of the face. As evaluated on several facial expression benchmark databases, the ED^{2}L framework showed its superiority over both traditional hand-crafted features based methods and CNN-based methods.

References

1. Li, S., Deng, W.: Deep facial expression recognition: a survey. arXiv preprint arXiv:1804.08348 (2018)
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
3. Matsumoto, D.: More evidence for the universality of a contempt expression. Motiv. Emot. **16**(4), 363–368 (1992)
4. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
5. Xie, W., Shen, L., Yang, M., Lai, Z.: Active AU based patch weighting for facial expression recognition. Sensors **17**(2), 275 (2017)
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). Software: http://www.csie.ntu.edu.tw/~cjlin/libsvm (2012)
7. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 558–565 (2017)
8. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
9. 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, pp. 2168–2177 (2018)
10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
11. Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X.: Exploring disentangled feature representation beyond face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2080–2089 (2018)
12. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 1 (2017)
13. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
14. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
16. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)

17. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, p. 5 (2005)
18. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR 2006), pp. 211–216. IEEE (2006)
19. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
20. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013. LNCS, vol. 8228, pp. 117–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42051-1_16
23. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 143–157. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16817-3_10
24. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
25. Wang, J., Yin, L., Wei, X., Sun, Y.: 3D facial expression recognition based on primitive surface feature distribution. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 1399–1406. IEEE (2006)
26. Berretti, S., Del Bimbo, A., Pala, P., Amor, B.B., Daoudi, M.: A set of selected SIFT features for 3D facial expression recognition. In: 2010 20th International Conference on Pattern Recognition, pp. 4125–4128. IEEE (2010)
27. Yang, X., Huang, D., Wang, Y., Chen, L.: Automatic 3D facial expression recognition using geometric scattering representation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–6. IEEE (2015)
28. Li, H., et al.: An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition. Comput. Vis. Image Underst. **140**, 83–92 (2015)
29. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
30. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition using longitudinal facial expression atlases. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 631–644. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_45
31. Zhao, X., et al.: Peak-piloted deep network for facial expression recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 425–442. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_27
32. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)