Facial Expression Recognition Based on Region-Wise Attention and Geometry Difference

Heran Du[1,2,3], Huicheng Zheng[1,2,3(✉)], and Mingjing Yu[1,2,3]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
zhenghch@mail.sysu.edu.cn
[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China
[3] Guangdong Key Laboratory of Information Security Technology, 135 West Xingang Road, Guangzhou 510275, China

Abstract. Facial expression is usually considered as a face movement process. People can easily distinguish facial expressions via subtle facial changes. Inspired by this, we design two models that are expected to better recognize facial expressions by capturing subtle changes in the face. First, we consider to re-calibrate the response of different facial regions to highlight several special facial areas. According to this idea, we constructed cross-channel region-wise attention network (CCRAN), which can underline the important information and mine the correlations between different facial regions effectively. Moreover, we use the feature subtraction method to obtain geographical facial difference information. Based on this idea, we constructed temporal geometric frame difference network (TGFDN), which accepts the facial landmark points as input. These points are extracted from the facial expression frames. This network can effectively extract the slight changes of geographical information on the expression sequences. Through properly fusing these two networks, we have achieved competitive results on the CK+ and Oulu-CASIA databases.

Keywords: Facial expression recognition · Attention mechanisms
Temporal difference

1 Introduction

Facial expressions are part of the human body's language. It is a physical and psychological response commonly used to convey feelings. Therefore facial expression recognition (FER) in the human-computer interaction is very important. In order to conduct the interaction, the machine needs to recognize the human facial expression to perceive their feeling. Considering that the expression often contains rich emotional information, the application of this task is very extensive.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 183–194, 2018.
https://doi.org/10.1007/978-3-030-03338-5_16

FER is generally considered as a classification problem. Many people have done a lot of research in this field before. Overall, these studies can be divided into two categories: frame-based methods and sequence-based methods [1,7,15, 20,24,28]. Because facial expressions are generally considered as a movement process, extracting useful temporal and spatial features is very helpful for facial expression recognition. Therefore, the recognition methods based on the image sequence are generally considered to be superior to the methods based on a still single frame [7,15].

However, the above methods are mainly based on the entire human face. In facial expression recognition tasks, the major changes in expression are often concentrated in several subtle facial regions. Humans can accurately recognize the category of expression through several key areas of the face, such as forehead, mouth, and brow. Therefore, the weights in different areas of the feature maps should be different.

In this paper, we first propose cross-channel region-wise attention network (CCRAN), trying to find the relationship between the different regions of the feature map. We hope to improve the network's ability to express specific image regions by introducing the cross-channel region-wise squeeze and excitation (CCSE) branch. Through this branch, we expect to re-calibrate features and enhance the image regional sensitivity of the network without introducing additional information.

Furthermore, we also propose temporal geometric frame difference network (TGFDN) to extract the temporal features from the facial landmarks. This network can effectively capture facial morphological changes and accurately describe facial movement characteristics. By performing feature extraction and frame difference for the landmarks of each frame separately, the network can extract low-level facial expression movement information from the landmarks. The result of the landmark difference is concatenated along the time axis and then input into the subsequent layers to further extract the high-level expression features. At the end of that, we can obtain the geometry information and movement characteristics of facial expressions.

The main contributions of this paper are divided into three parts.

- We propose CCRAN model, which accepts continuous frames as input, enhancing the network ability to recognize facial expressions by adding cross-channel region-wise attention mechanisms to the network.
- We propose TGFDN model, which can extract the inter-frame difference information from the facial landmarks points and can describe the motion process of expressions accurately.
- Finally, we fuse these two networks. The integrated deep spatial-temporal network takes into account geometry-appearance, regional-global, intra-frame and inter-frame information synthetically, improving the accuracy of expression recognition effectively.

2 Related Work

2.1 FER Based on Traditional Methods

Before the large-scale use of the deep learning-based method, it is a common practice to use hand-crafted features for facial expression recognition. These methods can be further divided into three kinds of methods based on local features extraction, facial action units (FAUs), and spatio-temporal information, respectively. Traditional methods based on local features, such as HOG, SIFT, LBP, and BoW have been extended to video. These methods also have their 3D cases [11,15,23,25,31]. In FAU based methods [12,13], facial action coding system (FACS) is used to detect and analyze FAUs to classify facial expressions. The methods based on spatio-temporal information are represented by the work of Liu et al. [15]. They have proposed an expressionlet-based spatio-temporal manifold descriptor.

2.2 FER Based on Deep Methods

In recent years, deep convolutional neural networks have achieved great success in image classification [4,5,27], object detection and localization [3,16,21,22], semantic segmentation [3,17], and other computer vision fields. Corresponding to these tasks, in the field of facial expression recognition, Liu et al. propose 3DCNN-DAP [14], which is based on 3D-CNN, constructing a deformable parts learning component to capture the expression features. Further, Jung et al. [8] trained two small deep networks with facial landmarks and image sequences separately. To achieve the better result, they performed joint fine tuning method to fuse these two networks. Based on this structure, Zhang et al. [29] introduce recurrent neural network to further analyze the facial landmarks. Ding et al. [2] use a large pre-trained face recognition network to help train a simple facial expression recognition network through a regularization mechanism. Based on this, Ofodile et al. [19] further improved the accuracy by introducing the motion trajectory of the landmark points into the network. In addition, Kim et al. [10] attempted to use a small deep encoder-decoder network pre-trained on a face database to obtain a contrastive representation between expression face and neutral face, which helps to distinguish expressions.

3 Approach

In summary, the proposed method uses a combination of two simple networks. First, we construct the TGFDN to capture the geographical inter-frame motion information. Then we use CCRAN to extract local appearance information in consecutive frames of the expression. Finally, these two networks are properly combined to improve the performance of facial expression recognition.

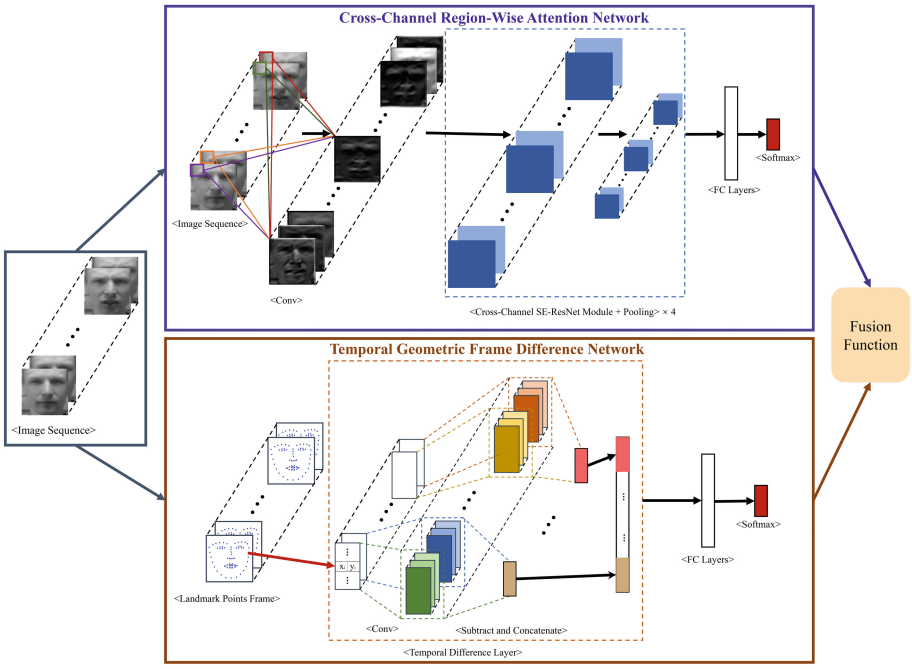

Fig. 1. Overview of our proposed architecture. The upper part of the figure shows the structure of CCRAN. The image sequence is fed into the network directly. Using a simple bottleneck (a convolution layer, a ReLU activation layer, and a batch normalization layer), the channels are increased to 64. After that, four cross-channel region-wise attention (CCRA) blocks are interleaved with four pooling layers and then followed by a fully connected layer to get logits. The lower part of the figure shows the structure of TGFDN. Facial landmark points are extracted from the frame sequence, reshaped into a matrix in which each row stores the coordinates of a point. Then the landmark matrices are fed into convolution layers separately. After the feature subtraction and difference concatenation, a fully connected layer is used to obtain logit values.

3.1 Cross-Channel Region-Wise Attention Network

In recent years, adding short connections to the network has proven to be an effective way to increase the efficiency of network information propagation [4,6,26]. So we use a simple CNN-Resnet structure as our backbone, which receives t frames of expression as input. The network includes four residual blocks interleaved with four pooling layers, and a fully connected layer at the end. Each residual block contains two convolutional layers. A batch normalization layer and a ReLU activation layer are between them, as shown in Fig. 3(a).

The whole Resnet block shown in Fig. 3(a) can be regarded as a unit that does not change the size and channels. The main problem with the backbone is that the convolutional operation takes equal considerations for the entire feature map and are less sensitive to subtle local changes. So we have joined the

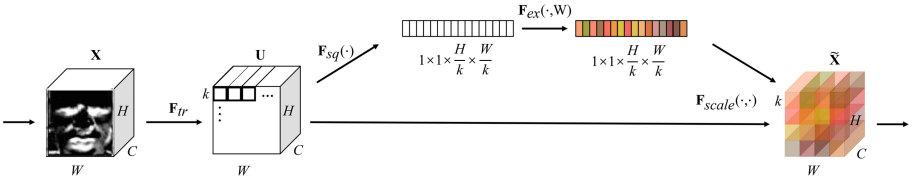

Fig. 2. Overview of the cross channel squeeze and excitation process.

cross-channel region-wise attention branch on the basis Resent block of this network. This branch draws on the squeeze and excitation network [5] and can be trained end-to-end, including a cross-channel squeeze and a cross-channel excitation operation as shown in Fig. 2.

The purpose of the squeeze operation is to compress the information of all feature maps within a layer into a one-dimensional vector. Specifically, we first compress all feature maps into a single feature map using average pooling. Then we use a $k \times k$ filter to do average pooling again on this entire compressed feature map. Each region of the compressed feature map is compressed to one value. We then flatten these values into a one-dimensional vector. The vector obtained in this way takes into account the context between the channels and the facial regions. Formally, a two-dimensional matrix $z \in \mathbb{R}^{\frac{H}{k} \times \frac{W}{k}}$ is generated by squeezing U through cross-channel $k \times k \times C$ sized average pooling window, where the z_{ij} is calculated by:

$$z_{ij} = F_{sq}(U) = \frac{1}{k \times k \times C} \sum_{c=1}^{C} \sum_{h=i \cdot k}^{i \cdot k+k-1} \sum_{w=j \cdot k}^{j \cdot k+k-1} u_c(i,j) \tag{1}$$

We further extract the contextual relationships between the regions contained in the vector through the excitation operation. Like SE-net [5], in order to reduce the complexity of the model while reducing over-fitting, we use two fully-connected layers as a bottleneck. One layer is the dimension-reduction layer, and the other is the dimension-restoring layer. Between these two layers, we use a ReLU as the activation layer to get more nonlinearity, so as to better fit and mine the complex correlations between different regions. We will use this branch to integrate with the original Resnet block. As we have shown in Fig. 3.

We obtain CCRAN by using the block in Fig. 3(b) to replace the block in Fig. 3(a). It can be seen from Fig. 3(b) that the cross-channel SE branch we proposed can be added flexibly to the original network structure. Here, we join the cross-channel SE branch before the identity addition operation.

3.2 Temporal Geometric Frame Difference Network

The entire network includes a temporal difference layer and two fully connected layers as shown in the upper part of Fig. 1. The TGFDN network receives the sequence of facial landmarks as input. We select t-frame facial landmarks to

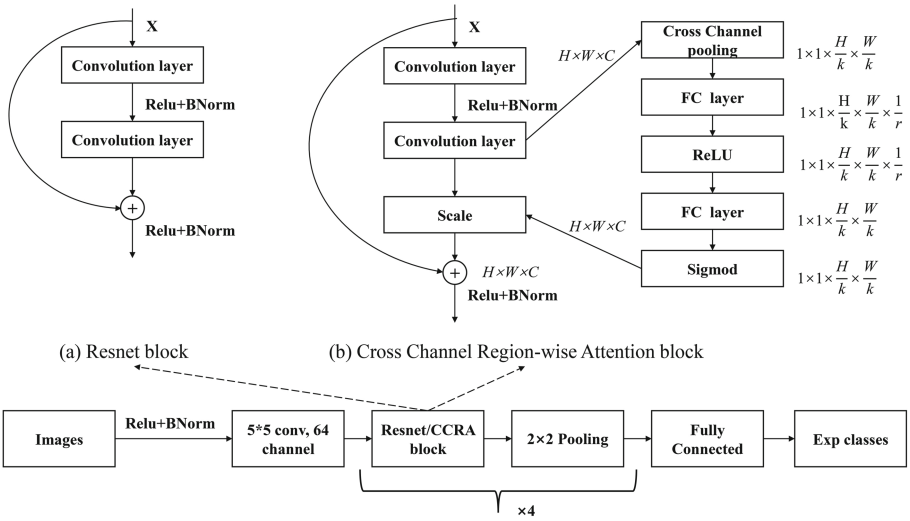

Fig. 3. Overview of the CCRAN architecture: (a) resnet block, (b) cross channel region-wise attention block, (c) the backbone network with two kinds of block.

describe the expression features. In Fig. 1, the landmarks selected for each frame are arranged in a matrix where each row stores the xy-coordinates of a point. Then t matrices are stacked and input into the network at the same time.

In the temporal difference layer, we use a convolutional operation to extract features frame-by-frame. The kernel size is $n \times 1$. Let $X = [x_1, x_2, \ldots, x_t]$ denote the input facial landmarks, where x_t refers to the landmark points extracted from the t-th facial expression frame. The set $U = [u_1, u_2, \ldots, u_t]$ represents a set of convolution kernels and $V = [v_1, v_2, \ldots, v_t]$ denotes the features extracted via convolution operation. Features v_t are extracted from x_t using its corresponding convolution kernels u_t,

$$v_t^s = u_t^s * x_t \tag{2}$$

where $*$ denotes convolution, while v_t^s denotes the s-th feature map of v_t and u_t^s represents the s-th kernel of u_t. The convolution operation is followed by a batch normalization layer and a ReLU activation layer. Then, we use the feature obtained in this frame minus the features obtained in the previous frame to obtain frame difference. After that, we concatenate all the differences and flatten them into the one-dimensional vector. Formally, Z represents the concatenation output, and C is the concatenation operation. Here we have:

$$Z = C(u_2 - u_1, u_3 - u_2, \ldots, u_t - u_{t-1}) \tag{3}$$

Then, the difference layers are passed through the two fully connected layers and finally classified using softmax function. The discussion on convolution kernel size and the hyper-parameter t is detailed in Sect. 4.4.

3.3 Model Fusion

We fuse the two networks together through a fusion function referring to the fusion method of Zhang et al. [29].

$$O(x) = \sum_{i=0}^{1} a_i(\beta A_i(x) + P_i(x)) \tag{4}$$

$P_i(x)$ $(0 < P_i(x) < 1)$ is the output of the softmax layer in the CCRAN and TGFDN. $P_0(x)$ comes from CCRAN and $P_1(x)$ comes from TGFDN. $A_i(x)$ is sorted according to the predicted value of each expression in $P_i(x)$. In addition, β $(0 \leq \beta \leq 1)$ acts as a weight parameter. When the value of β is close to 1, the fusion function will give priority to the sorting result of different expressions. When the value of β is close to 0, the fusion function will be a simple weighted-sum function. Finally, a_i is the balance factor between different models. We empirically set a_i to 0.5 and β to 0.1. This function considers the sorting results of the softmax output and actual value of the softmax output simultaneously.

4 Experiments

We evaluated the performance of our model on two widely used databases, including CK+ [18] and Oulu-CASIA [30]. The process and details of the experiments are shown in this section.

4.1 Implementation Details

The structure of CCRAN is I64-[B(5,64)+P2] $\times$ 4-FC1024-S7. I64 means that the size of input frames is 64×64, and B(5,64) refers to a cross-channel SE block with 64 channels and filters of size 5×5. Moreover, P2 refers to a 2×2 max pooling layer and FC1024 means a fully connected layer with 1024 nodes. The structure of TGFDN is L(68,2)-C((1,3),16)-FD-FC600-S7. L(68,2) means that landmarks of a frame are reshaped to 68×2 for input, and C((1,3),16) means a convolution operation with 16 output channels and filters of size 1×3. Moreover, FD means a frame subtraction layer and FC600 means a fully connected layer with 600 nodes. At last S7 is the softmax layer with seven outputs (in CK+ database).

4.2 Databases and Protocols

The CK+ Database. The CK+ database [18] is a representative database of facial expression recognition tasks. This database has a total of 539 sequences of facial expressions, corresponding to 123 subjects with different ages and genders. Among them, 327 expression sequences are marked and correspond to seven types: anger, contempt, disgust, fear, happiness, sadness, and surprise. Each expression sequence begins with a plain frame (neutral expression) and ends with the peak frame of expression. We follow the usual protocol of using 10-fold cross validation [8,15] for testing.

The Oulu-CASIA VIS Database. There are 80 individuals in this database. Each individual has six expressions, including anger, disgust, fear, happiness, sadness, and surprise. So the database has a total of 480 expression sequences. Like the CK+ database, we use 10-fold cross validation as our experimental method.

4.3 Data Preprocessing and Augmentation

The duration of the expression is not the same, but our network needs to accept a fixed-length image sequence as input. Therefore, we use the average sampling method to regularize the expression sequence along the time axis. From these sampled frames, the faces are detected, cropped and reshaped into 64×64. What's more, we use dlib [9] to further extract 68 facial landmarks. Then we regularize all the facial landmark points using the method described in [8]. We also follow the method of Jung et al [8], making data augmentation to the training data to alleviate the overfitting problem.

4.4 Experiment Results

Comparison with Other Methods. On the CK+, we can see that our method is very close to state-of-the-art [29] and better than three pre-trained models. The method with $*$ in Table 1 indicates that these methods use the face recognition database for pre-training and the facial expression database for fine-tuning, which introduces additional information to improve the result. On the Oulu-CASIA database, our method has also achieved very good results. The recognition ability of the fused network is higher than VGG-16 pre-trained network. Moreover, the recognition result obtained by CCRAN, which only uses the image frame as input, is surprisingly higher than the DTAGN, which uses both image frames and landmark points as input for recognition on the Oulu-CASIA database. It should be noted that there is no contradiction between our approach and the state-of-the-art [29]. It is very likely to further improve the performance by simply integrating the CCRA mechanism and the frame difference mechanism into the network to form a complementary relationship with our method.

Analysis and Discussion

Region-Wise Squeeze-and-Excitation Blocks. As we can see in Table 2, by adding the cross-channel region-wise attention (CCRA) mechanism to the Resnet block, the network performs better on two databases. This result shows that recalibration of the different region in feature maps can effectively help the network to learn facial expression features.

Facial Landmark Selection. The coordinates of facial landmarks extracted using the dlib [9] can only be integers, which are not accurate and can cause noise in the result. If the sampling frequency of expression frames is too high, the noise

Table 1. Comparisons of different methods on the CK+ and Oulu-CASIA database (where * indicates that the model use face recognition database for pre-training).

Method	Accuracy(CK+)	Accuracy(Oulu)
3DCNN [14]	85.9%	-
3DCNN-DAP [14]	92.4%	-
DTAN [8]	91.44%	74.38%
DTGN [8]	92.35%	74.17%
DTAGN(Weighted Sum) [8]	96.94%	80.62%
DTAGN(Joint) [8]	97.25%	81.46%
PHRNN-MSCNN [29]	98.50%	86.25%
VGG-16 Fine-Tune* [2]	89.9%	83.26%
FN2EN* [2]	96.8%	87.71%
GCNet* [10]	97.93%	86.39%
CCRAN	95.48%	81.58%
TGFDN	94.55%	77.38%
CCRAN-TGFDN	98.11%	83.54%

Table 2. Comparisons between resnet block and cross-channel region-wise attention block on the CK+ and Oulu-CASIA database.

Method	Explanation	Accuracy(CK+)	Accuracy(Oulu)
Baseline	Resnet block	94.39%	79.91%
CCRAN	CCRA block	95.48%	81.58%

Table 3. Comparisons between different input number and filter size of TGFDN on the CK+ and Oulu-CASIA database.

Input Size	Filter size	Accuracy(CK+)	Accuracy(Oulu)
7-frames	1×3	93.68%	74.12%
3-frames	2×2	92.99%	75.54%
3-frames	1×1	93.61%	77.13%
3-frames	1×3	94.55%	77.38%

will be large after frame difference operation. As shown in Table 3, we can see that using landmarks with only three frames ($t = 3$) for recognition has achieved better result than that with 7 frames. In addition, we also tried different filter sizes in the network. Through the display in Table 3, we can see that the results using 2×2 size filters on CK+ and Oulu-CASIA are significantly lower than the other two convolution kernels. We think the reason is that the correlation between the x-coordinate and the y-coordinate of the face landmark points is relatively small. So a single-column-size filter performs better.

Table 4. Confusion matrix of CK+ database.

	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	**97.78**	0	1.69	0	0	0	0
Contempt	2.22	**94.44**	0	0	0	3.57	0
Disgust	0	0	**98.31**	0	0	0	0
Fear	0	0	0	**92**	0	0	0
Happy	0	0	0	4	**100**	0	0
Sadness	0	5.56	0	4	0	**96.43**	0
Surprise	0	0	0	0	0	0	**100**

Table 5. Confusion matrix of Oulu-CASIA database.

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**78.75**	20	0	0	8.75	0
Disgust	12.50	**70**	12.5	0	2.5	0
Fear	0	0	**80**	2.5	2.5	8.75
Happy	1.25	0	6.25	**97.5**	1.25	0
Sadness	7.50	8.75	6.25	0	**85**	1.25
Surprise	0	1.25	6.25	0	0	**90**

Confusion Matrix. Tables 4 and 5 show the confusion matrices for our algorithm on the CK+ and Oulu-CASIA databases, respectively. The abscissa of the table represents prediction results and the ordinate represents labels. We can see that in the CK+ and Oulu-CASIA databases, the performance of our model for the fear is relatively poor, but the performance for happy and surprise is good.

5 Conclusion

In this paper, we try to improve the accuracy of expression recognition by capturing subtle facial movements. We propose CCRAN to extract the continuous, region-based, spatial appearance expression information and construct TGFDN to obtain temporal, global-based geographic expression features. After we fused these two networks, our model achieved better results on two different databases. In addition, other popular network structure may also explore the relationship between different areas of the feature map by simply adding the cross-channel region-wise attention mechanism. Therefore, our method is novel, effective, and general.

Acknowledgments. This work was supported by National Natural Science Foundation of China (U1611461), Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase, No. U1501501), and Science and Technology Program of Guangzhou (No. 201803030029).

References

1. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: development and applications to human computer interaction. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop, vol. 5, pp. 53–53. IEEE (2003)
2. Ding, H., Zhou, S.K., Chellappa, R.: FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 118–126. IEEE (2017)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
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. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)
6. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269 (2017)
7. Jeni, L.A., Lőrincz, A., Szabó, Z., Cohn, J.F., Kanade, T.: Spatio-temporal event classification using time-series kernel based structured sparsity. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 135–150. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_10
8. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE International Conference on Computer Vision, pp. 2983–2991. IEEE (2015)
9. Kazemi, V., Josephine, S.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874. IEEE (2014)
10. Kim, Y., Yoo, B., Kwak, Y., Choi, C., Kim, J.: Deep generative-contrastive networks for facial expression recognition. arXiv preprint arXiv:1703.07140 (2017)
11. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference, p. 275-1. British Machine Vision Association (2008)
12. Liu, M., Li, S., Shan, S., Chen, X.: AU-aware deep networks for facial expression recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6. IEEE (2013)
13. Liu, M., Li, S., Shan, S., Chen, X.: AU-inspired deep networks for facial expression feature learning. Neurocomputing **159**, 126–136 (2015)
14. 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
15. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756. IEEE (2014)
16. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

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. 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: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101. IEEE (2010)
19. Ofodile, I., et al.: Automatic recognition of deceptive facial expressions of emotion. arXiv preprint arXiv:1707.04061 (2017)
20. Pantic, M., Rothkrantz, L.J.: Facial action recognition for facial expression analysis from static face images. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **34**(3), 1449–1461 (2004)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM International Conference on Multimedia, pp. 357–360. ACM (2007)
24. Shan, C., Gong, S., McOwan, P.W.: Conditional mutual infomation based boosting for facial expression recognition. In: British Machine Vision Conference (2005)
25. Sikka, K., Wu, T., Susskind, J., Bartlett, M.: Exploring bag of words architectures in the facial expression domain. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7584, pp. 250–259. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33868-7_25
26. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)
27. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2015)
28. 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)
29. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
30. 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)
31. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)