Journal of Computers Vol. 33 No. 1, February 2022, pp. 13-19
doi: 10.53106/199115992022023301002

Multichannel Fusion Based on modified CNN for Image Emotion Recognition

Juntao Zhao*

Henan Institute of Economics and Trade, Zhengzhou 450000, China
sarkozyteague@foxmail.com

Received 27 September 2021; Revised 15 October 2021; Accepted 27 October 2021

Abstract. Social media networks are an integral part of people's daily life. Users share images and texts to express their emotions and opinions. Analyzing the image and text content published by these users can help understand and predict user behavior, so as to carry out marketing, public opinion monitoring and personalized recommendation. Weibo, Wechat and other social media are important ways of self-expression. Images are more intuitive than text. Therefore, more scholars begin to pay attention to the research of image emotion analysis. At present, image emotion analysis methods pay seldom attention to the influence of saliency object and face on image emotion expression. Therefore, we propose a multichannel fusion method based on modified CNN for image emotion recognition. Firstly, saliency target and face target region are detected in the whole image. Then feature pyramid is used to improve CNN to recognize saliency target emotion. Weighted loss CNN emotion recognition is constructed on multi-layer supervision module. Finally, the saliency target emotion, face target emotion and the directly recognized emotion on the whole image are fused to get the final result of emotion classification. Experimental results show that the proposed method can improve the accuracy of image emotion recognition.

Keywords: multichannel fusion, CNN, image emotion recognition, feature pyramid, saliency target emotion

1 Introduction

In the Internet era, images have become an effective way for people to express their emotions in daily life. Image emotion recognition is a complex and challenging problem, which can be applied to public opinion analysis [1, 2], robot emotion modeling, and network data analysis. The contents of the images are complex and rich, and the themes can be divided into landscapes, animals, humans and painting artworks, etc. Different images can intuitively display human emotions. Traditional image feature extraction methods usually require artificial setting of image features to be extracted, such as color features [3] and texture features [4]. Although such features can accurately express the features of the image, the operation of manual feature screening is very tedious, and multiple features cannot be taken into consideration at one time. Therefore, convolutional neural network (CNN) can be used to automatically extract multiple features of the image, and then continuously weighted and combined. Image classification [5] and target recognition [6] are carried out by using features most suitable for the current task. In this way, the tedious operation of artificial feature setting and screening can be avoided, and the features that are more consistent with the emotion recognition task can be selected to describe the emotion of the image. At the same time, features of different levels can be considered comprehensively.

Although CNN has been proved to be able to automatically recognize the target of the image, there are still several problems in the research of image emotion recognition. First of all, lack of labeled image emotion data will cause inadequate training of CNN model or lead to over-fitting problem in the trained network model. For example, the IAPSa image emotion dataset contains only 395 image data. Twitter image emotion data set [7] contains only 596 image data. The amount of data in these data sets is difficult to train an effective CNN model. Secondly, different images need different features to describe their emotions. The content of images is rich and varied. Different processing methods are needed to describe the emotion of images with different contents. For example, images with human facial expressions [8] need to focus on the impact of facial expressions on image emotions, while artistic paintings pay more attention to basic features such as color and texture. In literature [9], only basic features such as color and texture were used to analyze the emotion of images. The lack of multi-level feature description of image emotion is a key problem in image emotion recognition. At the same time, the relationship between multi-level features is not easy to obtain, and there is a huge gap between image features and human emotions, which also restricts the effect of image emotion recognition.

The contents of the two images are different, but because they exist in the same scene, the emotions expressed in the two images are similar. Sometimes, the two images have the same theme, but because of the difference in expression, the two images show opposite emotions. Image emotion is not only closely related to the target of the

* Corresponding Author

image, but also related to some basic features of the image. The object of the image can be separated from the whole image, so that we can focus on analyzing the effect of each part on the emotion of the image. Generally, CNN is used to extract high-level semantic information of foreground image and low-level basic features of background image respectively. To solve the above problem, we propose a multichannel fusion method based on modified CNN for image emotion recognition.

2 Related Works

Traditional image emotion analysis methods [10] mostly extract the underlying visual features of images such as color, texture, shape and outline for emotion analysis. Greenberg et al. [11] first used Plutchik's emotion wheel to construct a large visual emotion ontology, which was a semantic concept composed of more than 3000 emotive adjectives and nouns, and then selected 1200 well-performing adjective noun pairs as emotion detectors to train the emotion bank. Compared with the low-level visual feature method, the prediction using emotion bank has better accuracy, which can solve the semantic gap to a certain extent. In recent years, since deep neural networks can automatically extract image features, many researchers turn to adopt CNN to extract image emotion features. You et al. [12] first trained CNN by using Flickr data sets. Then it predicted the emotional polarity of images according to the trained CNN model, and filtered out the images that were not easily distinguished by CNN. Finally, a more discriminative emotional model PCNN (Progressive CNN) was trained using the newly obtained data. Experimental results show that this model was better than the simple CNN model. Jindal et al. [13] pre-trained model parameters in a large number of ImageNet data sets and used transfer learning to load model parameters into the image emotion recognition framework to reduce the training time.

With the deepening of image emotion analysis research, Song et al. [14] added the visual self-attention mechanism module into the CNN emotion classification framework for the first time. This method took the image saliency region as a prior knowledge to guide the learning of visual attention, thus making up for the deficiency of self-attention mechanism in learning emotion features. To mine the emotional areas of the image, Yang et al. [15] first adopted existing target detection methods to obtain N candidate areas, and combined the emotion scores and object points to select the top K emotional areas. Then it used the CNN to extract the features of the whole image and emotion areas. Finally, through the fusion strategy, it got the image emotional tendencies. Wu et al. [1] first used the target detection system to locate the saliency target region in the image, and then used VGGNet to train the saliency target and the entire image emotion recognition model respectively, and finally obtained the final emotion polarity by integrating the predicted results of the two. Experiments showed that saliency target could improve the accuracy of image emotion analysis. Zheng et al. [16] further explored the influence of the emotional region of saliency target on image emotion. Experiments showed that in saliency objects with faces, the emotion of images is often expressed through faces. By analyzing the existing image emotion analysis methods, it could be found that saliency target and face target play an important role in image emotion analysis.

However, there are few neural network models designed for saliency target and face target emotion characteristics, and the influence of saliency target and face emotion on the whole image emotion analysis are less. To solve the above problems, an image emotion analysis method based on multichannel fusion is proposed.

3 Proposed Image Emotion Recognition

In order to obtain the emotional polarity of saliency target emotion, the saliency target detection algorithm is firstly used to detect the position of saliency target, then the saliency target emotion characteristics are extracted through the constructed saliency target emotion recognition model. Finally, the emotional polarity of saliency target is obtained. Saliency target detection aims to detect the saliency target region in an image, which is applied in the fields of image understanding, image description generation and semantic segmentation. In this paper, SS-HED algorithm [17] is used to extract saliency target features at multiple scales of each layer by adding shortcut connections on the basis of depth supervised layer hopping structure. Specifically, first of all, it discards the last pooling layer and full connection layer of VGG16, then it conducts deconvolution after each convolutional block to get the side feature graph with the same size as the input image, and adds two convolution layers on the basis of each side feature graph to get the side output layer with the same size as the input image. The deep lateral output layer does a good job of locating the saliency target area, but it also loses some details. The shallow side output layer focuses on the low-level features, but lacks the overall information. Therefore, feature maps of the side output layer at different depths are fused to extract the saliency targets. Fig. 1 shows the specific process for obtaining the saliency target area. Firstly, a SS-HED saliency target detection model is trained by using

MSRA-B, and the image emotion data set is input into SS-HED to obtain the grayscale image of the saliency target region. Considering that grayscale images of multiple target areas may be connected, the grayscale images can be corroded to separate multiple target areas. Then it filters out target areas that are too small or have too wide a ratio. Since the emotion of an image is not only in the target area, the background around the target area also plays a key role in image emotion analysis. Therefore, in the grayscale map, the pixel of the minimum rectangular region surrounding the target is set as 255, and the remaining pixels are set as 0, as shown in Fig. 2(b). It adjusts the resulting binary graph to the same size as the original graph, and then performs bit-wise and operation on the original graph and binary graph. Finally, the contour detection is carried out, and the area belonging to the target rectangular frame is cut out. The mask operation process is as follows:

It is assumed that the pixel value a of the original Fig. 2(a) is RGB three-channel. Fig. 2(b) is a binary graph, in which the pixel value is $b \in (0,255)$. The target area in Fig. 2(c) containing saliency is obtained by bit-wise and operation of a and b, and its pixel value is c:

$$c = \begin{cases} (0,0,0) \; if \; b = 0 \\ a \qquad if \; b = 255 \end{cases} . \tag{1}$$

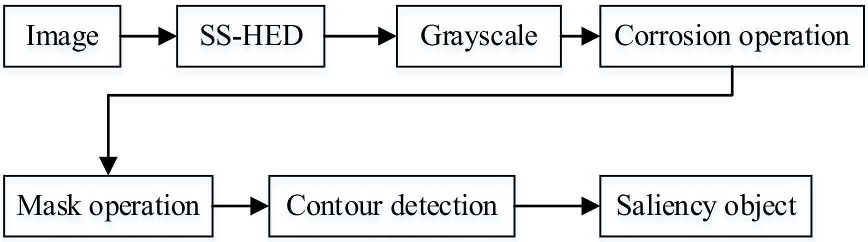

Fig. 1. Extracting saliency object region process

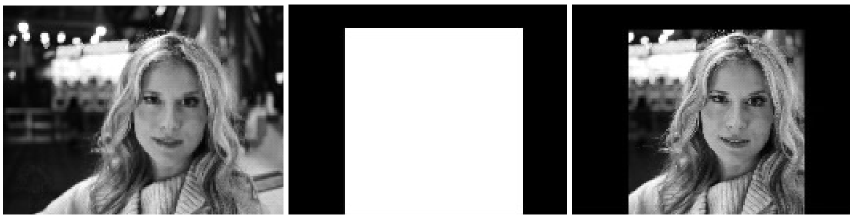
(a) raw image (b) binary image (c) saliency object image

Fig. 2. Raw image, binary image and saliency object results

Traditional VGG16 is a vertical network structure, and only deep convolutional layer can extract deep semantic features. In this paper, after the fifth convolution block of VGG16, the feature pyramid network (FPN) proposed by Lin et al. [18] was used to extract multiple deep scale feature images, and the multi-scale feature images were supervised and integrated to construct the saliency target emotion recognition model SFPN-CNN. In the first bottom-up stage, VGG16 is used as the base network structure, and the output of Conv2_2, Conv3_3, Conv4_3 and Conv5_3 is denoted as {C2,C3,C4,C5}. High-resolution shallow feature images in CNN have rich color, texture, shape and other features, while deep feature images are rich in semantic information. In the second top-down phase, FPN can combine not only low resolution and high level semantic features, but also high resolution and low level semantic features, in order to utilize different hierarchical feature maps at the same time. Fig. 3 is a schematic diagram of feature pyramid (FPN). C5 is convolved with 1×1 to obtain P5, and then a feature map of the same size as C4 is obtained through up-samping. Finally, the feature map of C4 convolved with 1×1 and the feature map of P5 up-sampled are added. In order to avoid aliasing, P4 is then obtained through 3×3 convolution, and iterated continuously. Finally, the multi-scale high-level semantic feature mapping {P5,P4,P3} of SFPN-CNN model is obtained.

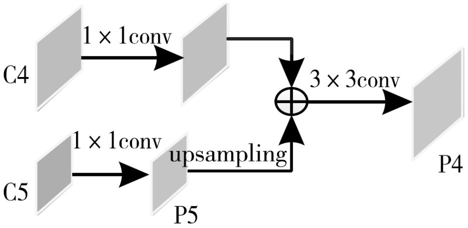

Fig. 3. FPN structure

The image target can be separated from the whole image through the detection of the saliency region, and then the image repair operation can be used to complete the vacant part of the image according to the characteristics and information of the background part [19-21]. This paper considers using foreground image to provide high-level semantic information and background image to provide low-level basic features. Zhang et al. [22] proposed a relational learning network, believing that humans could analyze two images at the same time and learned descriptive enhancement features for a certain category to analyze and compare the two images. The whole network consists of two branches, and the blocks of each branch are CNN networks with three convolutional layers. The two branches have their loss functions. The two branches generate feature images of foreground image and background image respectively, and they are combined and fused as input to the backbone network to learn the features of fused images and find out the differences and similarities between images. In the relational learning network, the two branches share parameters to ensure that the feature graph generated by the two branches is fixed size. The two branches have their own interest value, which indicates that humans pay more attention to the information of which branch in image analysis. The feature fusion of the two branches can be written as:

$$Conv = \lambda Conv_1 + (1-\lambda)Conv_2 \ . \tag{2}$$

Where, λ is the interest value of the branch in the foreground image. $(1-\lambda)$ is the interest value of the background image. $Conv_1$ represents the feature map generated by the branch of the foreground image. $Conv_2$ represents the feature map generated by the branch of the background image. Conv represents the feature graph after fusion. The loss function of the whole relational learning network is defined as:

$$l = \lambda l_{cls}(p, y^{(1)}) + (1-\lambda) l_{cls}(p, y^{(2)}) \ . \tag{3}$$

Where, p is the predicted probability after training. $y^{(1)}$ and $y^{(2)}$ are the image emotion markers of two branches respectively. l_{cls} is the cross loss function. Finally, the fused feature graph is input to the backbone network. Blocks in the backbone network are composed of CNN networks with three convolutional layers, which learn and fuse features in the feature graph for emotion recognition.

The two branches of the relational learning network use CNN network with the same number of layers to ensure the same size of feature maps generated by the two branches. If different CNN networks are used in the blocks of the two branches, the branch with fewer convolutional layers will generate larger feature images, while the branch with more convolutional layers will generate smaller feature images [23]. In order to fuse the feature maps of two branches, one way is to make the feature maps of small size up-sample to match the feature maps of large size, which will insert false values into the feature maps, it is equivalent to introducing noise to the feature maps. Another way is to pool the large size feature image, which reflects the low-level basic features of the image. After the pooling operation, a lot of information will be lost, and it is very likely that the feature image will directly lose the basic feature information such as color and texture. Through the front background image separation operation and the relationship learning network, it can effectively learn the similarity and difference between image features, as well as the relationship between different levels of image features, and recognize the emotion of the image by learning enhanced features.

4 Experiment and Analysis

4.1 Data Sets

In order to verify the effectiveness of image emotion analysis based on fusion of multiple visual objects, this paper uses three public image emotion datasets, including Twitter, ArtPhoto and Flickr and Instagram (FI) [24]. The three data sets are randomly sampled 80% as training set and 20% as testing set. ArtPhoto collects 806 art photos from photo-sharing sites, and each image is tagged with a real user. Twitter 1 collects 1269 images from the social networking site labeled by Amazon Mechanical Turk (AMT). FI data set firstly collects images on social networking sites through eight categories of emotion words. Then, 225 AMT workers further filter 90000 collected images, and finally selects 23308 images, including happiness, admiration, satisfaction, excitement, anger, disgust, fear and sadness. For the convenience of research, the first four categories are classified as positive and the last four as negative. Table 1 shows the number of categories for each dataset.

Table 1. Image emotion data

Data set	Positive	Negative	Total number
ArtPhoto	378	428	806
Twitter 1	769	500	1269
FI	16430	6878	23308

4.2 Experimental Design and Result Analysis

The experimental development environment is GeForce GTX 1060, 8GB video memory, TensorFlow 1.8.0, Python 3.6.4, and PyCharm. W+S+F+Max means that face emotion recognition model+saliency target emotion recognition model+whole image emotion recognition model+maximum fusion strategy to recognize image emotion.

To verify the performance of W+S+F+Max, it is compared with the following algorithms. FCNN is to recognize image emotion directly through the whole image. AR+concatenation [7] identifies image emotion by combining K emotion regions with the whole image. GMEI [10] refers to the recognition of image emotion through the saliency target model (VGG16) of the whole image fusion.

It can be seen from Table 2 that the accuracy of experiment 2 and experiment 3 is higher than that of Experiment 1, indicating that it is easier to identify the emotion of the image by introducing local emotion region into the whole image than by using only the whole image. While the whole image fusion saliency target recognition image emotional accuracy will be higher.

Compared with experiment 3, the accuracy of experiment 4 is improved by 4%, indicating that the accuracy of image emotion classification can be further improved by adding emotion region of face target on the basis of fusion of the whole image and saliency target. Meanwhile, the accuracy of SFPN-CNN model is higher than that of VGG16 in saliency target emotion recognition.

Table 2. Results comparison with different methods

No.	Method	Accuracy/%		
		FI	Twitter	ArtPhoto
1	FCNN	0.8431	0.7532	0.7299
2	AR+concatenation	0.8746	0.8217	0.7591
3	GMEI	0.8995	0.8276	0.7661
4	W+S+F+Max	0.9132	0.8626	0.7924

In the experiment, the saliency target detection algorithm SS-HED is applied to obtain the images containing saliency targets in each dataset, and the new dataset ArtPhoto_Obj, Twitter_Obj and FI_Obj are formed. The statistical results of the three saliency target images are shown in Table 3. As can be seen from Table 1 and Table 3, most images have saliency targets.

Table 3. Statistics on the number of saliency object

Saliency target data set	Number of saliency targets
ArtPhoto_Obj	590
Twitter_Obj	1144
FI_Obj	20079

In the saliency target data set, 80% are randomly selected as the training set and 20% as the testing set. Since ArtPhoto_Obj and Twitter_Obj data sets are relatively small, in order to accelerate the convergence speed, parameters of the FCNN model trained by the whole image are used as initialization parameters to train saliency target emotion analysis model. AdamOptimizer is used for the model, and the learning rate is 0.001. To prevent over-fitting, L2 regularization with a coefficient of 0.02 is added, and a Dropout layer with a coefficient of 0.5 is added after the fully connected layer.

In order to illustrate the effectiveness of splicing multi-scale semantic feature maps and the influence of supervised new feature mapping learning on saliency target emotion recognition, three groups of comparative experiments are designed respectively, as shown in Table 4. Experiment 1 is the result of saliency target emotion recognition with VGG16. Experiment 3 is the recognition result of saliency target emotion recognition model SFPN-CNN. In experiment 2, the recognition result of supervised new feature mapping is removed in experiment 3. Comparing experiments 1 and 2, we can see that the recognition effect of splicing multi-scale high-level semantic feature map is better than that of original VGG16. According to experiments 2 and 3, the accuracy of saliency target emotion classification can be further improved by simultaneously supervising the learning of feature mapping N_P3, N_P4 and N_P5.

Table 4. Experimental results of adding multi-scale feature map joint and supervised new feature maps

No.	Method	Accuracy/%		
		FI_Obj	Twitter_Obj	ArtPhoto_Obj
1	VGG16	0.8219	0.7311	0.7507
2	FPN-CNN	0.8259	0.7356	0.7715
3	SFPN-CNN	0.8338	0.7491	0.8132

5 Conclusion

Generally, the expression of emotion depends on the whole image. Studies have shown that saliency target and face target region are also important factors in emotion expression. Therefore, an image emotion analysis method based on multi-channel is proposed. Firstly, the multi-scale saliency target feature map obtained from the feature pyramid is supervised, and the multi-scale semantic feature map is integrated to construct the saliency target emotion recognition model. Secondly, the monitoring module is introduced into CNN, and different weight losses are applied to different monitoring modules to construct a face emotion recognition model. Finally, the above two models are fused with the whole image emotion recognition model. The validity of the proposed model in image emotion recognition is verified on three public data sets. The experimental results show that the emotion analysis method of proposed model can obtain higher accuracy than the method of emotion analysis directly from the whole image. However, in the emotion analysis stage of face target, the accuracy of face emotion recognition is reduced because the face data set contains a large number of side faces and partial occlusion data. So the next work will consider emotion recognition of side faces and occluded faces. In the stage of saliency target emotion analysis, the emotion region only considers the largest saliency target and ignores other small targets that affect the image emotion. Therefore, future works will combine multiple targets to improve the accuracy of saliency target emotion analysis.

6 Acknowledgement

The authors express their acknowledgement for the anonymous review.

References

[1] M. Hasan, E. Rundensteiner, X. Kong, E. Agu, Using Social Sensing to Discover Trends in Public Emotion, in: Proc. 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 2017.

[2] X. Zhang, X. Wang, S.-L. Yin, Multi-modal Data Transfer Learning-based LSTM Method for Speech Emotion Recognition, International Journal of Electronics and Information Engineering 13(2)(2021) 54-65.

[3] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, A. V. Palade, A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots, Neural Computing and Applications 29(2018) 359-373.

[4] J. H. Lui, H. Samani, K. Tien, An affective mood booster robot based on emotional processing unit, in: Proc. 2017 International Automatic Control Conference (CACS), 2017.

[5] Q. Shi, S. Yin, K. Wang, L. Teng, H. Li, Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation, Evolving Systems (2021). https://doi.org/10.1007/s12530-021-09392-3

[6] S.-L. Yin, H. Li, Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13(2020) 5862-5871.

[7] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.-Z. Wang, J. Li, J.-B. Luo, Aesthetics and Emotions in Images, IEEE Signal Processing Magazine 28(5)(2011) 94-115

[8] V. Mavani, S. Raman, K.-P. Miyapuram, Facial Expression Recognition Using Visual Saliency and Deep Learning, in: Proc. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

[9] B. Zhang, C. Quan, F. Ren, Study on CNN in the recognition of emotion in audio and images, in: Proc. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016.

[10]T. Rao, X. Li, M. Xu, Learning Multi-level Deep Representations for Image Emotion Classification, Neural Processing Letters 51(2020) 2043-2061.

[11]D. Greenberg, Large-Scale Social Experimentation in Britain: What Can and Cannot be Learnt from the Employment Retention and Advancement Demonstration, Evaluation 11(2)(2005) 223-242.

[12]Q.-Z. You, J.-B. Luo, H.-L. Jin, J.-C. Yang, Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI Press), 2015.

[13]S. Jindal, S. Singh, Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning, in: Proc. 2015 International Conference on Information Processing (ICIP), 2015.

[14]K.-K. Song, T. Yao, Q. Ling, T. Mei, Boosting Image Sentiment Analysis with Visual Attention, Neurocomputing 312(2018) 218-228.

[15]J. Yang, D. She, M. Sun, M. Cheng, P.-L. Rosin, L. Wang, Visual Sentiment Prediction Based on Automatic Discovery of Affective Regions, IEEE Transactions on Multimedia 20(9)(2018) 2513-2525.

[16]H. Zheng, T. Chen, Q. You, J. Luo, When saliency meets sentiment: Understanding how image content invokes emotion and sentiment, in: Proc. 2017 IEEE International Conference on Image Processing (ICIP), 2017.

[17]Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply Supervised Salient Object Detection with Short Connections, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18]T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[19]S.-L. Yin, H. Li, L. Teng, Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images, Sensing and Imaging 21(2020).

[20]X. Wang, S. Yin, K. Sun, H. Li, J. Liu, S. Karim, GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition, Journal of Applied Science and Engineering 23(3)(2020) 555-561.

[21]S. Yin, H. Li, L. Teng, M. Jiang, S. Karim, An optimised multi-scale fusion method for airport detection in large-scale optical remote sensing images, International Journal of Image and Data Fusion 11(2)(2020) 201-214.

[22]K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks, IEEE Signal Processing Letters 23(10)(2016) 1499-1503.

[23]Y.-Q. Miao, Q. Lei, W. Zhang, M. Zhou, Y.-M. Wen, Research on image sentiment analysis based on multi-visual object fusion, Application Research of Computers 38(4)(2021) 1250-1255. (In Chinese)

[24]Q.-Z. You, J.-B. Luo, H.-L. Jin, J.-C.Yang, Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI Press), 2016.