





Improving Ensemble Learning Performance with Complementary Neural Networks for Facial Expression Recognition

Xinmin Zhang  and Yingdong Ma ^(✉) 

The School of Computer Science, Inner Mongolia University, Hohhot, China
csmyd@imu.edu.cn

Abstract. Facial expression recognition has significant application value in fields such as human-computer interaction. Recently, Convolutional Neural Networks (CNNs) have been widely utilized for feature extraction and expression recognition. Network ensemble is an important step to improve recognition performance. To improve the inefficiency of existing ensemble strategy, we propose a new ensemble method to efficiently find networks with complementary capabilities. The proposed method is verified on two groups of CNNs with different depth (eight 5-layer shallow CNNs and twelve 11-layer deep VGGNet variants) trained on FER-2013 and RAF-DB, respectively. Experimental results demonstrate that the proposed method achieves the highest recognition accuracy of 74.14% and 85.46% on FER-2013 and RAF-DB database, respectively, to the best of our knowledge, outperforms state-of-the-art CNN-based facial expression recognition methods. In addition, our method also obtains a competitive result of the mean diagonal value in confusion matrix on RAF-DB test set.

Keywords: Convolutional Neural Networks · Ensemble learning
Expression recognition

1 Introduction

Facial Expression Recognition (FER) analyzes the category (e.g., happiness, sadness) of human expression based on face recognition. FER has been widely studied as accurate recognition of human facial expression is a fundamental step for many computer vision applications, such as medical security and human-computer interaction. Significant progress has been made in the last decade [1–5]. However, FER is a difficult task due to various illumination conditions, head position and occlusion in different face images. If feature extraction is carried out directly using these raw data, it would increase feature extraction error and eventually reduce FER performance. As a result, before feature extraction, preprocessing of facial images is necessary, such as face recognition, facial landmarks detection, face registration, histogram equalization, etc.

Despite the continues research efforts, FER under uncontrolled environment is still a challenging problem [5]. So far, most top performance approaches tend to utilize shallow neural networks with ensemble learning methods [5–7]. Ensemble of networks

not only makes use of strong feature learning ability of neural networks, but also explores the ability of different networks to complement each other during ensemble learning. As a result, ensemble of multiple networks usually has better FER performance than single classifier based methods. However, these methods have three main limitations: (1) shallow networks need more training overhead than deep networks to reach the same training termination condition; (2) because of the weak fitting ability, shallow networks are often inferior to deep networks in terms of performance; (3) most ensemble learning methods utilize all trained networks to make final decisions. But according to our experiment, ensemble of all networks does not necessarily achieve optimal performance. To solve these problems, in this paper, we propose a new ensemble learning method which combines complementary CNNs to achieve high performance with less time consumption. The method framework is shown in Fig. 1. The main steps of this method are summarized as follows:

- Twenty CNNs (including twelve deep CNNs and eight shallow CNNs) are trained as the candidate network set.
- An optimal deep network is selected to form our baseline system according to recognition performance.
- Candidate networks are added to or removed from our system until the best performance is achieved.

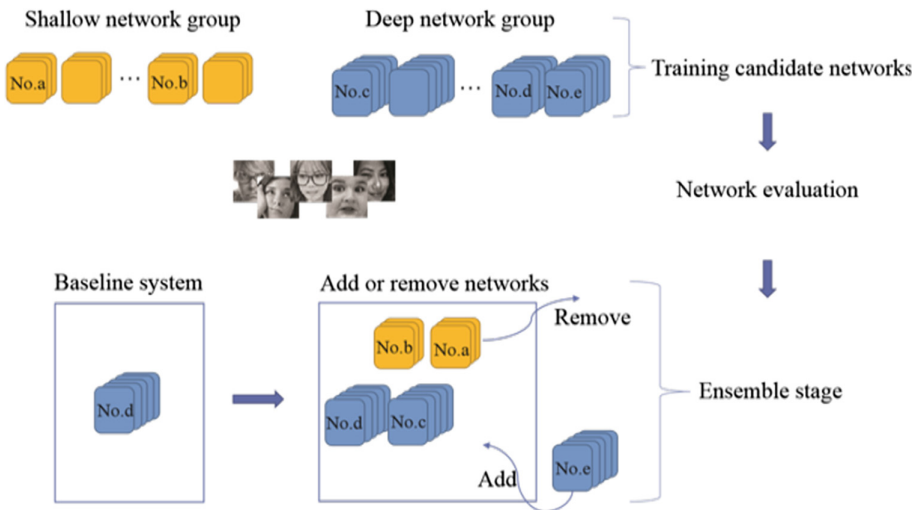


Fig. 1. Overview of our ensemble method.

The proposed method is evaluated on two real-world facial expression databases (FER-2013 [8] and RAF-DB [9]). To the best of our knowledge, our method outperforms state-of-the-art top performing works on FER-2013 and RAF-DB databases.

2 Related Works

2.1 Facial Landmarks Detection and Expression Recognition

Face images obtained under non-restrictive settings tend to have different degrees of occlusions and varied postures. To extract accurate facial features from these images, facial landmarks detection is usually required. Xiong and Torre proposed a Supervised Descent Method for minimizing a Non-linear Least Squares function. They also proposed a well-defined alignment error function which can be minimized using existing algorithms [10]. Sun et al. proposed an effective three-layer CNNs cascaded for facial landmarks detection [11]. Ren et al. learned a set of highly discriminative local binary features for each facial landmark independently [12]. These features are then used to jointly learn a linear regression to quickly locate facial landmarks. Zhu et al. proposed a 3D Dense Face Alignment and used cascaded CNNs to handle face alignment in the case of large pose variations and self-occlusions [13].

Kim et al. utilized alignable faces and non-alignable faces to improve FER performance [7]. They designed an alignment-mapping network to learn how to generate aligned faces from non-aligned faces. Rudovic et al. proposed a probabilistic method to implement facial expression recognition using head pose invariant [14]. The method performed head pose estimation, head pose normalization and facial expression recognition based on 39 facial landmarks.

2.2 Neural Networks

Krizhevsky et al. [15] proposed an eight-layer CNN in 2012 and made breakthrough progress in image classification. Because of their powerful feature representation ability, neural networks have been successfully applied to many computer vision applications, such as speech recognition [16] and semantic segmentation [17]. Recently, several FER methods utilized deep neural networks for improving performance. Liu et al. proposed a Boosted Deep Belief Network framework to carry out feature learning, feature selection and classifier construction iteratively [18]. Mollahosseini et al. proposed a deep neural architecture which applied the Inception layer [19] to address FER problem across multiple standard face databases [20]. In [21], Tang showed that significant gains can be obtained on several deep learning databases by simply replacing softmax with L2-SVMs. Meng et al. proposed an identity-aware convolutional neural network to alleviate high inter-subject variations [22]. They introduced an expression-sensitive contrastive loss and an identity-sensitive contrastive loss to show that learning features are not influenced by the variations of facial expression and different subjects. Vo et al. proposed CNN-based method to detect global and local facial expression features [23]. In their work, global features were computed to obtain possible candidate classification results for a face, and then, local features were utilized to reorder the previously obtained candidates to yield final recognition results.

2.3 Ensemble Learning

Ensemble learning builds a hypothesis set by training a series of learners [24]. It has been studied for a long time towards ensemble multiple neural networks in different visual fields [25–28]. As different neural networks provide complementary decision-making information, theoretically, the more diverse training networks are, the better performance they will be. Data preprocessing and different training configuration schemes can lead to network diversity (e.g., using different training sets, whether to adopt the dropout strategy [29]). In recent FER studies, combination of deep learning and ensemble learning has made remarkable progress. Yu and Zhang trained six 8-layer CNNs and automatically learned the ensemble weights among these networks by optimizing two loss functions [5]. Kim et al. constructed a hierarchical committee architecture with exponentially weighted decision fusion [6]. They combined nine 5-layer shallow CNNs with three 3-layer MLP classifiers (trained using features extracted from three alignment-mapping networks) in test stage [7]. Images in the training and test set were divided into alignable faces and non-alignable faces. The results on FER-2013 database showed that combination of alignable faces and non-alignable faces can improve FER performance.

3 Proposed Approach

3.1 Problem Analysis

Depth of Networks. In general, adding network layers leads to significant increasing of network parameters. As a result, it increases the training overhead in time and space. For this reason, many works have limited the training to shallow networks [5–7]. However, we observe counter-examples in our experiments. For example, when using “Xavier” [30] for parameter initialization and “ReLU” for activation to train FER-2013 database, shallow networks spend more training time than deep networks. Nevertheless, these networks do not get expected performance improvement. This fact shows that considering the time overhead and recognition accuracy, we should primarily train deep networks.

As some literatures have pointed out, the diversity of networks affects ensemble performance. However, to our best knowledge, most works did not explore the diversity of network depth. We believe that ensemble learning performance can be improved if shallow networks can also be trained to utilize network diversity.

Ensemble Strategy. Ensemble of all networks does not necessarily achieve optimal performance, which is mainly based on the following consideration: some networks do not provide complementary capabilities to other networks. In this case, addition of more networks might introduce negative effects for samples which had been predicted correctly.

3.2 Configurations of All Networks

Considering different data preprocessing, parameter initialization, activation function, training settings and network layer settings can lead to a variety of network models, we train eight 5-layer networks (shallow CNNs) and twelve 11-layer VGGNet [31] variants (deep CNNs) for ensemble stage.

The forward propagation process of 5-layer shallow CNNs is shown in Fig. 2. The architecture can be simplified as CPCPCPDF (C, P, F, and D stands for Convolution, Pooling, Fully connected layer, and Dropout, respectively). The detailed configurations of 5-layer networks are summarized in Table 1. All of these networks use ReLU [32] as activation function.

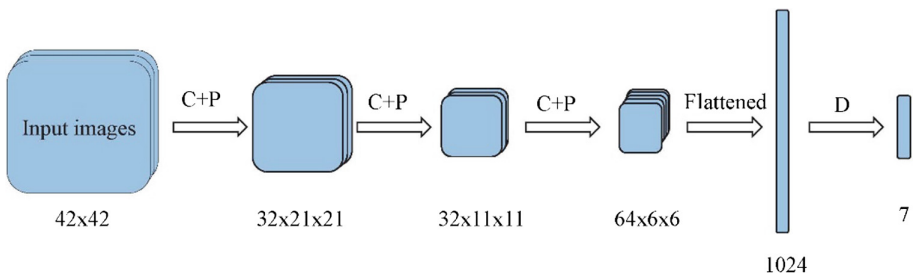


Fig. 2. The forward propagation process of 5-layer shallow CNNs.

Table 1. Configurations of eight 5-layer networks. Raw: Raw train data. Hist: Histogram equalization. Prep: Preprocessing methods. Stand: Standardization. M-M: Maximum-Minimum normalization. WIni: Weight Initialization. TruN: Truncation Normal distribution. Xav: Xavier initialization [30]. WRe: Weight Regularization. FCDrop: Dropout strategy used in Fully Connected layer (FC). FC₁: The first Fully Connected layer.

Config	Data	Prep	WIni	WRe	FCDrop
1	Raw	Stand	TruN	0.0001	FC ₁ = 0.5
2	Raw	Stand	Xav	0.0001	FC ₁ = 0.5
3	Raw	M-M	TruN	0.0001	FC ₁ = 0.5
4	Raw	M-M	Xav	0.0001	FC ₁ = 0.5
5	Hist	Stand	TruN	0.0001	FC ₁ = 0.5
6	Hist	Stand	Xav	0.0001	FC ₁ = 0.5
7	Hist	M-M	TruN	0.0001	FC ₁ = 0.5
8	Hist	M-M	Xav	0.0001	FC ₁ = 0.5

The forward propagation process of 11-layer VGGNet variants is shown in Fig. 3. Their architecture can be expressed as 4*(CCPD)FDF (4* indicates repeat four times). The detailed configurations of 11-layer CNNs are summarized in Table 2. The activation process of 11-layer CNNs uses BN+ReLU, ReLU+BN, and ReLU, respectively.

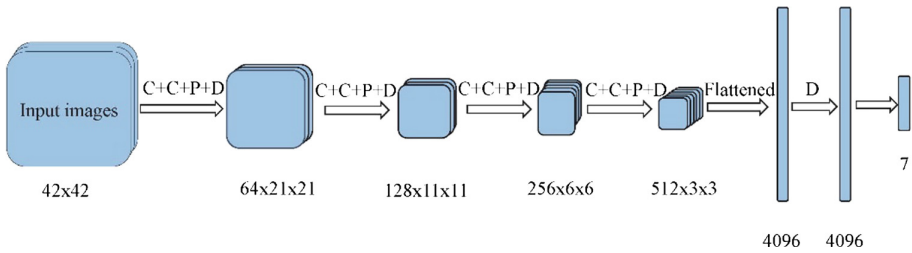


Fig. 3. The forward propagation process of 11-layer VGGNet variants.

Table 2. Configurations of twelve 11-layer VGGNet variants. BN: Batch Normalization [33]. Act: Activation function. BN+ReLU: Execute BN first, then ReLU. [ReLU+BN]: Execute ReLU first, then BN for all layers except for last FC layer (only ReLU). CCP: Successive Convolution, Convolution, and Pooling. CCPDrop: Dropout strategy used after every CCP.

Config	Data	Prep	Wini	Act	CCPDrop
9	Raw	Stand	Xav	BN+ReLU	0.2
10	Raw	Stand	Xav	[ReLU+BN]	0.2
11	Raw	Stand	Xav	ReLU	0.2
12	Raw	M-M	Xav	BN+ReLU	0.2
13	Raw	M-M	Xav	[ReLU+BN]	0.2
14	Raw	M-M	Xav	ReLU	0.2
15	Hist	Stand	Xav	BN+ReLU	0.2
16	Hist	Stand	Xav	[ReLU+BN]	0.2
17	Hist	Stand	Xav	ReLU	0.2
18	Hist	M-M	Xav	BN+ReLU	0.2
19	Hist	M-M	Xav	[ReLU+BN]	0.2
20	Hist	M-M	Xav	ReLU	0.2

In training stage, we use exponential decay learning to update a new learning rate. The learning rate of a network is updated as:

$$\eta = \eta_0 * (0.99)^N \quad (1)$$

where η_0 denotes the initial learning rate, η represents new learning rate, and N is the number of epochs. In order to fully learn the feature of training samples, a more severe termination condition must be satisfied to stop the training, that is, the training error of a batch does not exceed 10^{-6} for three consecutive times or the training reaches maximum number of iterations.

3.3 Ensemble Method

According to the above analysis, combination of complementary CNNs improves system performance. It can be achieved by gradually adding networks that improve

recognition accuracy and removing networks which cause system performance degradation.

In general, the fitting ability of deep networks is better than that of shallow networks. Therefore, at the beginning, we select a network with the best accuracy from deep network group as our baseline system.

In the next step, candidate networks from all shallow and deep networks are added to or removed from baseline system step by step until the best performance is achieved. The system ensemble mode in this paper is majority vote. It is important to note that the evaluation scores of all networks on validation and test data have been calculated in advance, so we do not need to spend a long time in the process of ensemble selection. All we need to do is matrix addition.

4 Experiments on the FER-2013 Database

4.1 FER-2013

FER-2013 [8] is one of the largest facial expression databases so far. It has 28,709 images for training, 3,589 images for public test, and 3,589 images for private test. To reduce training errors, we remove 46 non-face images and 11 non-number filled images from original database.

We use IntraFace [10, 34] to detect facial landmarks. We label an image as Non-Alignable Faces (NAF) if its detection score is smaller than a given threshold, and otherwise, label it as Before Registered Alignable Faces (BRAf). The affine transformation principle is applied to adjust two eyes to horizontal position. We refer to the After Registered Alignable Faces as ARAf.

Data increment is implemented for training, validation and test set following the method introduced in [7]. Specifically, 10 times increment are used in this work (four $42 * 42$ corners and a resize of original image, as well as their horizontal flip images).

4.2 Training and Evaluation

In training stage, the initial learning rate of shallow and deep network group is set to 0.05. The maximum number of iterations for shallow and deep network groups is 600000 and 200000, respectively.

During validation and testing stage, the score of each image is the mean of 10 corresponding incremental images. For Alignable Faces (AF), we evaluate Before Registered Alignable Faces (BRAf) and After Registered Alignable Faces (ARAf) respectively and average the two values. After evaluating Non-Alignable Faces (NAF), results of all validation (testing) samples are combined using the following formula:

$$\text{acc} = \text{acc}(\text{AF}) * \alpha + \text{acc}(\text{NAF}) * \beta \quad (2)$$

where α is the proportion of alignable faces in validation (testing) set, and β is the proportion of non-alignable faces in validation (testing) set.

4.3 Ensemble and Analysis

For FER-2013, we conduct network ensemble experiments on validation set. After determining the optimal network combination, testing set is used as the final performance evaluation.

Baseline System. In deep network group, network No. 11 is selected as the baseline system as it has the highest validation accuracy (70.52%).

Ensemble Process. In this experiment, ensemble of all deep and shallow networks is utilized to explore the change of system performance. Candidate networks are selected from deep and shallow network groups. At the beginning, network No. 18 is selected as combination of No. 18 and baseline system set has top performance (71.79%). In the second step, network No. 13 is selected as combination of network No. 13 and new system has best performance (72.35%). This process continues until system performance is no longer growth. In the seventh step, after removing network No. 3, the highest performance (72.64%) is obtained. The ensemble process is summarized in Table 3. We observe performance reduction when more networks are added. For example, ensemble of all 20 networks yields 72.30% validation accuracy. Finally, the system achieves 74.14% test accuracy with an ensemble of five deep CNNs.

Table 3. Ensemble process on FER-2013.

Steps	System	Acc	Select	Candidate
1	11	70.52	18	1–20
2	11 18	71.79	13	1–20
3	11 18 13	72.35	9	1–20
4	11 18 13 9	72.42	12	1–20
5	11 18 13 9 12	72.64	3	1–20
6	11 18 13 9 12 3	72.62	–	1–20
7#	11 18 13 9 12	72.64	3	1–20

To prove the feasibility of our method, we list ensemble accuracy of the shallow network group, the deep network group, and all networks on the validation set in Table 4.

Table 4. Performance comparison of different combinations on FER-2013.

Networks	Accuracy
Shallow network group	70.52
Deep network group	72.16
All 20 networks	72.30

Result Analysis. Table 5 lists performance comparison of ours and state-of-the-art works on the FER-2013 database. The proposed method only combines five deep

CNNs (11-layer) to achieve 74.14% test accuracy. To our best knowledge, this method outperforms other state-of-the-art CNN-based FER methods. Moreover, the proposed method is efficient than other methods. The method is implemented on a personal computer with i7-7700k CPU, 16 GB memory and a GTX 1080Ti GPU. The average time to process a test image using a shallow network and a deep network is 12.3 ms and 14.1 ms, respectively. Ensemble of five deep networks consumes 72.7 ms. As a contrast, [5] and [7] spend 76.3 ms and 146.6 ms to process the same image on our personal computer.

Table 5. Performance comparison of the proposed method and state-of-the-art works on FER-2013.

	Methods	Accuracy	Average time (ms)
[21]	A DCN using L2-SVM Loss.	71.16%	–
	A DCN using cross-entropy Loss	70.1%	–
[5]	Ensemble of six 8-layer CNNs using learned weights	72%	76.3
[6]	Ensemble of 36 DCNs in a hierarchical committee	72.72%	–
[7]	Ensemble three MLP classifiers and nine 5-layer CNNs	73.73%	146.6
Ours	Five 11-layer CNNs	74.14%	72.7

5 Experiments on the RAF-DB Database

5.1 RAF-DB

RAF-DB [9] is also a real-world facial expression database that used the crowdsourcing technology for facial annotation. The database contains about 30000 images of basic 7 single-class expressions and 11 compound expressions. In our experiment, we use only 15339 registrated images of single-class expressions, including 12271 training images and 3068 test images. We tripled the training and test set, including an original image, and its horizontal mirror and vertical mirror.

5.2 Training and Evaluation

In training stage, the initial learning rate of shallow and deep network group is set to 0.01 and 0.05, respectively. The maximum number of iterations for shallow and deep network group is 200000 and 20000, respectively.

During testing stage, the score of each image is the mean of three corresponding incremental images.

5.3 Ensemble and Analysis

Baseline System. Similar to FER-2013, the best performing network No. 19 (83.41%) from deep network group is selected as baseline system.

Ensemble Process. Candidate networks are also selected from deep and shallow network groups. In the third step, the system performance increases to 85.46% when three networks are combined. Please see Table 6 for detail information. Since then, adding more networks leads to system performance reduction. For instance, adding network No. 19 reduces system performance from 85.46% to 85.23%. However, the highest ensemble performance can be observed after removing network No. 19 from system network set. Finally, the system achieves 85.46% accuracy with an ensemble of two deep CNNs and one shallow CNNs.

Table 6. Ensemble process on RAF-DB.

Steps	System	Acc	Select	Candidate
1	19	83.41	13	1–20
2	19 13	84.88	3	1–20
3	19 13 3	85.46	19	1–20
4	19 13 3 19	85.23	–	1–20
5#	19 13 3	85.46	19	1–20

In Table 7, we list the ensemble performance of shallow network group, deep network group, and all 20 networks on the RAF-DB database.

Table 7. Performance comparison of different combinations on RAF-DB.

Networks	Accuracy
Shallow network group	83.54
Deep network group	84.39
All 20 networks	84.42

Table 8. Our method is compared with the existing methods on two evaluation criteria: diagonal average of confusion matrix (Ave) and recognition accuracy (Acc). The results of center loss [35] + LDA, center loss + mSVM, DLP-CNN [9] + LDA and DLP-CNN + mSVM are tested in [9]. Seven numbers in the second line represent the number of samples of different expressions on original training set. Sur: Surprise, Fea: Fear, Dis: Disgust, Hap: Happy, Ang: Anger, Neu: Neutral.

Methods	Sur	Fea	Dis	Hap	Sad	Ang	Neu	Ave	Acc
	1290	281	717	4772	1982	705	2524		
Our	80.24	47.30	45	94.68	82.22	74.07	90.59	73.44	85.46
center loss + LDA	76.29	54.05	49.38	92.41	74.90	64.81	77.21	69.86	79.96
center loss + mSVM	79.63	54.05	53.13	93.08	78.45	68.52	83.24	72.87	82.86
DLP-CNN + LDA	74.07	52.50	55.41	90.21	73.64	77.51	73.53	70.98	78.81
DLP-CNN + mSVM	81.16	62.16	52.15	92.83	80.13	71.60	80.29	74.20	82.84

Result Analysis. The proposed ensemble method not only achieves the best recognition accuracy, but also have competitive results for the average accuracy of seven single-class expressions (the mean diagonal value of confusion matrix). After ensemble of three networks, the values of diagonal in the confusion matrix are shown in Table 8. As shown in the table, four existing methods are listed for comparison. All of them apply different loss functions to train neural networks, and then use feature vectors extracted to train LDA and SVM classifier. In contrast, our method only uses softmax loss for training, and directly uses neural networks to present competitive classification performance.

6 Conclusion

In this paper, we propose a new ensemble learning based method for improving facial expression recognition. Specifically, two groups of CNNs (eight 5-layer CNNs and twelve 11-layer CNNs) are trained with various configurations. On this basis, a new network ensemble method is proposed to combine complementary CNNs to improve FER performance. Extensive experiments on FER-2013 and RAF-DB show that the proposed method achieves excellent recognition accuracy with less time overhead. Performance comparison of the proposed method and state-of-the-art works demonstrates that our method reaches the best recognition accuracy (74.14%) on the FER-2013 database. On RAF-DB database, our ensemble method also achieves the highest recognition accuracy (85.46%) and competitive performance of diagonal mean value of confusion matrix (73.44%) without complicated training process.

References

1. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: IEEE International Conference on Image Processing, vol. 2, pp. II-370 (2005)
2. Liu, W., Wang, Z.: Facial expression recognition based on fusion of multiple Gabor features. In: 18th International Conference on Pattern Recognition, vol. 3, pp. 536–539 (2006)
3. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **6**(1), 1–12 (2015)
4. 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)
5. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442 (2015)
6. Kim, B.K., Roh, J., Dong, S.Y., Lee, S.Y.: Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *J. Multimodal User Interfaces* **10**(2), 173–189 (2016)
7. Kim, B.K., Dong, S.Y., Roh, J., Kim, G., Lee, S.Y.: Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 48–57 (2016)

8. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. *Neural Netw.* **64**, 59–63 (2015)
9. Li, S., Deng, W., Du, J.P.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2584–2593 (2017)
10. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 532–539 (2013)
11. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476–3483 (2013)
12. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692 (2014)
13. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3D solution. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 146–155 (2016)
14. Rudovic, O., Pantic, M., Patras, I.: Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1357–1369 (2013)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
16. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Sig. Process. Mag.* **29**(6), 82–97 (2012)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015)
18. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805–1812 (2014)
19. Szegedy, C., et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
20. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: *IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10 (2016)
21. Tang, Y.: Deep learning using linear support vector machines. *Comput. Sci.* (2013)
22. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 558–565 (2017)
23. Vo, D.M., Sugimoto, A., Le, T.H.: Facial expression recognition by re-ranking with global and local generic features. In: *23rd International Conference on Pattern Recognition*, pp. 4118–4123 (2016)
24. Zhou, Z.H.: Ensemble learning. In: Li, S.Z. (ed.) *Encyclopedia of Biometrics*, vol. 1, pp. 270–273. Springer, Berlin (2009)
25. Hansen, L.K.: Neural network ensemble. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990)
26. Guan, Y., Li, C.T., Roli, F.: On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(7), 1521–1528 (2015)

27. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Pedestrian detection with spatially pooled features and structured ensemble learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(6), 1243–1257 (2016)
28. Ding, C., Tao, D.: Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 1002–1014 (2018)
29. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. *Comput. Sci.* **3**(4), 212–223 (2012)
30. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256 (2010)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Comput. Sci.* (2014)
32. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814 (2010)
33. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, pp. 448–456 (2015)
34. Fernando, D.L.T., Chu, W.S., Xiong, X., Vicente, F., Ding, X., Cohn, J.: Intraface. In: *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8 (2015)
35. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016. LNCS*, vol. 9911, pp. 499–515. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_31