Discriminative Feature Learning Using Two-Stage Training Strategy for Facial Expression Recognition

Lei Gan[1], Yuexian Zou[1,2(✉)], and Can Zhang[1]

[1] ADSPLAB, School of ECE, Peking University, Shenzhen, China
{ganlei,zouyx,zhangcan}@pku.edu.cn
[2] Peng Cheng Laboratory,Shenzhen, China

Abstract. Although deep convolutional neural networks (CNNs) have achieved the state-of-the-arts for facial expression recognition (FER), FER is still challenging due to two aspects: class imbalance and hard expression examples. However, most existing FER methods recognize facial expression images by training the CNN models with cross-entropy (CE) loss in a single stage, which have limited capability to deal with these problems because each expression example is assigned equal weight of loss. Inspired by the recently proposed focal loss which reduces the relative loss for those well-classified expression examples and pay more attention to those misclassified ones, we can mitigate these problems by introducing the focal loss into the existing FER system when facing imbalanced data or hard expression examples. Considering that the focal loss allows the network to further extract discriminative features based on the learned feature-separating capability, we present a two-stage training strategy utilizing CE loss in the first stage and focal loss in the second stage to boost the FER performance. Extensive experiments have been conducted on two well-known FER datasets called CK+ and Oulu-CASIA. We gain improvements compared with the common one-stage training strategy and achieve the state-of-the-art results on the datasets in terms of average classification accuracy, which demonstrate the effectiveness of our proposed two-stage training strategy.

Keywords: Convolutional neural networks · Facial expression recognition · Two-stage training strategy · Discriminative feature learning

1 Introduction

Recently, facial expression recognition (FER) has received increasing attention [1,5,10–12] due to its wide range of applications among health care, social interaction, human-computer-interaction systems, and etc. Essentially, FER is a multi-class classification problem, which aims at classifying expression images as one of several basic expression labels, i.e., anger, disgust, fear, happiness, sadness,

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11729, pp. 397–408, 2019.
https://doi.org/10.1007/978-3-030-30508-6_32

Fig. 1. Illustration of class imbalance and easy/hard expression examples. The pie chart (left) shows the class imbalance problem, i.e., different expressions account for different proportion. The picture (right) indicates the hard expression examples problem. The easy expression example can be obviously recognized as happiness expression. In comparison, the hard expression example is too ambiguous to be directly recognized.

surprise, which are first defined in [2] and have been universally adopted to represent facial expressions.

There are two problems existing in the field of FER including class imbalance and hard expression examples, as illustrated in Fig. 1. On the one hand, the imbalance problem exists because some expressions show frequently, such as *happiness*, meanwhile, some expressions rarely display, such as *anger*. On the other hand, people sometimes display some subtle expressions which are usually too ambiguous to be correctly recognized. For example, the *disgust* and *sadness* expressions can be mixed with each other because they display similarly sometimes. The two problems limit the performance of the FER system.

In this paper, we attempt to mitigate these problems by exploiting the recently proposed focal loss [8] and presenting a two-stage training strategy using CE loss in the first stage and focal loss in the second stage. The focal loss contains two hyper-parameters α and γ, targeting at the class imbalance and hard expression examples problems, respectively. Besides, the learned model after the first training stage can be utilized to further explore discriminative features in the learning process.

2 Related Work

Recently, many studies have been conducted to recognize facial expression from raw images. In [15], a two-stage fine-tuning method follows a transfer learning approaches. Based on a network pre-trained on the ImageNet [7] dataset, it trains on a larger FER dataset in the first stage and then narrows down to train on a smaller FER dataset in the second stage. In [14], three inception convolutional structures are constructed in order to classify the registered facial images into several expression labels, indicating the effectiveness of the inception layers for the FER problem.

In [12], both contrastive loss and softmax loss are jointly utilized to learn features for FER, which aims at developing effective feature representations for identity-invariant FER by balancing the distribution of intra- and inter- class variations.

However, most of the previous methods fail to either incorporate the previously learned information or pay attention to the class imbalance and hard expression examples problems. In this paper, we delicate to investigating the potential improvements using the two-stage training strategy, utilizing the CE loss in the first stage and the focal loss in the second stage.

3 Our Approach

We apply a deep learning method to recognize facial expression. Specifically, We adopt the popular inception-v3 [18] convolutional neural network as our basic network architecture due to its good balance between accuracy and efficiency on many image classification tasks. Most one-stage training strategies apply CE loss already achieve competitive results. Our aim is to further extract discriminative features by focusing on the class imbalance and hard expression examples problems. Thus, we exploit a recently proposed focal loss subsequently and propose a two-stage training strategy utilizing the cross-entropy (CE) loss in the first stage and the focal loss in the second stage.

3.1 CE Loss

Most existing CNN-based FER methods train the model using cross-entropy (CE) loss, which perform reasonably well on the FER task in terms of the average classification accuracy. However, it actually has limited capability to tackle the class imbalance and hard expression examples problems because each training example is assigned equal weight of loss. The cross-entropy (CE) loss is defined in (1) as follows:

$$Loss_1 = -\sum_{i=1}^{c} y_i log(\tilde{y_i}) \tag{1}$$

where y_i is the i-th value of the ground truth label, and $\tilde{y_i}$ is the i-th output value of the softmax of the network. c is the total number of expression classes. The $\tilde{y_i}$ is defined as (2) using logit values of the network:

$$\tilde{y_i} = \sigma(l_i) \tag{2}$$

where l_i is the i-th logit value of the network, and the $\sigma(\cdot)$ is a softmax activation function.

3.2 Focal Loss

Focal loss was first proposed in [8] by reshaping the CE loss. Specifically, it introduces a weighting factor α and a focusing parameter γ to the cross-entropy loss, as formulated in (3):

$$Loss_2 = -\alpha \sum_{i=1}^{c} (1 - \tilde{y_i})^{\gamma} y_i log(\tilde{y_i}) \tag{3}$$

where $\alpha \in [0, 1]$ and $\gamma \geq 0$ are two hyper-parameters. α is used to balance the importance among imbalanced examples, and γ is used to differentiate between easy and hard expression examples. The y_i and $\tilde{y_i}$ are of the same meanings as in the CE loss. Final decision $\tilde{o}$ is obtained in (4) using the softmax output of the network:

$$\tilde{o} = arg \max_i \tilde{y_i} \tag{4}$$

We introduce the focal loss to the existing FER system in order to mitigate the class imbalance and hard expression examples problems.

3.3 Two-Stage Training Strategy

Considering that the focal loss can help the network to further extract discriminative features based on the learned feature-separating capability, we propose a two-stage training strategy aiming at mitigating the class imbalance and hard expression examples problems. Since most existing deep learning-based FER methods have performed reasonably well training in a single stage, our training strategy utilizes CE loss in the first stage and focal loss in the second stage in order to boost the FER performance.

For the class imbalance problem, the hyper-parameter α balances the importance of the examples in different categories. In this way, the total loss will not be dominated by those categories which have the number advantage. For the hard expression examples problem, the focusing hyper-parameter γ adjusts the rate of reducing the relative loss for those well-classified examples and concentrates more on those misclassified examples.

For further illustrate the effectiveness of our proposed two-stage training stage, we also perform extensive ablation study in §4.4 to give enough evidence.

4 Experiments and Analysis

We conduct extensive experiments in a supervised training manner on two well-known FER datasets: CK+ [13] and Oulu-CASIA [20], and compare our two-stage training strategy with other state-of-the-art algorithms for FER. Details of each dataset are tabulated in Table 1. The AN, CO, DI, FE, HA, SA, SU are short for Anger, Contempt, Disgust, Fear, Happiness, Sadness and Surprise, respectively. Some images from each dataset are also listed in Fig. 2.

Table 1. The number of images for each expression in the datasets

	AN	CO	DI	FE	HA	SA	SU	All
CK+	135	54	177	75	147	84	249	981
Oulu-CASIA	240	-	240	240	240	240	240	1440

Fig. 2. Cropped facial expression images from the CK+ (1^{st} row) and Oulu-CASIA dataset (2^{nd} row). From the left to the right column, the displaying expressions are anger, disgust, fear, happiness, sadness, surprise and contempt, respectively.

4.1 Implementation Details

Preprocessing. After detecting and cropping face from a raw image by applying the MTCNN [19] module, we perform various data augmentation techniques to obtain more training data for each epoch and make our trained model more robust. In the training session, each image processed after the MTCNN module is first resized to 299×299, then horizontally flipped with a probability of 0.5, flipped in the vertical direction with a probability of 0.5, and randomly rotated by $10°$. In the testing phase, the processed image after MTCNN module is resized and cropped at the center into size 299×299.

Initialization. The weights of convolutional layers are initialized from the adopted Inception-v3 model pre-trained on ImageNet [7]. The weights of the last fully-connected layer in each training stage are randomly initialized using Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. We set the value of μ and σ to 0 and 0.001, respectively. The output neuron number of the last fully-connected layer is modified as the total number of expression classes.

Training Setup. The network is trained with a mini-batch size of 32 images using back-propagation method and Stochastic Gradient Descent (SGD) optimization algorithm in a supervised training manner. The momentum is 0.9. Also, the dropout method is used for reducing the over-fitting problem, and the dropout ratio is set to 0.5. The weight decay is also used for regularization with a factor of 0.001.

Following the previous work [5,10], a 10-fold subject-independent cross validation protocol is adopted for evaluation in all the experiments in order to avoid the same subject existing in both the training and testing data. Nine subsets are used for training while the remaining subset is used for validation. In the first training stage, the whole network is trained for 80 epochs in total with an initial learning rate of 0.01. In the second training stage, the network is further trained starting from the best model in the first stage, and the training stops at the 80th epoch. The learning rate in this stage is set as 0.01 at first and decay $10\times$ after 60-th epoch. All the settings are the same in the experiments. Our

proposed two-stage training strategy is implemented with PyTorch deep learning framework on a single NVIDIA GeForce GTX 1080 GPU with 8 GB memory.

Testing and Evaluation Metric. Given a raw image in the testing phase, the network will predict the final expression label of it. we adopt average classification accuracy as the evaluation metric, which is widely used in the field of FER research.

4.2 Experiments on the CK+ Dataset

Dataset Description. The Extended Cohn-Kanade Dataset (CK+) [13] is a widely used dataset for facial expression recognition. It contains 118 subjects with 327 facial expression sequences in total, ranging from 18 to 30 years old. Each image sequence is annotated with one discrete expression label out of seven expressions, including six basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) and one non-basic expression (contempt). In this dataset, each image sequence begins with a neutral expression and gradually reaches a peak expression at the last frame. The same as [10], we construct our image-based CK+ dataset by extracting the last three frames from each image sequence. Finally, the image-based CK+ dataset consists of 981 images with the resolution of 640×490. This dataset is divided into ten subject-independent subsets by sampling subject ID in ascending order with a step size of 10.

Class Imbalance Problem. From Table 1, we can learn that the class imbalance problem seriously exists in the CK+ dataset, in other words, images number varies among different expression classes. Clearly, the three expressions with least images are contempt, fear and sadness, respectively. In the experiments, the hyper-parameters of α and γ are experimentally set as 0.25 and 0 respectively on this dataset.

Results. The confusion matrix comparisons on the CK+ dataset are reported in Fig. 3. It compares the confusion matrix of the baseline (left) using the CE loss and our two-stage training strategy (right) using the CE loss in the first stage and the focal loss in the second stage. From the results reported from Fig. 3, we observe that the contempt expression is the most difficult one to be classified if we train the network with only the CE loss, because it has the least training data and the network with only the CE loss has limited ability to tackle the class imbalance problem. In contrast, after further training the whole network driven by the focal loss, the network acquires enhanced feature-discriminating capability by consequently paying attention to the expressions with less training images. Clearly, it can be observed from the right confusion matrix in Fig. 3 that the contempt expression is further perfectly classified with an accuracy increase of 5.6%. Besides, for the fear and sadness expressions which account for the second least and third least proportion of all the dataset, the

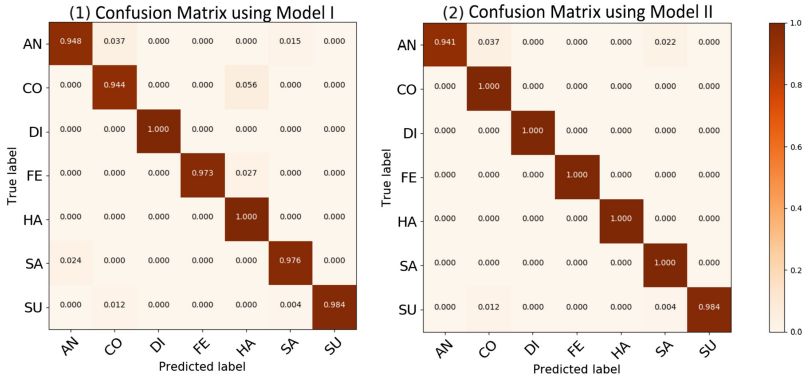

Fig. 3. Confusion matrix of our two-stage training strategy on the CK+ dataset in the first stage (left) and in the second stage (right). The darker the color is, the higher the accuracy reaches. The x-labels and y-labels stand for prediction results and ground truth, respectively. (Color figure online)

classification accuracy of them increase 2.7% and 2.4%, respectively, verifying that our proposed two-stage training strategy is able to learn more discriminative features by mitigating the class imbalance problem. As for the classification accuracy of the anger expression, which has a decrease of 0.7% after the second stage, we speculate that there may be some underlying factors that undermine the system performance, and we will discuss it later in our future work.

The overall accuracy of 10-fold cross validation on the CK+ dataset is shown in Table 2[1]. The best result is underlined in boldface and the second best result is marked with an underline. As shown in Table 2, although the CNN model with only the CE loss perform reasonably well compared with most FER methods, our two-stage training strategy can boost its performance by mitigating the class imbalance problem.

4.3 Experiments on the Oulu-CASIA Dataset

Dataset Description. Oulu-CASIA dataset [20] is a more challenging benchmark dataset for facial expression recognition. It contains 480 image sequences in total, including 80 subjects between 23 and 58 years old. Each image sequence has one of the six basic expression labels: anger, disgust, fear, happiness, sadness and surprise. It is captured by a VIS camera under three different illumination conditions: Strong, Dark and Weak. Similar to the CK+ dataset, each image sequence in this dataset starts from neutral expression to peak expression in the last few frames. Following the previous work in [1,5], we only use the image sequences captured under normal illumination. Then we construct the image-based Oulu-CASIA dataset with the resolution of 320 × 320 by collecting the

[1] For fair comparison, we compare our method with other state-of-the-arts which also use the SGD optimization algorithm.

Table 2. Overall accuracy of different methods on the CK+ dataset. The best result is underlined in bold face. The second best result is underlined.

Method	Accuracy (%)
3DCNN-DAP [9]	92.4
STM-Explet [10]	94.2
BDBN [11]	96.7
Exemplar-HMMs [17]	94.6
DTAGN(Joint) [5]	97.3
2B(N+M)Softmax [12]	97.1
Baseline	98.2
Ours (two stage)	**98.8**

Table 3. Overall accuracy of different methods on the Oulu-CASIA dataset. The best result is underlined in bold face. The second best result is underlined.

Method	Accuracy (%)
HOG 3D [6]	70.6
Atlases [3]	75.5
STM-Explet [10]	74.6
Exemplar-HMMs [17]	75.6
DTAGN(Joint) [5]	81.5
ExprGAN [1]	84.7
Baseline	87.7
Ours (two stage)	**88.3**

last three frames of each sequence and obtain 1440 images in total at last. Then we split the dataset into ten subject-independent groups by sampling subject ID with a step size of 8.

Hard Expression Examples Problem. The Oulu-CASIA dataset is challenging because it contains some hard expression examples that are too ambiguous to be recognized. Some hard expression examples in the Oulu-CASIA dataset are listed below in Fig. 4. In the experiments, the hyper-parameters of α and γ are experimentally set as 0 and 0.5 respectively on this dataset.

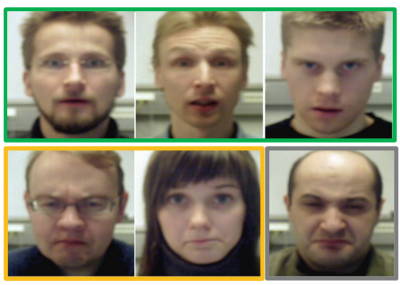

Fig. 4. Some hard expression examples. Images in the green, yellow and gray boxes are actually surprise, sadness and disgust expressions, respectively. (Color figure online)

Results. Detailed confusion matrix comparisons on the Oulu-CASIA dataset are presented in Fig. 5. It also compares the confusion matrix of the baseline (left) using the CE loss and our two-stage training strategy (right) using the CE loss in the first stage and the focal loss in the second stage. It is proved that our two-stage training strategy has the ability to distinguish the features extracted from

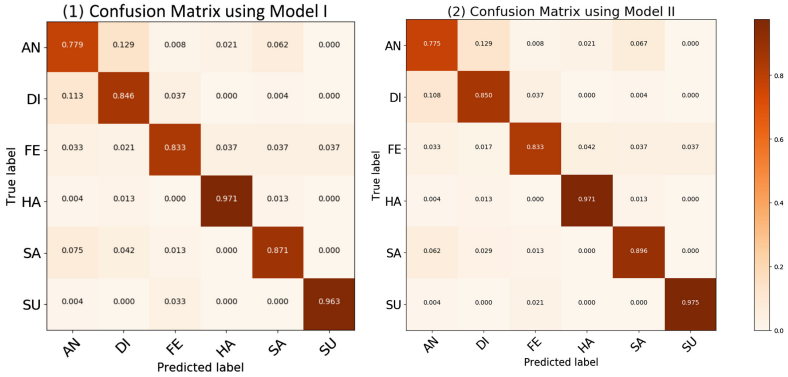

Fig. 5. Confusion matrix of our two-stage training strategy on the Oulu-CASIA dataset in the first stage (left) and in the second stage (right). The darker the color is, the higher the accuracy reaches. The x-labels and y-labels stand for prediction results and ground truth, respectively. (Color figure online)

hard expression examples. For example, the classification accuracy of sadness expression is 87.1% using only the CE loss in the first training stage, and it has a gain of 2.5% when further training using the focal loss in the second stage. Besides, for the surprise expression, our training strategy also gains an increase of 1.2%. Since there are few hard expression examples in the happy and fear expressions after analyzing this dataset, there is little improvements in terms of the classification accuracy of these expressions. For the classification of the anger expression which has a decrease of 0.4% after the second stage, we will investigate the potential influence in our future work.

The overall accuracy of 10-fold cross validation on the Oulu-CASIA dataset is shown in Table 3. The best result is underlined in boldface and the second best result is marked with an underline. Obviously, our two-stage training strategy has some advantages over the common one-stage training method using only the CE loss in terms of the overall classification accuracy.

4.4 Ablation Study

We conducted ablation study to explore and identify the best training strategy for the facial expression recognition (FER) task. Extensive experiments have been conducted on the CK+ and Oulu-CASIA datasets and we drew the conclusion that our proposed two-stage training strategy, utilizing the cross-entropy (CE) loss in the first stage and the focal loss in the second stage, is the best training strategy. Additionally, Our proposed training strategy also works well in other CNN models.

Training Strategy. To explore the best training strategy for the FER task, we conducted five different experiments on each dataset including one-stage training

with the focal loss only, one-stage training with the CE loss only, one-stage training with the joint CE loss and focal loss, and two-stage training with the focal loss followed by the CE loss or vice versa. Table 4 shows the detailed results of these experiments according to the 10-fold cross validation protocol. The results indicate that our proposed two-stage training strategy, utilizing CE loss in the first stage and focal loss in the second stage, achieves the best recognition performance in terms of the average classification accuracy.

Table 4. Accuracy comparisons using different training strategies with different loss in each training stage. ●: focal loss, ○: CE loss, ◐: joint CE loss and focal loss.

Dataset	1st stage	2nd stage	Accuracy (%)
CK+	●	✗	98.0
	○	✗	98.2
	◐	✗	97.8
	●	○	98.5
	○	●	**98.8**
Oulu-CASIA	●	✗	88.0
	○	✗	87.7
	◐	✗	87.1
	●	○	85.1
	○	●	**88.3**

Network Architecture. To further illustrate the effectiveness of our proposed two-stage training strategy, we conducted more experiments using two different CNN models, including the VGG-Face [16] and the Resnet-18 [4]. Detailed experimental results are shown in Table 5. It is observed that our proposed two-stage training strategy performs better than the one-stage training using only

Table 5. Accuracy comparisons using the VGG-Face and Resnet-18 network architecture with different loss in each training stage. ●: focal loss, ○: CE loss.

Network	Dataset	1st stage	2nd stage	Accuracy (%)
VGG-Face	CK+	○	✗	94.6
		○	●	97.2
	Oulu-CASIA	○	✗	83.3
		○	●	83.7
Resnet-18	CK+	○	✗	98.4
		○	●	98.6
	Oulu-CASIA	○	✗	87.8
		○	●	88.3

the CE loss in terms of average classification accuracy. Besides, the effectiveness of two-stage training strategy is independent of the CNN model we used.

5 Conclusion

In this paper, we introduce the focal loss to the existing FER system and propose a two-stage training strategy to recognize facial expression from raw images. Specifically, we utilize the cross-entropy (CE) loss in the first stage and the focal loss in the second stage. Due to the fact that the focal loss contains a weighting factor α and a focusing parameter γ, so the network is able to mitigate the class imbalance and hard expression examples problems. Extensive experiments on the CK+ and Oulu-CASIA datasets show that our proposed two-stage training strategy achieves improved performance compared with one-stage training with only the CE loss. Additionally, our proposed training strategy also works well in other CNN models. We believe this work will provide a new perspective in capturing discriminative features for FER using the two-stage training strategy.

Acknowledgment. This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs. (No: JCYJ20170306165153653) and National Engineering Laboratory for Video Technology - Shenzhen Division, Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). Special acknowledgements are given to Aoto-PKUSZ Joint Lab for its support.

References

1. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: facial expression editing with controllable expression intensity. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
3. 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
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. 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)
6. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC 2008–19th British Machine Vision Conference, pp. 275–1. British Machine Vision Association (2008)
7. 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)

8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
9. 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
10. 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)
11. 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)
12. Liu, X., Vijaya Kumar, B., You, J., Jia, P.: Adaptive deep metric learning for identity-aware facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–29 (2017)
13. 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, pp. 94–101. IEEE (2010)
14. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
15. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 443–449. ACM (2015)
16. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: Bmvc, vol. 1, p. 6 (2015)
17. Sikka, K., Dhall, A., Bartlett, M.: Exemplar hidden Markov models for classification of facial expressions in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 18–25 (2015)
18. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
19. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. **23**(10), 1499–1503 (2016)
20. 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)