Semi-supervised Learning of Deep Difference Features for Facial Expression Recognition

Can Xu, Ruyi Xu, Jingying Chen(✉), and Leyuan Liu

National Engineering Research Center for E-Learning,
Central China Normal University, Wuhan, China
chenjy@mail.ccnu.edu.cn

Abstract. Facial expression recognition (FER) is an important means of detecting human emotions and is widely applied in many fields, such as affective computing and human-computer interaction. Currently, several methods for FER heavily rely on large amounts of manually labeled data, which are costly and not available in real-world applications. To address this problem, this paper proposes a semi-supervised method based on the deep difference features. First, a cascaded structure is introduced to the original safe semi-supervised SVM (S4VM) to solve the multi-classification task. Then, multiple deep different features are fed to the cascaded S4VM to train the six basic facial expressions using the information of the unlabeled data safely. Extensive experiments show that the proposed method achieved encouraging results on public databases even when using a small labeled sample set.

Keywords: Facial expression recognition · Deep learning · Cascaded S4VM
Semi-supervised method

1 Introduction

Analyzing facial expressions is one of the most important methods of human emotion recognition and facial expressions are defined as the corresponding facial changes in response to a person's inner emotional state and intentions [1]. Nowadays, automatic facial expression recognition (FER) has miscellaneous applications, such as affective computing, interactive games, social psychology, synthetic animation, and intelligent robots [2].

Automatic FER systems can be divided into two categories: those that based on static images and those that based on dynamic image sequences [3]. The static-based method only contains information of the currently input image, while the sequence-based method can use temporal information from multi frames to identify the expression. FER systems receive static images or dynamic sequences as input and then output the corresponding expression category. This work focuses on methods based on the key frames extracted from dynamic image sequences.

In the past two decades, many attempts have been made to recognize facial expressions, and the effectiveness of these attempts depends largely on the size of the labeled training set. A large-scale training set can better reflect the real distribution of samples and hence acquire a better generalization error. However, manual annotation is

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 245–254, 2018.
https://doi.org/10.1007/978-3-030-03338-5_21

demanding, time consuming and expensive [4]. A semi-supervised method can simultaneously use labeled and unlabeled data to improve the classification performance with small datasets, reduce the workload of manual labeling and enhance the practicability of FER [5].

There have been few attempts to recognize facial expressions using a semi-supervised method. Existing methods can be roughly divided into two categories: semi-supervised learning (SSL) [6–8] and semi-supervised clustering [9–11]. SSL exploits the distribution of the unlabeled data to enhance training. Semi-supervised clustering sets the pairwise constraints with labeled data for cluster analysis. In 2004, Cohen et al. [6] were the first to apply SSL to facial expression recognition. They trained probabilistic classifiers with labeled and unlabeled data based on Bayesian networks and achieved an average recognition accuracy of 74.8% on the Cohn-Kanade dataset. Hady et al. [7] mentioned a learning framework to exploit the unlabeled data by the combination of the Co-Training and the one-against-one output-space decomposition approach, which uses Tri-Class SVMs as binary classifiers. The average recognition accuracy on the four basic expressions of the Cohn-Kanade dataset was 86.95%. Jiang et al. [8] focused on the problem of multi-pose facial expression recognition by bringing transfer learning into SSL. Liu et al. [9] addressed the expression recognition in the wild under a semi-supervised frame that combined reference manifold learning with Semi-Supervised Non-negative Matrix Factorization to select discriminant unlabeled data for enhanced training. Liliana et al. [10] proposed a semi-supervised clustering method based on Fuzzy C-means (FCM) to consider the level of ambiguity of facial expressions. Araujo et al. [11] mentioned a semi-supervised temporal clustering method and applied it to the complex problem of facial emotion categorization.

Although the unlabeled samples are helpful to construct the exact model for facial expression classification, experiments show that the effect of some SSL methods is even worse than simply using the methods employed for labeled samples [12, 13]. To address this problem, Li and Zhou presented the safe semi-supervised vector machine (S4VM) [14] to explore multiple candidate low-density separators, estimate the decision boundary closest to the real situation and ensure the best classification effect. The researchers define S4VM as a safe semi-supervised classifier whose performance never degenerates, even when using unlabeled data.

Inspired by Li and Zhou, this work proposes a semi-supervised learning method based on the DPND feature. The DPND feature proposed in our previous work [15] extract the deep representations of the peak (the fully expressive) frame and the neutral frame, respectively, and use the difference between them to represent the facial expression. In this paper, to further improve the robustness, a set of DPND features is extracted from each facial expression sequence which select the key frames near to the cluster centroids. Then, a cascaded semi-supervised classifier is constructed to classify facial expressions with both labeled and unlabeled samples. The final classification result of each sequence is decided by the voting of all key-frame pairs.

The rest of this paper is organized as follows. The details of the semi-supervised FER method are presented in Sect. 2. The experimental setup is described in detail, and the experiment results are given in Sect. 3. Section 4 concludes the paper.

2 The Proposed Method

In this section, the proposed semi-supervised FER approach will be described in detail. The proposed method consists of two main parts: (1) Multiple DPND feature extraction from expression sequences and (2) construction of a cascaded semi-supervised classifier for FER.

2.1 Multiple DPDN Feature Extraction

To address the FER problem, researchers have proposed many elaborate features to represent facial expressions during past decades [16]. However, some recent works show that features learned from millions of training samples by deep learning outperform manually designed features in face-related tasks, such as face detection [17] and face recognition [18]. Encouraged by these advancements, the popular VGG-16 [19] is adopted as the network architecture for deep representation extraction in this study. The VGG-16 is pre-trained on the VGG face dataset, which contains 2.6 M face images from 2,622 subjects. When face images are put into the VGG-16, the output of neuron responses by one of the intermediate layers of the VGG-16 network can be extracted as images' deep representation. In this paper, the DPND feature is employed to describe the change between the neutral frame and the peak frame as our previous work [15]:

$$f_{DPND} = \left(f^P - f^N\right)/N \tag{1}$$

where f^P and f^N are deep representation features extracted from the peak frame and neutral frame, respectively, and N is the normalized factor. The DPND feature can effectively retain facial expression information while eliminating individual differences and environmental noises.

For some standard facial expression datasets, such as CK+ [20], in which each sequence begins with the neutral expression and ends with the peak expression, the DPND feature can be easily obtained by the deep representation feature of the beginning frame and the end frame. However, the neutral frame and the peak frame of an expression sequence are not directly available in some datasets, such as the BU-4DFE [21]. To extract the DPND feature from expression sequences, a joint method of K-means clustering and rank-SVM is presented.

However, a single DPND feature [15] from each sequence to represent the facial expression has two limitations: first, the extraction of key frames has a certain randomness due to the random initialization of cluster centroids; second, the extracted key frames can only approximately represent the neutral frames and peak frames. In order to further improve the robustness, in this work, a set of DPND features is extracted from each facial expression sequence which select the key frames near to the cluster centroids obtained using K-means. The final classification result of each sequence is decided by the voting of all key-frame pairs. In this way, the multiple DPND feature can effectively avoid the problem caused by the inaccurate selection of key frames. And the subsequent experiments prove that, compared to the single DPND feature, the multiple DPND feature can indeed improve the accuracy of FER.

2.2 Construct a Cascaded Multi-class Classifier for FER

In this subsection, a cascaded classifier is introduced to the S4VM construct to recognize the six basic facial expressions using the proposed DPND feature. The original S4VM proposed by Li and Zhou [14] is an inductive binary classifier. For applying it to FER tasks, a set of S4VMs is combined with a cascaded structure, and each S4VM divides a kind of facial expression from the given dataset. A brief introduction of S4VM is first given.

Safe Semi-Supervised Support Vector Machine (S4VM). Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{\pm 1\}$ be the label space. A set of labeled data as $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{\boldsymbol{l}}$ and a set of unlabeled data are given as $\left\{\hat{x}_j\right\}_{j=1}^{u}$. Semi-Supervised learning SVM (S3VM) aims to find a decision function $\boldsymbol{f} : \mathcal{X} \rightarrow \{\pm 1\}$ and a label assignment on unlabeled instances $\mathbf{y} = \left\{\boldsymbol{y}_{l+1}, \ldots, \boldsymbol{y}_{l+u}\right\} \in \mathcal{B}$ such that the following objective function is minimized,

$$h(f, \hat{y}) = \frac{\| f \|_H}{2} + C_1 \sum_{i=1}^{l} l(y_i, f(x_i)) + C_2 \sum_{j=1}^{u} l(\hat{y}_j, f(\hat{x}_j)) \tag{2}$$

S4VM focuses on the safeness of SSL algorithms. Its main idea is to generate multiple low-density separators to approximate the ground truth decision boundary and maximize the improvement in performance of inductive SVMs for any candidate separator. To generate a pool of diverse separators $\{f_t\}_{t=1}^{T}$, the following function is minimized:

$$\underset{\{f, \hat{y}_t \in \beta\}_{t=1}^{T}}{min} \sum\nolimits_{t=1}^{T} h(f_t, \hat{y}_t) + M\Omega(\{\hat{y}_t\}_{t=1}^{T}), \tag{3}$$

where T is the number of separators, Ω is a penalty coefficient about the diversity of separators, and M is a large constant to ensure diversity. A variety of methods can be adopted to solve this optimization problem, such as global simulated annealing search and representative sampling.

To learn a label assignment **y** such that the performance against the inductive SVM, $\boldsymbol{y}^{svm}$, is improved, the worst-case improvement over inductive SVM is maximized and $\bar{\mathbf{y}}$ is denoted as the optimal solution:

$$\bar{\boldsymbol{y}} = \underset{y}{\operatorname{argmax}} \min_{\hat{y}} gain(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{y}^{svm}) - loss(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{y}^{svm}) \tag{4}$$

where $gain(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{y}^{svm})$ and $loss(\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{y}^{svm})$ are the gained and lost accuracies compared to the inductive SVM, respectively. It has been shown that the accuracy of $\bar{\mathbf{y}}$ is never worse than that of $\boldsymbol{y}^{\boldsymbol{svm}}$ and achieves the maximal performance improvement over that of $\boldsymbol{y}^{\boldsymbol{svm}}$ in the worst cases.

Multi-class Classification with the Cascaded S4VM. The original S4VM is typically designed for binary classification problems; thus, S4VM must be extended into a multi-class classifier for FER. The most common strategies are called one-against-one and one-against-all, however, S4VM, as an inductive method, cannot use one-against-one to construct a multi-class classification, while adoption of one-against-all is ineffective due to the same large training set for each binary classification.

This paper constructs multi-class classification based on a cascaded structure [22, 23], which can hold inductive and effective to unlabeled data. In detail, the training set that contains labeled and unlabeled data is put into the cascaded classifier, and samples of the specified class are picked out for each S4VM classifier. The identified unlabeled data and the corresponding labeled data are removed from the training set, while the remaining samples are passed to the next S4VM classifier.

It is worth noting that the performance of multi-class classifiers varies widely according to different cascaded order. To design a more effective cascaded classifier, the order of the S4VM classifiers is determined according to a discriminant measure of labeled data. The ratio of the inner-class distance and the inter-class distance is defined as the separable measure:

$$\mathrm{S}_p = \frac{D_{pp}}{\sum_{q \neq p} D_{pq}} \tag{5}$$

where $\mathrm{D}_{pq} = \frac{1}{|p||q|} \sum\limits_{i \in p, j \in q} d_{ij}$ is the average distance between any two samples in the class p and q. The class p is separated from the training set according to the ascending order of S_p. The corresponding classes are sorted to $p_1, p_2, \ldots, p_m$. Then, a classifier with a cascaded structure is constructed, such as that shown in Fig. 1.

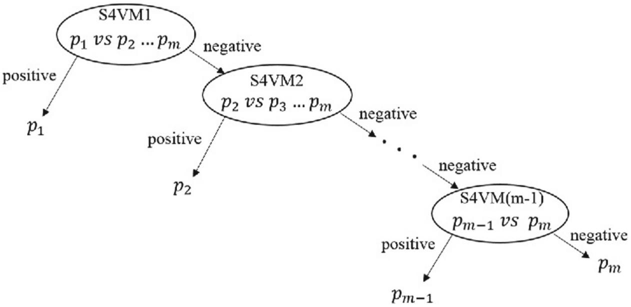

Fig. 1. Multi-class classification based on a cascaded structure.

Samples of class p_1 are assigned to the positive category, and samples of the rest classes are assigned to the negative category; then, the first sub-binary classifier S4VM1 is trained. After that, samples of class p_1 are removed from the training set. Similarly, samples of class p_2 are assigned to the positive category, and the rest of the samples are assigned to the negative category; then, the second sub-binary classifier SVM2 is trained until all the sub-classifiers are trained. Finally, a cascaded S4VM is obtained.

3 Experiments

3.1 Experimental Protocol

To evaluate the effectiveness of the proposed algorithm, two public sequence-based datasets, CK+ [20] and BU-4DFE [21], were chosen for the experiment; the CK+ dataset has been used in [10]. The details of these two datasets are listed in Table 1. In our experiment, only six basic expressions (angry, disgust, fear, happy, sad and surprise) were considered, and we extracted a subset of 53 subjects from the CK+ and a subset of 64 subjects from the BU-4DFE. Some samples of the two databases are shown in Fig. 2. For the CK+ dataset, the DPND feature is the difference between the deep representation feature of the first frame and the last frame; for BU-4DFE, the DPND feature is extracted from the facial sequences directly by our proposed method.

Table 1. Details of the CK+, BU-4DFE dataset.

Dataset	Subjects	Sequences	Gender(F/M)	Age	Ethnicity
CK+	97	486	65%/35%	18–30	Multiethnic
BU-4DFE	101	606	56%/44%	18–70	Multiethnic

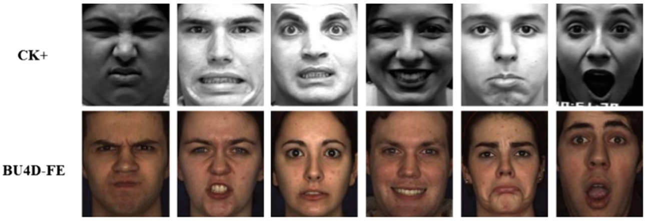

Fig. 2. Exemplar expression images in the CK+, BU-4DFE dataset.

3.2 Comparison Among the Multiple DPND, the Single DPND and the DPR Feature

In order to show the effectiveness of the DPND feature, we compared it with the static feature that the deep representations of peak frames (DPR feature) extracted from the VGG-16 network. Then, the proposed cascaded S4VM was employed to evaluate the effects of the different features. For BU-4DFE, the multiple DPND feature was extracted from a set of key-frame pairs near to the cluster centroids. It is noteworthy that the labeled samples only accounted for 10% of the training set in the experiment. The average accuracies of the different features are listed in Table 2. The results indicate that the accuracy of the single DPND feature on the CK+ and BU-4DFE are 8.5% and 21% higher than that of the DPR feature, and the performance of the multiple DPND feature is 3.4% higher than that of the single DPND feature on the BU-4DFE, which strongly proves the excellence of the DPND feature, especially the multiple DPND feature.

Table 2. Average accuracy of the DPND and DPR features.

Feature	CK+	BU-4DFE
Multiple DPND	——	71.8%
Single DPND	89.4%	68.4%
DPR	80.9%	47.4%

3.3 Comparisons with the State-of-the-Art Method

In this subsection, we compare the proposed method (the cascaded S4VM with the DPND feature) with the current state-of-the-art method [10] on the CK+ dataset. The method [10] is based on an SSL algorithm. It first employed an Active Appearance Model to detect human facial points for feature extraction and then utilized semi-supervised Fuzzy C-Means to work as the classifier system; we refer to the method as SSFCM. It selected 329 images of eight emotions from the CK+ dataset, of which 63% were used as a training set and the remaining samples were used for testing. The average accuracies of the proposed method and SSFCM method are shown in Table 3. The proposed method outperforms the SSFCM method [10] even though the SSFCM method selected the peak frames out from the sequences manually and used more labeled data than our method.

Table 3. Average accuracies of the proposed method and the current state-of-the-art method on the CK+.

Method	CK+
Proposed method	89.4%
SSFCM	80.7%

3.4 Comparison with the Supervised Classification

In this subsection, we aimed to use the CK+ and BU-4DFE dataset to evaluate the capability of the SSL method for FER. To this end, the proposed cascaded S4VM and SVM were used as expression classifiers and SVM was considered the baseline because it has been demonstrated as a successful approach for FER tasks. The performance of the cascaded S4VM was calculated based on its outputs, including the list of generated labels for unlabeled data. Using the same data, SVM was applied as a fully supervised version of the cascaded S4VM (see Table 4) for comparison of the semi-supervised learning and supervised learning. The results demonstrate that although a small proportion of each dataset was labelled (10%), the accuracy of the cascaded S4VM for FER on the CK+ and BU-4DFE are 5% and 12% higher than that of SVM.

For more evaluation, the accuracy of the cascaded S4VM was considered with different amounts of labeled data (10%, 12.5%, 17%, 20%, 25% and 50%), as shown in Fig. 3. In all these experiments, the cascaded S4VM achieved better accuracy than SVM, especially in the case of few labelled data, which confirms the cascaded S4VM's efficiency. The results illustrate that combined with information from labeled and unlabeled samples, the cascaded S4VM can predict the distribution of data more reasonably and then adjust the decision boundary to improve the classification accuracy. Figure 3 also shows that as the number of labeled data increases, the accuracy of the cascaded S4VM and SVM also increase and match.

Table 4. Accuracy of the cascaded S4VM compared to SVM.

Dataset (10%)	SVM	Cascaded S4VM
CK+	84.9%	89.4%
BU-4DFE	59.9%	71.8%

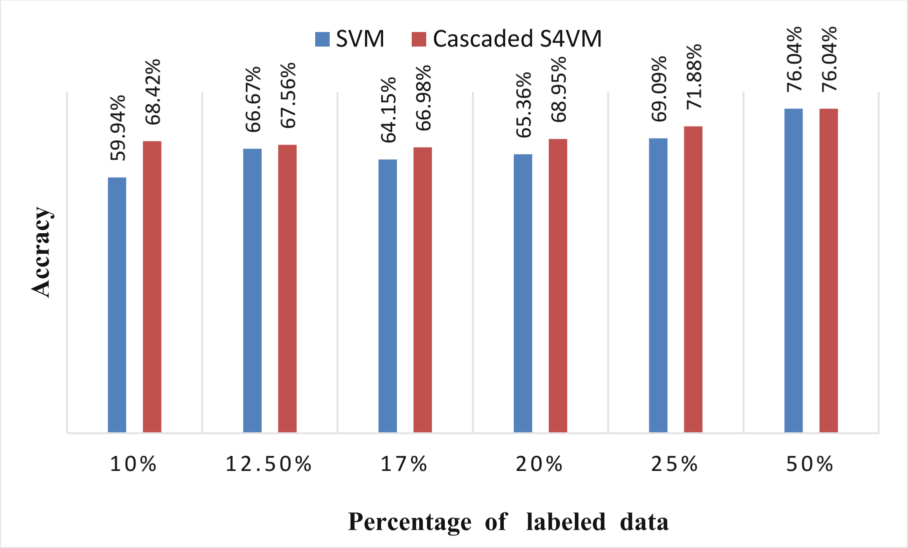

Fig. 3. Accuracy with different percentages of labelled data.

4 Conclusion

In this paper, we propose a semi-supervised method based on the multiple DPND feature for FER. The DPND feature tends to emphasize the facial parts that are changed in the transition from the neutral to the expressive face and to eliminate differences in individual face identities and environmental noises. In this work, the multiple DPND feature are extracted from each sequence to improve the robustness of feature representation. Then, a cascaded semi-supervised classifier is constructed to recognize six basic facial expressions using both labeled and unlabeled data. The proposed method achieves an accuracy of 89.4% on the CK+ dataset and an accuracy of 71.8% on the BU-4DFE dataset when only 10% of the training samples are labeled. The encouraging results on public databases suggests that our method has strong potential to recognize facial expressions in real-world applications.

Acknowledgment. This work was supported by Research funds from the National Key Research and Development Program of China (No. 2018YFB1004500, No. 2018YFB1004504), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU17ZDJC04) Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJAZH005), National Natural Science Foundation of China (No. 61702208), and Natural Science Foundation of Hubei Province (No. 2017CFB504).

References

1. Li, S.Z., Jain, A.K.: Handbook of Face Recognition, vol. 132, no. 3, pp. 470–487. Springer, Heidelberg (2011)
2. Fang, T., Zhao, X., Ocegueda, O., Shah, S.K.: 3D facial expression recognition: a perspective on promises and challenges. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, vol. 28, pp. 603–610. IEEE (2011)
3. Lopes, A.T., Aguiar, E.D., Souza, A.F.D., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2016)
4. Jiang, B., Jia, K., Sun, Z.: Research on the algorithm of semi-supervised robust facial expression recognition. In: Yoshida, T., Kou, G., Skowron, A., Cao, J., Hacid, H., Zhong, N. (eds.) AMT 2013. LNCS, vol. 8210, pp. 136–145. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-02750-0_14
5. Jadidi, Z., Muthukkumarasamy, V., Sithirasenan, E., Singh, K.: Flow-based anomaly detection using semi supervised learning. In: International Conference on Signal Processing and Communication Systems, pp. 1–5. IEEE (2016)
6. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semi supervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Trans. Pattern Anal. Mach. Intell. **26**(12), 1553–1566 (2004)
7. Hady, M.F.A., Schels, M., Schwenker, F., Palm, G.: Semi-supervised facial expressions annotation using co-training with fast probabilistic tri-class SVMs. In: Diamantaras, K., Duch, W., Iliadis, Lazaros S. (eds.) ICANN 2010. LNCS, vol. 6353, pp. 70–75. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15822-3_8
8. Jiang, B., Jia, K.: Semi-supervised facial expression recognition algorithm on the condition of multi-pose. J. Inf. Hiding Multimed. Sig. Process. **4**(3), 138–146 (2013)
9. Liu, M., Li, S., Shan, S., Chen, X.: Enhancing expression recognition in the wild with unlabeled reference data. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7725, pp. 577–588. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-15822-3_8
10. Liliana, D.Y., Widyanto, M.R., Basaruddin, T.: Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means. In: International Conference on Advanced Computer Science and Information Systems, pp. 439–445. IEEE (2017)
11. Araujo, R., Kamel, M.S.: A semi-supervised temporal clustering method for facial emotion analysis. In: IEEE International Conference on Multimedia and Expo Workshops, pp. 1–6. IEEE (2014)

12. Wang, L., Chan, K.L., Zhang, Z.: Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I-629–I-634. IEEE (2003)
13. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi-supervised support vector machines. J. Mach. Learn. Res. **9**(1), 203–233 (2008)
14. Li, Y.F., Zhou, Z.H.: Towards making unlabeled data never hurt. IEEE Trans. Pattern Anal. Mach. Intell. **37**(1), 175–188 (2015)
15. Chen, J., Xu, R., Liu, L.: Deep peak-neutral difference feature for facial expression recognition. Multimed. Tools Appl. **77**(2), 1–17 (2018)
16. Corneanu, C.A., Oliu, M., Cohn, J.F., Escalera, S.: Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1548–1568 (2016)
17. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: Computer Vision and Pattern Recognition, pp. 5325–5334. IEEE (2015)
18. Hu, G., Yang, Y., Yi, D., Kittler, J., Christmas, W., Li, S.Z., et al.: When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition, pp. 384–392 (2015)
19. Parkhi, O.M, Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 1–12 (2015)
20. Tian, Y.L., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97 (2001)
21. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: International Conference on Automatic Face and Gesture Recognition, vol. 2006, pp. 211–216. IEEE (2006)
22. Saatci, Y., Town, C.: Cascaded classification of gender and facial expression using active appearance models. In: International Conference on Automatic Face and Gesture Recognition, vol. 47, pp. 393–398. IEEE (2006)
23. Li, L., Gao, Z.P., Ding, W.Y.: Fuzzy multi-class support vector machine based on binary tree in network intrusion detection. In: International Conference on Electrical and Control Engineering, vol. 28, pp. 1043–1046. IEEE (2010)