Signal, Image and Video Processing (2021) 15:263–270
https://doi.org/10.1007/s11760-020-01753-w

ORIGINAL PAPER

LKRNet: a dual-branch network based on local key regions for facial expression recognition

Dandan Zhu[1,2] · Gangyi Tian[1,2] · Liping Zhu[1,2] · Wenjie Wang[1,2] · Bingyao Wang[1,2] · Chengyang Li[1,2]

Received: 18 January 2020 / Revised: 21 May 2020 / Accepted: 19 July 2020 / Published online: 28 July 2020

Abstract
The task of facial expression recognition (FER) is riddled with many challenges, such as face occlusion, head posture, illumination angle, and intensity. Due to the development of deep learning and large FER datasets in recent years, most methods have achieved notable success. This paper aims to solve the problem that general classification models are difficult to distinguish, for some easily confused expressions (such as anger and surprise). To this end, we make two contributions in this paper: (1) The model extracts weighted local key regions as local information on the final feature maps, and fuses the global information for multi-task recognition. (2) Triplet loss function is used to make the intra-class feature distance significantly reduced from the inter-class feature distance. It can enhance the discriminability of features while fitting the sample distribution. The experiments confirm that two contributions are combined to gain another round of performance boost. For instance, the results on CK+ and FER2013 datasets demonstrate the superiority of the proposed method.

Keywords Facial expression recognition · Convolutional neural networks · Local key region · Triplet loss

1 Introduction

At present, many FER datasets face many challenges such as facial occlusion, different head postures, different illumination angle, and individual differences. Besides, the boundaries between some expressions are vague and difficult to separate. Similar expression categories are often confused, especially in datasets collected in an uncontrolled environment (natural scenes), such as FER2013. What is more, different personal identity attributes produce huge noise, like human age and gender.

Researches show that facial expression recognition is closely related to the features of face local key regions. The facial features such as eyes and mouth are effective for FER. In related face recognition tasks, some methods combine facial landmark detection [19,33] to predict several facial attributes. Thus, key local information in the face image can well assist the extraction of relative features. However, many studies [6,10,27] only train the relevant tasks together, or simply introduce the face key features. These methods do not analyze contributions of key local features of different positions for face recognition.

Although the traditional cross-entropy loss can effectively fit the distribution of sample space, it is not accurate enough for the hard samples that are difficult to classify. For these hard samples, the inter-class distance is not large enough and the intra-class distance is not small enough. To deal with it, the triplet network uses triplet loss function. Based on the distance of samples in the feature space, triplet loss aims to make the inter-class distance far greater than the intra-class distance. It uses metric learning rather than label learning to extract more effective features, which provides a new idea for classification tasks. Research [7] has shown that the improvement of classic networks by the triplet network is beneficial to classification tasks.

For the above-mentioned reasons, this paper proposes a facial expression recognition method based on local key regions, named LKRNet. The total architecture is shown in Fig 1. Contributions of this paper are summarized as follows:

✉ Gangyi Tian
2019211265@student.cup.edu.cn

Dandan Zhu
zhu.dd@cup.edu.cn

1 College of Information Science and Engineering, China University of Petroleum (Beijing), Beijing, China

2 Key Lab of Petroleum Data Mining, China University of Petroleum (Beijing), Beijing, China

(1) First, we design a dual-branch architecture based on local key regions. The positions of facial landmarks from the input face image are obtained through MTCNN. The features of the weighted regions near the landmarks are extracted from the last feature maps based on the distance to the center landmark. One branch is to classify only the features of local key regions, and the other is to classify the whole feature maps. The two branches process simultaneously in the network, which together promotes performance.
(2) Second, inspired by FaceNet [20], we use triplet loss to improve traditional cross-entropy loss function. The purpose is to make same expressions closer and different expressions farther. The triplet loss uses the relative constraint rather than the absolute constraint on the distribution of features. Meanwhile, cross-entropy loss is consistent with the sample distribution of the hypothesis space, which is better for the stability of network training. By combining triplet loss and cross-entropy loss, the classification accuracy is further improved.

The rest of this paper is organized as follows: Sect. 2 provides an overview of FER. Section 3 introduces the proposed method. Section 4 shows the results and analysis of experiments. Section 5 makes a conclusion.

2 Related work

Previous studies have shown that many traditional feature extraction methods can be applied to facial expression recognition, such as optical flow [4,15], Gabor wavelet transform [24,29], LBP [13], HOG [18], etc. Besides, the methods also include hidden Markov model [17,34], artificial neural network [1], Bayesian network [21], support vector machine, Adaboost [5], etc. These traditional feature extractors have achieved certain results on datasets. However, these extractors cannot automatically extract features. Especially in large datasets collected in an uncontrolled environment, they are susceptible to many factors. Therefore, these feature extraction methods are not applicable in complex datasets.

FER2013 [3] and emotion recognition in the wild [2] indicates that FER is developed from laboratory environment to reality since 2013. Convolutional neural network (CNN) is widely used in image classification tasks, including FER. The current research is mainly divided into two directions: (1) increase input information, to provide more prior knowledge to the network; (2) optimize network structure.

Using only RGB images as input may lose important information such as texture, rotation, translation, scaling, occlusion, and illumination. Therefore, some methods extend the input of networks to solve this problem. Levi [11] combines the original image with its mapping LBP features based on 3D metric space as the input of CNN. This can remove some illumination noise from the input image. Luo [14] uses edge, texture, and angle features to generate three different feature maps as the input of CNN. Zhang [32] uses the SIFT feature vectors to compose feature matrix as input. By training deep learning network, they can obtain the mapping relationship of the SIFT feature vector and its corresponding semantic information.

Other methods enhance expression feature extraction by optimizing the network structure design. Yao [28] propose HoloNet, using concatenated rectified linear units (CReLU) to reduce redundant filters and enhance nonlinearity in the lower convolutional layer. By combining the residual block with CReLU to construct the middle layers, it can increase the depth of the network without causing gradient disappearance or gradient explosion. Hu [8] proposes the supervised scoring system (SSE), which embeds three supervised network blocks into the whole network. The inter-class scores of the three layers are combined as the input of the secondary supervision to output the final result.

3 Proposed method

3.1 Local key regions

The global characteristics of the entire image tend to get a lot of non-critical information. This will lead to the lack of generalization ability. Local key regions can make a good use of effective features for FER. We intend to obtain local features that are closely related to facial expression changes, such as eyes, nose, and mouth.

According to the excellent performance of MTCNN [31], it is used to extract the bounding box and facial landmarks from the face image. The facial area outside the critical range of the landmarks contributes little to the expression recognition, and may even introduce noise. Therefore, these regions are not treated as key regions. The size of local areas is computed, as shown in Eqs. 1 and 2:

$$w_{\text{local}} = \alpha w_{\text{global}} \tag{1}$$

$$h_{\text{local}} = \alpha h_{\text{global}} \tag{2}$$

Here, w_{global} and h_{global} are the width and height of the face bounding box detected by MTCNN. w_{local} and h_{local} are the width and height of the assumed local areas. We find the extraction effect is relatively better when $\alpha = 0.3$.

MTCNN also outputs the coordinates of 5 facial landmarks (left eye, right eye, nose, left mouth corner, and right mouth corner) $(x_c, y_c, (c = 1, 2, 3, 4, 5))$. Each key region from the upper left coordinate $(x_c - \frac{w_{\text{local}}}{2}, y_c - \frac{h_{\text{local}}}{2})$ to the lower right coordinate $(x_c + \frac{w_{\text{local}}}{2}, y_c + \frac{h_{\text{local}}}{2})$ is used to

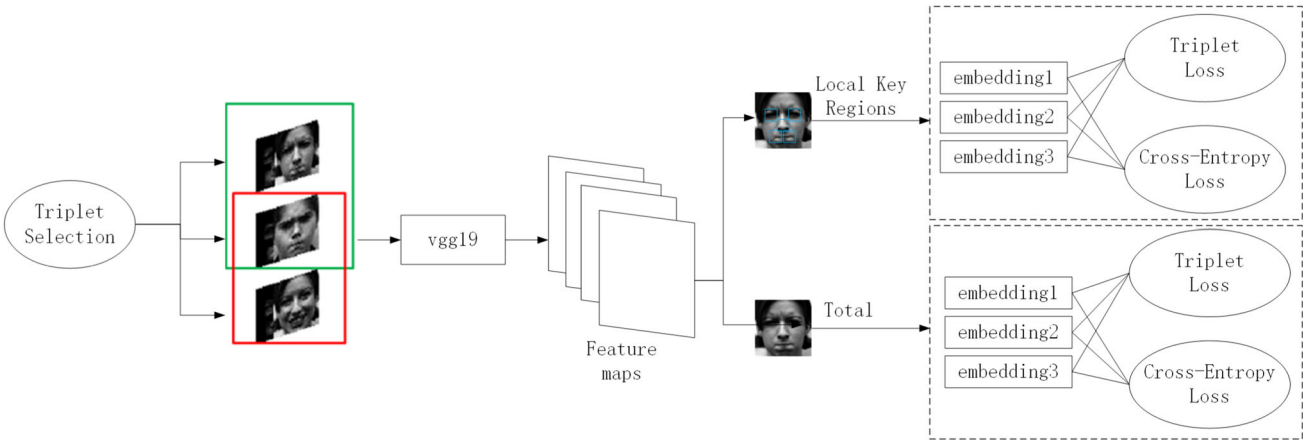

Fig. 1 The total architecture of our method

generate mask image from the original expression image. In order to get more representative features of the regions close to the facial landmarks, a simple weight-normalized linear Euclidean distance to the center facial landmark is used for each pixel (i, j) in each local key region, as shown in Eq. 3:

$$w_{ij} = 1 - \frac{d_E((i, j), (x_c, y_c))}{d_E((x_c^{'}, y_c^{'}), (x_c, y_c))} \tag{3}$$

Here, d_E is the Euclidean distance. $d_E((i, j), (x_c, y_c)) = \sqrt{(i - x_c)^2 + (j - y_c)^2}$ indicates the Euclidean distance from any pixel to its center landmark in a local region. $d_E((x_c^{'}, y_c^{'}), (x_c, y_c)) = \sqrt{(x_c^{'} - x_c)^2 + (y_c^{'} - y_c)^2}$ indicates the Euclidean distance from the farthest pixel (the four corners) to its center landmark in the local region. The farther the pixel is from the center point, the smaller the corresponding weight is, and vice versa. Specially, the weight of the farthest pixel (the four corners) of a region is 0, and the weight of the center is 1.

Given the final feature maps p, the weighted mask $M(p)$ is generated. Then, downsampled $M(p)$ is multiplied by p to obtain a weighted feature maps of local key regions. It is shown in Eq. 4:

$$p^{'} = p \bigodot M(p) \downarrow \tag{4}$$

Figure 2 shows examples of seven expressions, including anger, contempt, disgust, fear, happiness, sadness, and surprise from top to bottom, respectively. p and $p^{'}$ form dual branches. Then, predictions are made through the same structure but not shared classification networks, respectively. The use of the extracted key region feature maps $p^{'}$ can introduce the key local information of face to assist the network feature extraction. In the deep convolutional neural network, embedding the position information of pixels relative to the landmarks can effectively improve the recognition accuracy.

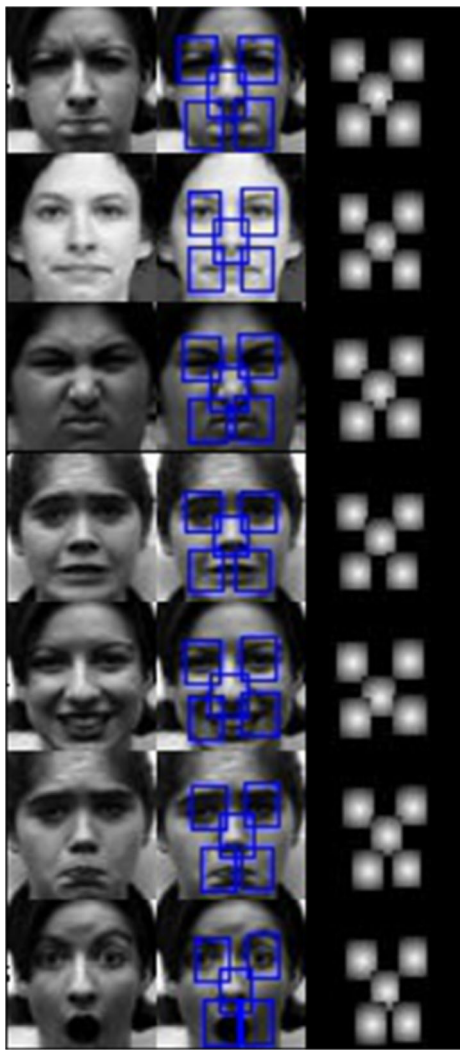

Fig. 2 Local key regions of seven expressions

Traditional expression recognition only uses the cross-entropy loss to supervise the training process. The cross-entropy loss helps maintain the discriminability of deep features between different classes. However, there are still serious differences within each class. Moreover, facial expressions in natural scenes have significant intra-class differences. In order to obtain better expression classification performance, the extracted features should be characterized by a large distance between different classes and a small distance within same class.

Due to the excellent performance of the triplet loss, this paper proposes a novel loss function that introduces the triplet loss into the traditional cross-entropy loss. The triplet network consists of three samples, anchor sample a, positive sample p, and negative sample n, which are put into the same feedforward network with shared parameters.

The proposed loss function combines the weighted value of both the cross-entropy loss and the triplet loss. The hyper-parameter λ adjusts the weight of the two loss functions, as shown in Eq. 5:

$$L = L_{\text{cross}} + \lambda L_{\text{triplet}} \tag{5}$$

Here, L_{cross} is the sum of the cross-entropy loss of the anchor sample, the positive sample and the negative sample in a triplet. Its implementation is in Eq. 6:

$$\begin{aligned} L_{\text{cross}} = &\sum_{j}^{K} \log(p_i^a(j)) q_i^a(j) + \sum_{j}^{K} \log(p_i^p(j)) q_i^p(j) \\ &+ \sum_{j}^{K} \log(p_i^n(j)) q_i^n(j) \end{aligned} \tag{6}$$

Here, K represents the number of expression categories; p is the probability of each expression class; y is the ground truth label of the expression image; q is the indicator function (Eq. 7).

$$q(j) = \begin{cases} 1, j = y \\ 0, j \neq y \end{cases} \tag{7}$$

Triplet loss is set so that the distance between the anchor sample and the positive is significantly smaller than the distance between the anchor sample and the negative. It is shown in Eq. 8:

$$\left|\left| f(x_i^a) - f(x_i^p) \right|\right|_2^2 + \alpha < \left|\left| f(x_i^a) - f(x_i^n) \right|\right|_2^2, \tag{8}$$

where x_i^a represents the anchor example; x_i^p represents the positive sample; x_i^n represents the negative sample; α represents the margin of the distance between the positive pair and the negative pair. Therefore, the loss function L_{triplet} is shown in Eq. 9:

$$L_{\text{triplet}} = \left[\left|\left| f(x_i^a) - f(x_i^p) \right|\right|_2^2 - \left|\left| f(x_i^a) - f(x_i^n) \right|\right|_2^2 + \alpha \right]_+ \tag{9}$$

In Eq. (9), $[z]_+ = \max[z, 0]$. This loss function allows the anchor sample to be closer to the positive example and farther from the negative example in the feature space.

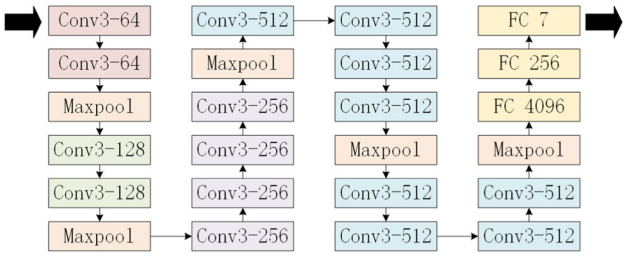

Fig. 3 The modified structure based on VGG19

3.2 Network architecture

The model structure is reasonably adjusted based on the classical deep CNN, VGG19 [23]. The final fully connected layers are removed, and two fully connected layers are added to reduce layer dimension. Finally, a seven-dimensional fully connected layer is added to classify the expression into 7 expression categories. The modified vgg19 is shown in Fig. 3.

In order to obtain the detailed analysis of the proposed architecture's performance, the following three configurations are evaluated. Therefore, we can demonstrate the effects of different modules in the proposed method.

Baseline1. Only use local key regions in Sect. 3.1. The model is trained only by the traditional cross-entropy loss function.

Baseline2. Do not use local key regions. The proposed weighted loss function is calculated instead of only using the cross-entropy loss function.

4 Experiments

4.1 Datasets

In order to evaluate the recognition effect of the proposed model, we select two typical expression datasets CK+ and FER2013. The two datasets represent different scene, laboratory environment and natural environment. Samples of the datasets are shown in Fig. 4. The left is examples of the CK+ dataset, and the right is examples of the FER2013 dataset.

Extended CohnKanade (CK+) is based on the Cohn-Kanade Dataset. The images are collected in a laboratory

Fig. 4 CK+ dataset and FER2013 dataset

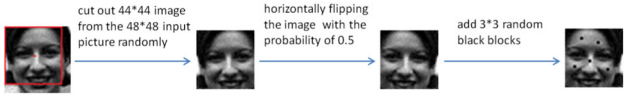

Fig. 5 The process of facial image preprocessing

Table 1 Hyperparameter values of the model

Hyperparameter	Value
Learning rate	0.001
Learning rate decay	0.8
Momentum	0.9
Weight decay	0.0005
Batch size	128

environment. This dataset includes seven basic expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. It contains 593 video sequences from 123 subjects. Among it, 327 video sequences from 118 subjects have expression category labels.

FER2013 is a large, natural dataset collected automatically by the Google Image Search API. All images are 48×48 pixels. This dataset includes seven expression categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality. It contains a total of 28,709 training images, 3589 verification images, and 3589 testing images.

4.2 Implementation details

The model is based on the deep learning framework PyTorch. In order to avoid over-fitting, data augmentation is used on the images, as shown in Fig. 5. We resize the input image to 48×48, randomly select a cutting center, and cut out the image of 44×44 in size and then horizontally flip the image with a probability of 0.5 and add 3×3 random black blocks in the image. This can make the training dataset more abundant, obtaining stronger generalization ability of the model. In addition, we set the learning rate to 0.0001 to warm up the model (20 epochs), and we also used dropout strategy, which is set to 0.8.

The stochastic gradient descent (SGD) is used as the optimizer in the training process of the model. Other main hyperparameter values are shown in Table 1.

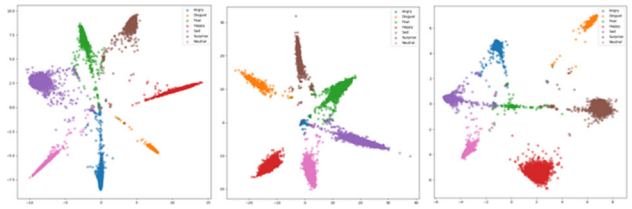

Fig. 6 Feature vector visualization of the proposed model. The value of λ from left to right is set to 0, 0.2, and 0.4, respectively

4.3 Results presentation

4.3.1 Parameter validation

In the proposed loss function, the hyperparameter λ is a key parameter to balance the cross-entropy loss and the triplet loss. When λ decreases, the cross-entropy loss has a greater impact. In particular, when $\lambda = 0$, the model reduces to **baseline1**.

On the FER2013 testing images, the feature vectors the model extracted under different λ values are visualized as shown in Fig. 6. As λ increases, the inter-class distance of samples in the feature space increases.

- When λ is small, the cross-entropy loss plays a leading role. The distribution of samples between classes in the feature space is scattered. For example, when $\lambda = 0$, some categories such as anger (blue samples), fear (green samples), sadness (purple samples), and neutral (pink samples) have some samples which cannot be separated from other categories.
- As λ increases, the distribution of different classes in the sample space becomes more and more dispersed, so that the classes can be better distinguished. For example, when $\lambda = 0.2$, different classes are well separated, and the intra-class sample distribution is aggregated.
- When λ is too large, the cross-entropy loss has a small impact. Not only the distribution of samples in the same class is relatively concentrated, but also the feature vectors of some classes are close. In this way, the classification effect is degraded. For example, when $\lambda = 0.4$, the distribution of a certain samples such as anger (blue samples), fear (green samples), sadness (purple samples) is relatively viscous.

4.3.2 Hard sample classification

From the experiment results, the proposed model has a higher improvement degree of accuracy on FER2013 than CK+. A possible explanation is that CK+ is a dataset collected in a controlled laboratory environment with less noise information (occlusion, head posture, etc.). Using a general deep

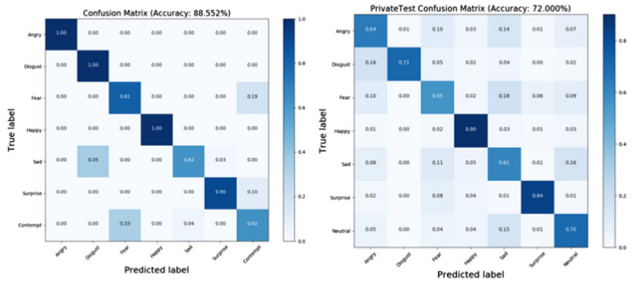

Fig. 7 Confusion matrix of the model on CK+ and FER2013 datasets

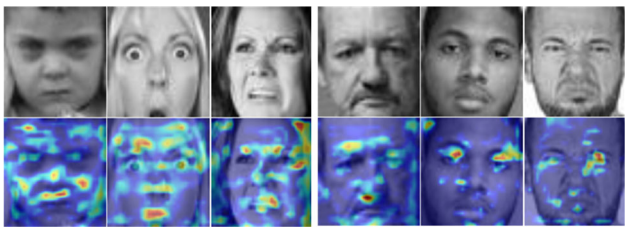

Fig. 8 Feature extraction visualization of VGG19 and our method

CNN can already obtain good results. For the dataset collected in the natural environment such as FER2013, there are more interferences in the image. The boundary between the expressions is more blurred. This makes the expression recognition more difficult. Using local key regions can filter noise to some extent. The loss function aggregates samples of the same class and separate samples of different classes in the feature space to some extent.

On both CK+ and FER2013, the accuracy improvement of the proposed method on some difficult-to-classify expressions (such as "anger," "fear," "surprise") is higher than that of other easily classified expressions (such as "disgust" and "neutral"). In addition, through the confusion matrix (Fig. 7), some expressions that are easily confused with each other (such as "happy" and "surprised") are also better recognized. The proposed method only mistakenly recognizes "contempt" as "fear," "sadness" as "disgust" on CK+ in a small range, and "'fear" as "sadness" on FER2013 in a small range. The other expressions can be almost predicted correctly.

4.3.3 Feature visualization

In order to visualize the effectiveness of the proposed model, the feature visualization model highlights the important regions of the image. We use the Grad-CAM algorithm [22] to obtain the heatmap of the proposed model on images. The heatmap indicates the intensity of features extracted from the model in an image. In this way, the key regions of expression recognition can be analyzed.

Figure 8 shows the feature visualization results on some samples of FER2013 dataset. The left image shows the feature extraction effect of VGG19, and the right image shows that of our method. It can be observed that VGG19 extracts features that are not closely related to FER (such as hair and background). In contrast, the features extracted by our method are mainly the facial features that are highly relevant to expressions. The result clarifies how the proposed method enhances the interpretability of the network. It also verifies the rationality and effectiveness of the proposed method.

4.3.4 Ablation study and comparison

We evaluate the two baselines and our proposed method on CK+ and FER2013. In addition, we compare with some state-of-the-art methods in recent years, such as Zeng et al. [30], Lpoes et al. [12], DeRL [26], Xiang et al. [25], Hua et al. [9], and Minaee et al. [16]. The recognition accuracy results are shown in Tables 2 and 3.

Both baseline 1 and baseline 2 are better than VGG19 in classification. The accuracy of baseline 1 and the accuracy of baseline 2 are higher than that of VGG19 on CK+ by 0.336% and 1.010%, respectively. There are also consistent conclusions on FER2013. The accuracy of baseline 1 and baseline 2 is higher than that of VGG19 on FER2013 by 0.250% and 1.616%, respectively. This shows that not only the use of key local regions can guide expression recognition, but also the fusion of cross-entropy loss and triplet loss can improve the result of expression recognition.

By integrating the advantages of baseline 1 and baseline 2, the proposed method has a recognition accuracy that exceeds two baselines. In the expression recognition, the accuracy of the proposed method on CK+ is 2.678% higher than that of VGG19. The accuracy is higher than that of baseline 1 and baseline 2 by 2.521% and 2.124%, respectively. Besides, the accuracy of the proposed method on FER2013 is 3.132% higher than that of VGG19. The accuracy is higher than that of baseline 1 and baseline 2 by 2.844% and 1.894%, respectively. The accuracy of different models iterated 60 epochs on the CK+ dataset is shown in Fig. 9.

Compared with state-of-the-art methods, our proposed method has improved the accuracy of "anger" and "disgust," but "sadness," "surprise," and "contempt" are a bit weaker than other methods on CK+. Meanwhile, on FER2013, our method has a significant improvement in "disgust," which improves the accuracy by 6% at least. This proves the effectiveness of our proposed method.

5 Conclusions

This paper makes two contributions to tackling FER problem. First, we propose a dual-branch architecture based on local key regions for learning the facial key information. Despite the fact that it is accurate and effective, it is yet to be

Table 2 Comparison of different models on CK+ dataset

CK+	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)	Contempt (%)
VGG19	95	99	71	88	87	86	78
Baseline 1	100	99	71	100	72	91	64
Baseline 2	69	99	76	100	100	88	78
Zeng et al. [30]	90	99	87	100	87	98	94
Lopes et al. [12]	91	99	92	100	82	98	–
DeRL [26]	96	96	90	99	96	99	100
Our method	100	100	83	100	87	91	81

Table 3 Comparison of different models on FER2013 dataset

FER2013	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)	Neutral (%)
VGG19	60	75	53	85	60	81	68
Baseline 1	60	73	52	88	59	81	68
Baseline 2	64	75	54	88	58	82	70
Xiang et al. [25]	60	41	33	81	52	74	54
Minaee et al. [16]	53	66	46	69	63	68	80
Hua et al. [9]	66	69	48	88	65	79	74
Our method	65	75	55	90	61	84	70

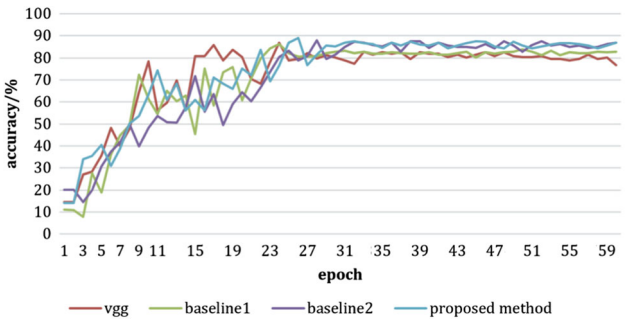

Fig. 9 Accuracy of different models iterated 60 epochs on CK+ dataset

improved. We introduce the triplet loss into the FER process. This can increase the distance between classes and decreases the distance within each class. The proposed method is evaluated on CK+ and FER2013 datasets. The experiment results show that this method is superior to the general CNN and significantly improves the accuracy of FER.

Future work will consider the effective extraction of time information on FER video datasets. FER based on 3D local regions will be studied as well.

References

1. Cohn, Y.l.T.J.: Recognizing facial actions by combining geometric features and regional appearance patterns. Carnegie Mellon University, the Robotics Institute (2001)
2. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426. ACM (2015)
3. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: a report on three machine learning contests. In: International Conference on Neural Information Processing, pp. 117–124. Springer (2013)
4. Guojiang, W., Guoliang, Y., Kechang, F.: Facial expression recognition based on extended optical flow constraint. In: 2010 International Conference on Intelligent Computation Technology and Automation, vol. 2, pp. 297–300. IEEE (2010)
5. Ai, H.Z., Xiao, X., Xu, G.: Face detection and retrieval. Chin J Comput. Chin. Ed. **26**(7), 874–881 (2003)
6. Hasani, B., Mahoor, M.H.: Facial expression recognition using enhanced deep 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 30–40 (2017)
7. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, pp. 84–92. Springer (2015)
8. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.: Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 553–560. ACM (2017)
9. Hua, W., Dai, F., Huang, L., Xiong, J., Gui, G.: Hero: human emotions recognition for realizing intelligent internet of things. IEEE Access **7**, 24321–24332 (2019)
10. Huang, R., Xie, X., Feng, Z., Lai, J.: Face recognition by landmark pooling-based CNN with concentrate loss. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1582–1586. IEEE (2017)
11. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Proceedings

of the 2015 ACM on International Conference on Multimodal Interaction, pp. 503–510. ACM (2015)
12. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit. **61**, 610–628 (2017)
13. Luo, Y., Wu, C.M., Zhang, Y.: Facial expression recognition based on fusion feature of PCA and IBP with SVM. Opt.-Int. J. Light Electron Opt. **124**(17), 2767–2770 (2013)
14. Luo, Z., Chen, J., Takiguchi, T., Ariki, Y.: Facial expression recognition with deep age. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (2017)
15. Mase, K.: Recognition of facial expression from optical flow. IEICE Trans. Inf. Syst. **74**(10), 3474–3483 (1991)
16. Minaee, S., Abdolrashidi, A.: Deep-emotion: facial expression recognition using attentional convolutional network. arXiv:1902.01019 (2019)
17. Otsuka, T., Ohya, J.: Spotting segments displaying facial expression from image sequences using HMM. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 442–447. IEEE (1998)
18. Ouyang, Y., Sang, N., Huang, R.: Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. Neurocomputing **149**, 71–78 (2015)
19. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 121–135 (2017)
20. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015)
21. Sebe, N., Lew, M.S., Cohen, I., Garg, A., Huang, T.S.: Emotion recognition using a cauchy naive bayes classifier. In: Object recognition supported by user interaction for service robots, vol. 1, pp. 17–20. IEEE (2002)
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
24. Wu, T., Fu, S., Yang, G.: Survey of the facial expression recognition research. In: International Conference on Brain Inspired Cognitive Systems, pp. 392–402. Springer (2012)
25. Xiang, J., Zhu, G.: Joint face detection and facial expression recognition with mtcnn. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 424–427. IEEE (2017)
26. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2168–2177 (2018)
27. Yang, H., Yin, L.: CNN based 3d facial expression recognition using masking and landmark features. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 556–560. IEEE (2017)
28. Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y.: Holonet: towards robust emotion recognition in the wild. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 472–478. ACM (2016)
29. Ye, J., Zhan, Y., Song, S.: Facial expression features extraction based on gabor wavelet transformation. In: 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), vol. 3, pp. 2215–2219. IEEE (2004)
30. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)
31. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
32. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K.: A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans. Multimed. **18**(12), 2528–2536 (2016)
33. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision, pp. 94–108. Springer (2014)
34. Zhou, X., Huang, X., Xu, B., Wang, Y.: Real-time facial expression recognition based on boosted embedded hidden Markov model. In: Third International Conference on Image and Graphics (ICIG'04), pp. 290–293. IEEE (2004)