Pose-Invariant Facial Expression Recognition Based on 3D Face Morphable Model and Domain Adversarial Learning

Xiao Ma[1], Kaige Zhang[2(✉)], and Xuan Yang[3]

[1] School of Computer Science and Engineering,
South China University of Technology, Guangzhou 510641, China
201630610632@mail.scut.edu.cn
[2] Department of Computer Science, Utah State University, Logan 84322, USA
kg.zhang@aggiemail.usu.edu
[3] College of Software, Taiyuan University of Technology, Taiyuan 030600, China

Abstract. Pose is one of the most important factors affecting performance of face related recognition algorithms including facial expression recognition (FER). Traditionally, non-frontal FER is conducted by either performing face formalization or designing separate models for different poses. Different from those methods, we propose a one-stage FER approach by training a pose invariant deep convolutional network (DCNN) with the following novelties: First, we introduce the 3D face morphable model to reconstruct high fidelity 3D faces for data augmentation which increases the pose variety without losing expression information. Second, we employ domain adversarial learning to eliminate the influence of domain difference between real 2D face images and 3D synthetic face images at feature level, which realizes a one-stage deep FER approach that is robust to different face poses. Third, the proposed approach provides a solution for cross-domain problems involving data from different sources, which can be applied to other face related recognition problems. The method is validated using three FER datasets FER2013, multi-PIE and BU-3DFE; and it outperforms the current state-of-the-art methods.

Keywords: Facial expression recognition · Deep learning · 3D morphable model · Domain adaptation

1 Introduction

Computer vision based facial expression recognition (FER) has been a research topic for many years. In the early time, Ekman et al. [1] defined six basic facial expressions shared among human beings to express their emotions (i.e., anger, disgust, fear, happiness, sadness and surprise). With the fast development of artificial intelligence, FER has gained increasing attentions because of its valuable applications such as driver fatigue monitoring, human-computer interaction, medical care, digital entertainment, etc. In general, FER is conducted by

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 491–502, 2019.
https://doi.org/10.1007/978-3-030-34113-8_41

feature extraction and machine classification. Traditionally, feature extraction is conducted by using hand-crafted feature extractors such as local binary patterns (LBPs) [2], histogram of gradients (HOG) [3], and histogram of optical flow (HOF) [11]; and for machine classification, it can be support vector machine, Bayesian classifiers, random forest, etc. These methods achieved some success on FER; however, in-the-wild facial expression recognition task is still a challenge due to the high appearance changes.

Over the last several years, deep learning has achieved great success and dominated the state-of-the-arts in many challenging problems [4], in which an artificial neural network (ANN) with multiple hidden layers is used; and it has been viewed as the most promising means for solving in-the-wild FER problem. However, deep learning based approach is a data-driven strategy which relies on large amount of relevant training data; in practice, the available datasets mainly consist of frontal face images, that makes the FER with large pose face images remain a challenge.

There has been methods addressing non-frontal face expression recognition problems, which involves 3D and 2D methods [5–7]. Traditionally, the 3D method was suffering from the computation efficiency and hard-to-converge problems [5]; and the 2D based methods are usually designed as view-specific and they required a similar amount of data in different poses for the training, which limits the performance of the methods because that there exists a disproportion in the availability of frontal and non-frontal view facial expression data [6]. Such problem is more distinct in deep learning based method because of the inherent data-hungry and data-dependence. During the past several years, the 3D face morphable model has achieved impressive progress and the model has been used to reconstruct high fidelity tridimensional faces which can not only construct the 3D face shape from a 2D face image but can also reserve the other attributes such as face expressions [7,8]. With the fact that 3D model is able to produce 2D face images from arbitrary views, it can be used to improve the face related recognition algorithms.

In this paper, we propose a one-stage pose-invariant FER approach by training a DCNN. First, the 68-landmarks are located accurately from a 2D face image. Then, the state-of-the-art 3D morphable model (3DDFA) [8] is used to reconstruct the 3D face model from the 2D image without losing the original facial expression information; and the reconstructed 3D face is used to produce high fidelity 2D face images with the same emotion as the original face at multiple poses. At last, the augmented facial image data, including original and 3D generated face images, are utilized to perform a joint, alternative training for FER where feature-level domain adversarial learning is utilized for network training to eliminate the influence of domain difference between 3D-generated face images and the real 2D face images. The experimental results demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. In Sect. 2, it introduces the works related to the proposed method. In Sect. 3, it discusses the proposed method. In Sect. 4, it presents the experimental settings and the results. In Sect. 5, it concludes the paper.

2 Related Works

Systematic FER has been discussed in some survey papers [9–11]. Here, we first review some FER works related to this work. Then, we give a brief introduction to other related works including 3D face morphable model and generative adversarial learning.

2.1 Facial Expression Recognition

Deep learning based methods have been used for FER with the availability of large amount training data. Mollahossein et al. [12,13] introduced deep neural networks for automatic feature extraction which outperformed the hand-crafted feature extraction methods. Jung et al. [14] fine-tuned a deep neural network for FER. Hasani et al. [15] proposed a spatio-temporal FER method using DCNN and conditional random field. Based on local facial action units, Liu et al. [16,17] extract facial features on several key points for FER. These works achieved very good performances for frontal face expression recognition; however, they are suffering from performance degradation with large pose face images. Pose-invariant FER has been studies for many years of which the traditional pattern classification methods were employed [18–20]. Recently, deep learning based methods have also been utilized to address this issue. Zhang et al. [21] combined traditional feature extraction method (SIFT) with deep neural network for multi-view FER. Liu et al. [22] proposed a multi-channel pose-aware CNN for multi-view FER. Lai et al. [23] employed generative adversarial networks (GANs) to perform face frontalization before expression recognition of non-frontal faces. A common issue in deep learning based methods is the lacking of sufficient labeled face data which results a network either biased to the frontal face recognition or with poor performance.

2.2 3D Face Morphable Model

Blanz et al. [24] proposed the 3D face morphable model which can be directly matched to a 2D image, where the head pose, expression, illumination and other parameters are free variables subject to be optimized. Based on accurate landmark detection, it can be used to reconstruct photo-realistic face from a 2D face image [7]. Recently, by introducing deep neural networks, the state-of-the-art of 3D face alignment has been pushed to a new level. For example, Zhu et al. [8] utilized deep neural network to regress the 3DMM parameters from 2D face images to reconstruct high fidelity 3D faces where the landmark information is used. Chang et al. [25] proposed ExpNet which regress the 3D expression parameters directly from the face images without using landmark detection. In this work, we used the 68-landmark detection as a proxy step and reconstruct 3D face from a 2D face image with expression label to generate face images at multiple poses for data augmentation.

2.3 Domain Adversarial Learning

Goodfellow et al. [26] proposed the generative adversarial network (GAN) which can be trained to generate real-like images by minimizing the distribution difference of the data generated from random noise or from real-world by playing a max-min game. While GAN provided a strategy to generate real-like data at the image level, Ganin et al. [27] proposed deep domain adaptation which can transfer two or multiple domains to a common domain at feature level. The com-mon thing is that they both used the adversarial loss to make the generated images/features indistinguishable by the discriminator (i.e., eliminate the domain difference between two datasets at image/feature level). In this work, we introduced the domain adversarial learning for a joint training to eliminate the domain difference between 3D-generated face images and the real face image which reserves the facial expression information in the meanwhile.

3 Proposed Method

3.1 Overview

This approach employs the 3DMM for data augmentation and proposes a novel learning strategy to overcome the domain problem that has troubled the network training. Figure 1 is an overview of the proposed method. Given a 2D face image, it first performs 68-landmark detection with a well-trained deep convolutional neural network. Then the landmarks and the original 2D face image is sent to 3DDFA to fit the 3D morphable model (3DMM) and reconstruct a 3D face with high fidelity where the expression parameters are also well regressed. The 3D model is then used to produce 2D face images with different poses for non-frontal-face data augmentation. Ideally with sufficient data, the deep classification network can be trained to recognition different facial expressions; however, it is found that the network can well classify the 3D generated face images but suffered performance degradation when dealing with the real 2D data. In order to eliminate the influence of domain difference between the generated face images and the original 2D face image, it introduces feature-level domain adaptation to perform a joint training which updates the network parameters with original 2D face image and generated 2D face image alternatively, following the training of a generative adversarial network (GAN).

3.2 2D Facial Landmark Detection

Landmark detection has been a proxy step for 3DMM fitting of which a 2D image and the corresponding key points are necessary to construct a high fidelity 3D face. Comparing to regressing a complicated 3D face morphable model (some recent work has tried to regress 3D face model directly from 2D image [25]), the landmarks detection using a DCNN is relatively easy and the effectiveness has been studied and verified by many references [28]. In this work, we start from performing the 68-facial landmarks detection from 2D face images (i.e.,

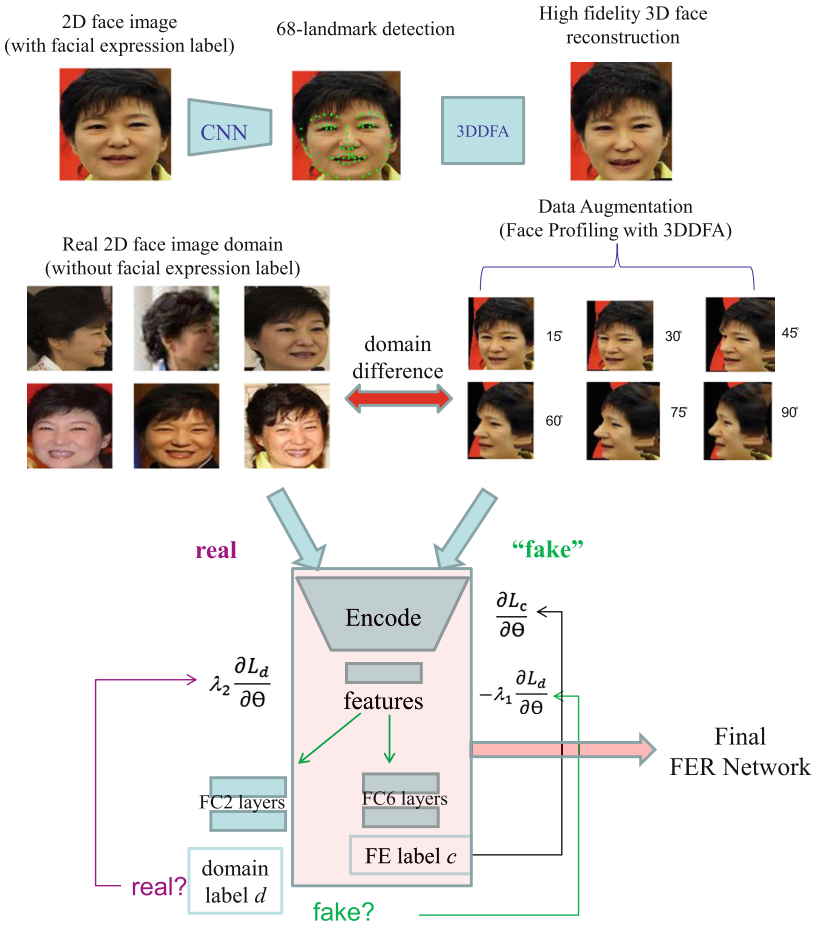

Fig. 1. Overview of the proposed method.

performing a 2D face alignment) with a deep convolutional network. Then the landmark information and the original 2D image are used to reconstruct the corresponding 3D face with 3D face morphable model.

For landmark alignment, the most widely used dataset 300-W-LP [29] is introduced; it is obtained by expanding the 300-W dataset where the landmark annotation is available. Following [30], we directly train a deep convolutional network to regress the landmark coordinates with the normalized mean error loss:

$$NME = \frac{1}{N}\sum_{k=1}^{N}\frac{\|x_k - y_k\|_2}{d} \tag{1}$$

where x is the ground truth (GT) landmarks of a 2D face image, y is the estimated landmark coordinates obtained by the deep network, and d is the square-

root of the GT bounding box. VGG-16 is used as the base network for the training and the network is fine-tuned based on the state-of-the-art face recognition network [31].

3.3 Multi-pose Data Augmentation with 3D Model

3D Face Reconstruction. Blanz et al. [24] developed the 3D morphable model for human-face description with the following formulation:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \tag{2}$$

where S is a 3D face model, $\bar{S}$ is a mean face model, A_{id} is the principle axes trained on the 3D face scans with neutral expression, α_{id} is the shape parameter, A_{exp} is the principle axes trained on the offsets between expression scans and neutral scans and α_{exp} is the expression parameter. It has been studied in many works [7,8] that with the accurate 68-landmarks and the original face image, the 2D image can be mapped to the 3D model to obtain a 3D face. In this paper, it employs the state-of-the-art 3D face alignment method [8] where a deep convolutional network is trained for parameter optimization via minimizing the difference of the generated 2D face and the real 2D face. By this way, a high fidelity 3D face can be constructed from the related real 2D face image with the same facial emotion. Refer [8] for more details about the face reconstruction.

Data Augmentation. After the 3D face is constructed, it can be projected onto specific 2D plane with the scale orthographic projection:

$$V(p) = f * Pr * R * (\bar{s} + A_{id}\alpha_{id} + A_{exp}\alpha_{id}) + t_{2d} \tag{3}$$

where $V(p)$ is the model construction and projection function, leading to the 2D positions of model vertices, f is the scale factor, Pr is the orthographic projection matrix, R is the rotation matrix and t_{2d} is the translation vector.

Based on the 3D-2D geometric mapping described above, it can generate 2D face images of arbitrary poses with the same expression label as the original image which can be used to augment the image data of non-frontal face. In our experiments, yaw angles of 15°, 30°, 45°, 60°, 75°, 90° are produced for each face; and they are used to train a deep neural network for FER.

3.4 Joint Training with Domain Adaptation

Ideally, if the related data is sufficient (i.e., non-frontal face images with different poses), the network could be trained to recognize facial expressions under different poses with a classic 6-class Sofmax loss. However, in our initial attempts, it was found that the network achieved very bad performance on the testing set for both frontal and non-frontal face images, even worse than the network trained without using data augmentation [refer Sect. 4.3]. After further exploration, it was found that the network did well on the 3D generated face images. That

indicates: (1) the 2D images generated from the 3D model is facial expression distinguishable; (2) the 3D model was well aligned which reserved the expression information; (3) the generated 2D images are not totally the same as the images captured directly using a camera.

In conclusion, there is a domain difference between the generated 2D images from the 3D model and the real 2D images which has failed the network training. That is, the "well-trained" network is a biased one which tends to deal with the generated images well; however, the true target, FER from real 2D face images, was not achieved. While the network can be used for the FER task by first fitting a 2D image to a 3D model and then generates a face image for FER, it is time consuming and tedious.

To develop a straight-forward FER deep network from 2D face images, we introduce the domain adaptation where the domain adversarial loss is utilized:

$$L_{adv} = -E_{x \in I}[logD(x)] \tag{4}$$

where x is the input face image, I is the 3D generated image dataset. It is combined with the regular softmax loss to perform a joint training which eliminates the domain difference between the generated face images and the real 2D face images. Different from other joint learning approaches which combined the losses as whole for the training, we treated them separately and updated the parameters alternatively using the images from the two domains. Thus, when the input image is a real 2D face image, it updates the network with the gradient computed from the domain loss $\frac{\partial L_d^r}{\partial \theta}$; however, when the input image is a 3D generated image, it updates the network using both the domain loss and the expression class loss, of which the negative gradient of the domain loss is used which aims at eliminating the domain difference between generated face images and real face images. Finally, the full objective is

$$L_{final} = 1\{x \in I\} * (L_{adv} + \lambda_1 L_d) + 1\{x \in R\} * \lambda_2 L_d \tag{5}$$

where I is the 3D generated dataset, R is the real dataset, and λ_1 and λ_2 are experimental results 0.30 and 0.65, respectively.

3.5 Implementation

As illustrated in Fig. 1, the network architecture is detailed as follows:
C_96_7_2 - ReLU - C_128_3_2 - ReLU - C_128_3_1 - ReLU - C_256_3_2 - ReLU - C_256_3_1 - ReLU - C_256_3_2 - ReLU - C_512_3_1 - ReLU - C_512_3_2 - ReLU - C_512_3_1 - ReLU - C_512_3_2 - ReLU - FC_512 - ReLU - FC_6/FC_2- SF_6/SF_2
The naming rule follows the format: "layer type_channel number_kernel size_stride"; "C" denotes convolution; "FC" is fully connected layer; and SF is the softmax layer. For instance, "C_64_7_2" means that the first layer is a convolutional layer and the number of channels is 64, the kernel size is 7 and the stride is 2. Note that the FC_6 is designed for the classification of 6 expressions; FC_2 is designed for the adversarial learning where the 2 channels represent for the two domains.

For training, different from other joint learning approaches which combined the losses as whole for the training, we adopt the training strategy of GAN and treated the two losses separately and update the network parameters by alternatively backpropogating the domain loss and the 6-class classification loss where the adversarial loss with negative gradients is back-propagated when the input sample is from 3DMM model.

4 Experiments

4.1 Dataset and Settings

To evaluate the proposed method, we compare with different methods including the traditional handcrafted-feature extraction and deep learning based methods. The confusion matrix and overall accuracy on the multi-view FER datasets Multi-PIE [32] and BU-3DFE [33] are computed quantitatively to verify the effectiveness of the propose method.

Multi-PIE database [32] contains the face images of 337 subjects, 235 male instances and 107 female instances; there are more than 750000 images captured with fifteen cameras on different illuminations and viewpoints. Each subject was asked to give six different kinds of expressions: neutral (NE), smile (SM), squint (SQ), surprise (SU), disgust (DI), and scream (SC). Finally, the data of 100 subjects were selected which includes complete six emotions with good quality, and 13 poses are used in the experiments which includes 0°, ±15°, ±30°, ±45°, ±60°, ±75°, 90° face views. The training testing split is 80:20.

BU-3DFE database [33] is a 3D facial emotion dataset which contains face images of 100 subjects, 56 female instances and 44 male instances. Slightly different form the Multi-PIE data, there are seven different facial expressions including anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), surprise (SU), and neutral (NE) in four levels. In the experiments, the models are used to generate 2D face images under different views including 0°, ±30°, ±45°, ±60°, 90°. The 100 subjects are randomly divided into 80 training subjects and 20 testing subjects.

4.2 Experimental Results

To provide an objective evaluation of the proposed method, we compare the proposed method with nine existing methods including kNN, LDA, LPP, D-GPLVM, GPLRF, GMLDA, GMLPP, MvDA, and DS-GPLVM as reported in [34]. As shown in Table 1, the proposed method achieves new state-of-the-art in average accuracy; and it also outperforms the other methods at different poses including the recognition from frontal faces. Moreover, different from most previous methods, the proposed method achieves close results (around 93%) on different poses which demonstrates the pose robustness of the propose method.

Table 1. Overall accuracy of different methods on Multi-PIE

Method	Pose					Avg.
	−30	−15	0	15	30	
KNN	80.88	81.74	68.36	75.03	74.78	76.15
LDA	92.52	94.37	77.21	87.07	87.47	87.72
LPP	92.42	94.56	77.33	87.06	87.68	87.81
D-GPLVM	91.65	93.51	78.70	85.96	86.04	87.17
GMLDA	90.47	94.18	76.60	86.64	85.72	86.72
GMLPP	91.86	94.13	78.16	87.22	87.36	87.74
MvDA	92.49	94.22	77.51	87.10	87.89	87.84
Zhang [35]	90.97	94.72	89.11	93.09	91.30	91.80
Proposed	93.10	94.96	92.80	94.60	92.20	93.53

From Table 2, the proposed method achieves very good recognition performance for different expressions. But the recognition accuracies of DI and SQ are lower than the other expressions; it is similar as the results reported in some references [36] which indicates that the recognition of the emotions is inherently difficult. For the results on BU-3DFE, since it is 3D data, the proposed method achieves very good result because the method itself is based on 3D model.

Table 2. Confusion matrix on Multi-PIE and BU-3DFE

Dataset	Multi-PIE							BU-3DFE					
DI	91.92	0	1.90	5.60	0	2.80	AN	95.30	2.70	0	1.30	0	1.07
SC	0.38	98.00	0	0	2.48	0	DI	0.27	85.48	0.20	4.36	2.10	5.93
SM	3.40	0	94.25	2.00	1.05	0	FE	0.50	0	92.82	5.44	0.17	0.20
SQ	4.20	0	1.80	91.9	0	5.70	HA	1.53	2.20	5.25	82.50	5.53	3.40
SU	0	2.00	0.45	0	95.42	1.20	SA	2.40	1.08	1.19	3.22	89.00	3.90
SE	0.10	0	1.60	0.50	1.05	90.3	SU	0	8.54	0.54	3.18	3.20	85.50
	DI	SC	SM	SQ	SU	NE		SU	SA	HA	FE	DI	AN

4.3 Ablation Study

The advantage of the proposed method is effective data augmentation which introduced 3D generated data to train the network. Here, we perform the ablation studies with the following settings: (1) directly training a CNN; (2) directly introduce extra data to train the CNN; (3) introduce the 3D model to augment the data at different poses and train the CNN; (4) train the CNN with 3D augmentation and domain adversarial learning. They are used to demonstrate the effectiveness of 3D model and the necessary of the proposed domain adversarial learning.

Table 3. Ablation study

Method	Pose					Avg.
	−30	−15	0	15	30	
CNN	90.67	94.53	89.01	93.01	91.10	91.66
CNN + FER2013	90.01	94.21	93.21	92.57	90.47	92.09
CNN + FER + 3D	70.10	71.10	72.32	71.20	70.13	70.97
CNN + 3D + DA (Proposed)	93.10	94.96	92.80	94.60	92.20	93.53

In Table 3, it shows that directly training a CNN achieved similar results as the deep learning method in [35], in which the recognition of large yaw angle faces are better than those frontal faces. By introducing the extra data FER2013, the CNN are trained to achieve better results on frontal faces (pose with 0°); however, there is performance degradation on the non-frontal faces. The results indicate that the deep network is data-dependence and the network was trained biased to the frontal faces since the FER2013 contains more frontal faces. For the approach that directly training the CNN with 3D augmented data, as discussed in the method part, the method achieved very bad results on all different poses due to the domain difference. Finally, the proposed method with domain adversarial learning overcomes the problem and achieves new state-of-the-art performance.

5 Conclusion

In this paper, we have proposed a novel deep learning strategy for pose-invariant facial expression recognition. It introduced 3DMM for data augmentation which has first shown that there exists a domain difference between the real 2D images and the generated images with 3DMM and such difference can fail the training. More important, it proposed a solution for this problem by introducing the feature-level domain adversarial learning to train the network which eliminated the influence of the domain difference without losing the expression information. The experimental results demonstrated the effectiveness of the proposed method and it achieved new state-of-the-art on two multi-view FER datasets. In the future, we will apply the proposed method to other face related recognition tasks.

References

1. Ekman, P., Friesen, W.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **172**, 124–129 (1971)
2. Zhong, L., Liu, Q., Yang, P., Metaxas, D.: Learning active facial patches for expression analysis. In: CVPR, pp. 2562–2569. IEEE (2012)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE CVPR, San Diego (2005)

4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
5. Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L.: Static and dynamic 3D facial expression recognition: a comprehensive survey. Image Vis. Comput. **30**(10), 683–697 (2012)
6. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4295–4304 (2015)
7. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.: High-fidelity pose and expression normalization for face recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 787–796 (2015)
8. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.: Face alignment across large poses: a 3D solution. In: Proceedings of Conference Computation Vision Pattern Recognition, Las Vegas (2016)
9. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
10. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
11. Li, S., Deng, W.: Deep facial expression recognition: a survey. arXiv preprint arXiv:1804.08348v2 (2018)
12. Mollahosseini, A., Chan, D., Mahoor, M.: Going deeper in facial expression recognition using deep neural networks. In: Applications of Computer Vision (WACV), pp. 1–10 (2016)
13. Mollahosseini, A., Hasani, B., Mahoor, M.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(99), 1010–1022 (2017)
14. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE ICCV, pp. 2983–2991 (2015)
15. Hasani, B., Mahoor, M.: Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In: Automatic Face & Gesture Recognition, pp. 790–795 (2017)
16. Liu, M., Li, S., Shan, S., Chen, X.: AU-aware deep networks for facial expression recognition. In: IEEE Automatic Face and Gesture Recognition (FG), pp. 1–6 (2013)
17. 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
18. Pantic, M., Rothkrantz, M.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1445 (2000)
19. Rudovic, O., Pantic, M., Patras, I.: Coupled Gaussian processes for pose-invariant facial expression recognition. TPA- MI **35**(6), 1357–1369 (2013)
20. Zheng, W., Tang, H., Lin, Z., Huang, T. S.: A novel approach to expression recognition from non-frontal face images. In: IEEE ICCV, pp. 1901–1908 (2009)
21. Zhang, Z., Luo, P., Chen, C.L., Tang, X.: From facial expression recognition to interpersonal relation prediction. Int. J. Comput. Vision **126**(5), 1–20 (2018)
22. Liu, Y., Zeng, J., Shan, S., Zheng, Z.: Multi-channel pose-aware convolution neural net-works for multi-view facial expression recognition. In: Automatic Face & Gesture Recognition, pp. 458–465 (2018)

23. Lai, Y., Lai, S: Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In: Automatic Face & Gesture Recognition, pp. 263–270 (2018)
24. Blanz, V., Romdhani, S., Vetter, T.: Face identification across different poses and illuminations with a 3D morphable model. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 192–197 (2002)
25. Chang, F., Tran, A., Hassner, T., Masi, I., Nevatia1, R., Medioni, G.: ExpNet: landmark-free, deep, 3D facial expressions. arXiv preprint arXiv:1802.00542v1 (2018)
26. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
27. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 (2014)
28. Bulat, A. and Tzimiropoulos, G.: How far are we from solving the 2D & 3D Face Alignment problem? In: IEEE ICCV (2018)
29. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: IEEE CVPR (2013)
30. Parkhi, O. M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
31. 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
32. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)
33. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.: A 3D facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition, pp. 211–216 (2006)
34. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. IEEE TIP **24**(1), 189–204 (2015)
35. Zhang, F., Zhang, T, Mao, Q., Xu, C.: Joint pose and expression modeling for facial expression recognition. In: IEEE CVPR (2018)
36. Zhang, T., et al.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimedia **18**(12), 2528–2536 (2016)