Journal on Artificial Intelligence
DOI:10.32604/jai.2021.027069
Article

Tech Science Press

Facial Expression Recognition Based on the Fusion of Infrared and Visible Image

Jiancheng Zou[1], Jiaxin Li[1,*], Juncun Wei[1], Zhengzheng Li[1] and Xin Yang[2]

[1]North China University of Technology, Beijing, 100144, China
[2]Department of Computer Science, Middle Tennessee State University, Murfreesboro, 37132, USA
[*]Corresponding Author: Jiaxin Li. Email: lijiaxin10061108@163.com
Received: 10 January 2022; Accepted: 14 January 2022

Abstract: Facial expression recognition is a research hot spot in the fields of computer vision and pattern recognition. However, the existing facial expression recognition models are mainly concentrated in the visible light environment. They have insufficient generalization ability and low recognition accuracy, and are vulnerable to environmental changes such as illumination and distance. In order to solve these problems, we combine the advantages of the infrared and visible images captured simultaneously by array equipment our developed with two infrared and two visible lens, so that the fused image not only has the texture information of visible image, but also has the contrast information of infrared image. On the other hand, we improved the WGAN by adding SSIM and LBP loss functions to ensure the structural similarity between the fused image and infrared image, and also the texture similarity between the fused image and visible image respectively. Finally, a facial expression recognition model Pyconv-SE18 with pyramid convolution and attention mechanism module is designed to extract the important feature information of facial expression in multiple scales. We add cosine distance loss function to reduce the feature difference within the class. Experiment results show that the robustness of expression recognition algorithm to illumination is improved based on the fused images. The accuracy of this model on FER2013 and CK+ public data sets are 69.3% and 94.6%, respectively.

Keywords: Image fusion; expression recognition; pyramid convolution; attention mechanism

1 Introduction

In recent years, facial expression recognition has made great progress in emotion analysis, and has become one of the key technologies of emotion analysis. Facial expressions contain rich feature information. Mehrabian et al. [1] and others' research showed that expression accounts for up to 55% of the information in human daily communication, while language accounts for only 7%. Therefore, facial expression is an important factor in human communication to help us understand the intentions of others. In 1971, American psychologists Ekman and Friesen defined seven basic facial expressions, namely happiness, sadness, anger, fear, surprise, disgust, and neutrality [2].

At present, the classification of visible expression recognition is mostly based on Ekman, in which expressions are divided. Although there are many expression recognition algorithms, the specific process is divided into three parts: face detection, feature extraction and classification recognition. Among them, feature extraction is the most critical step in facial expression recognition, which directly affects the effect of expression recognition. Jung et al. [3] designed a dual channel network structure. The first channel extracts the dynamic apparent features of face image sequence, the second channel extracts the dynamic geometric features of face, and the two channels fuse at the end for facial expression recognition, which

has better recognition effect.

Near-infrared facial expression recognition research is less, mainly due to the lack of near-infrared facial expression database. In 2009, Zhao et al. [4] published the Oulu -CASIA Nir-vis facial expression video database. In 2011, Zhao et al. [5] divided the human face into five regions, combined geometric and apparent features, and used LBP-TOP feature descriptor to extract near-infrared facial expression features for each region. Then, support vector machine is trained to classify and recognize the extracted features; the results show that Near-infrared facial expression recognition has stronger robustness to illumination difference.

At present, facial expression recognition in infrared image mainly focuses on facial expression recognition with thermal feature points. Khan et al. [6] marked the thermal feature points on the infrared face image, took the difference between the thermal intensity values (TIVs) of each thermal feature point as the feature vector, and used multivariable test and linear discriminant analysis for classification and recognition. Finally, they successfully recognized the three expressions of happiness, sadness and disgust. Ahmad et al. [7] used Gauss Laguerre (GL) filter to filter the infrared image to obtain the complex texture features of the infrared image, and used k-nearest neighbor for classification and recognition to achieve high accuracy.

In practical application, the acquisition of near-infrared image is easier than that of thermal infrared image. At the same time, near-infrared images are not sensitive to illumination [8]. However, due to the fuzzy surface texture of near-infrared face image, it cannot well present the specific information of facial features, while visible image has clear texture details. Therefore, through the fusion of visible image and infrared image, the complementarity of visible image and near-infrared image is realized to improve the accuracy of expression recognition [9]. Therefore, this paper studies facial expression recognition based on near-infrared image and visible image fusion.

In the infrared and visible image fusion model, we improve the structure of WGAN. Firstly, the network of the fusion model's generator directly connects the infrared image and the visible image. Secondly, SSIM loss function and LBP loss function are added to the loss function of the generator to retain more information in the source image; In the facial expression recognition model Pyconv-SE18, we modify the original residual block based on the original ResNet18, and extract the multi-scale feature information of face image by adding pyramid convolution and SE module, and add cosine distance function to reduce the difference of intra-class features in feature space and increase the feature distribution between classes to improve the accuracy of facial expression recognition.

2 Related Work

2.1 WGAN

GAN [10] has been successfully applied to image processing. However, GAN is unstable due to network optimization, so it is difficult to train GAN. Therefore, Arjovsky et al. [11] improved GAN by Wasserstein distance, known as WGAN. WGAN uses Wasserstein distance instead of Jason Shannon divergence to compare the distribution of real data and generated data [12]. In addition, in WGAN, the weight of the discriminator should be clipped into a compact space $[-C, C]$, and Lipschitz constraints should be imposed on the discriminator. WGAN still has a disadvantage that the model is difficult to converge, because weight shearing may lead to gradient explosion or disappearance. Therefore, Gulrajani et al. [13] proposed an improved WGAN as Eq. (1).

$$\min_{G}\max_{D} L_W(D,G) = E_{x\sim P_r}[D(x)] + E_{z\sim P_z}[D(G(z))] + \mu E_{x}[(\left\|\nabla_{x} D(x)\right\|_2 - 1)] \tag{1}$$

2.2 Pyramid Convolution

Pyramid Convolution [14] can process information entered through multiple convolution kernels with different scales. The main advantage of PyConv is multi-scale processing with different spatial resolution and depth. From level 1 to level n, the convolution kernel size increases and the depth decreases. In order to use kernels with different depths at each level of PyConv, using grouping convolution for reference, the

input characteristic mapping is divided into different groups, and the inner core is applied independently for each input characteristic mapping group.

2.3 Channel Attention Mechanism Module

Convolutional neural network assumes that each channel is equally important, but in practice, the importance of different channels is different, and some channels have a great impact on the final classification results. Therefore, it is particularly important to assign more weights to important feature channels. In this paper, the compression excitation module (SENet) [15] is introduced to the features obtained by pyramid convolution. to enhance the response to important feature channels. The structure of SENet is shown in Fig. 1.

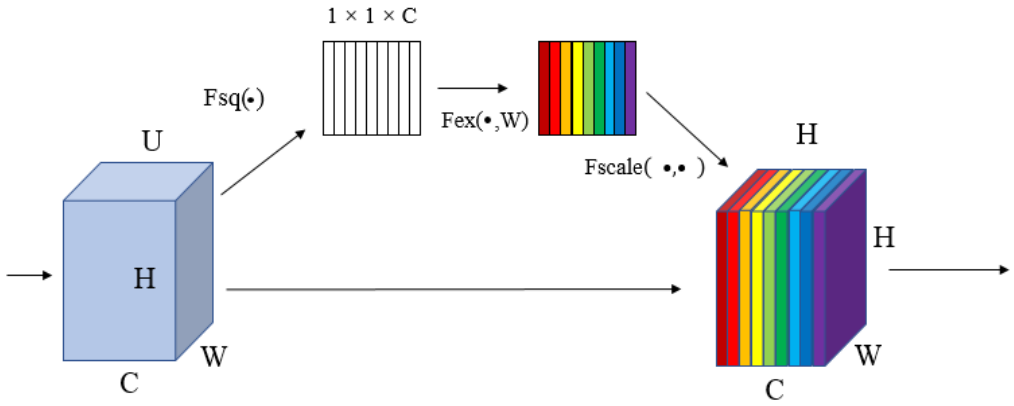


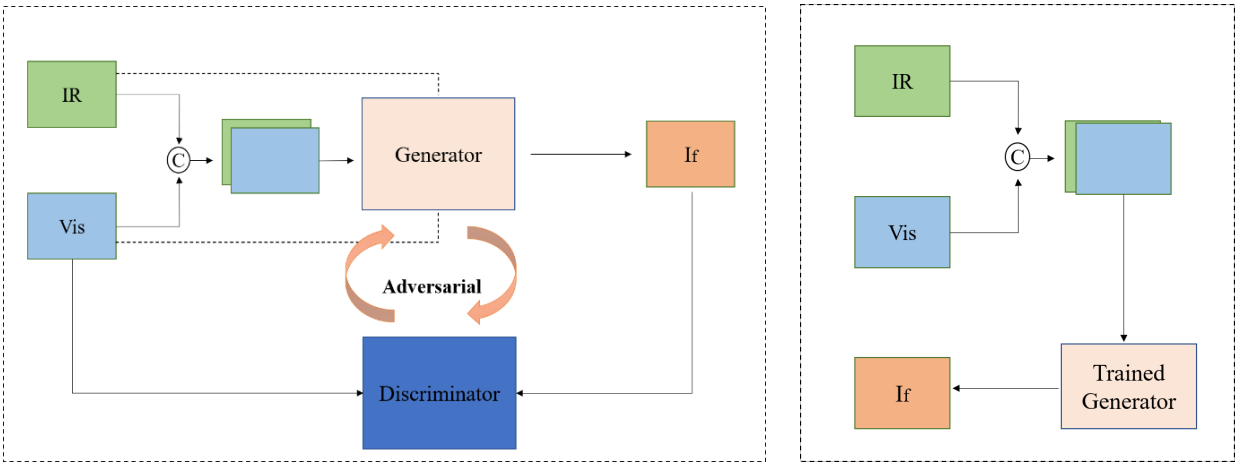
Figure 1: The structure of SENet

By introducing the channel attention mechanism into the multi-scale feature extraction layer, we learn the importance of different feature channels. In the training process of the model, for the useful features related to the expression recognition results, the SE module will increase the weight value of the corresponding feature channel to enhance the feature response. For useless or interfering features, the SE module will reduce the weight value of the corresponding feature channel to weaken the feature response. By introducing the channel attention mechanism, the feature representation ability of the model is improved.

3 Infrared and Visible Image Fusion Model

In this section, we first introduce the framework of our fusion network, which includes generator and discriminator. Then, we discuss the structure of generator and discriminator. Finally, we introduce the details of loss function design.

The structure of the network is shown in Fig. 2. Our network includes two parts: generator and discriminator. In the training phase, firstly, the infrared image (IR) and visible image (VIS) are connected in series in the channel dimension as the input data of the generator. Then, infrared image is directly connected to the shallow layer of the generator network, and the input of the second layer and the third layer. Network is added to extract more edge feature information of the infrared image, such as contrast, brightness information, etc. Directly connect the visible image in the deep layer of the generator network, increase the input of the fourth layer and the fifth layer network, and retain more texture information of the visible image. After the generator, we get the preliminary fusion image. Finally, the fusion image obtained by the generator and the visible image are input into the discriminator at the same time for antagonistic training. With the increase of the number of iterations, the discriminator forces the generator to obtain more detailed texture information from the visible image and obtain an image with better fusion result. In the test phase, we use the trained generator to fuse the input visible and infrared image.



(a) Training process (b) Testing process

Figure 2: Infrared and visible image fusion model

3.1 Network Overview

3.1.1 Structure of Generator

The structure of the generator is shown in Fig. 3. Firstly, the visible image and infrared image are connected as the input image of the generator network. The first and second layers contain 5 × 5 convolution kernels to extract shallow features. The third to fourth layers use 3 × 3 convolution kernels to extract image features. The fifth layers adopt 1 × 1 convolution kernel to reduce the dimension of the spliced features into a single channel image for realizing feature fusion, and obtain the fused image in an end-to-end manner. In the first four layers, we use LeakyReLu [16] activation function to effectively avoid negative threshold neurons in CNN. The fifth layer uses the tanh activation function. The stride of all convolution layers is set to 1. In order to make the input and output have exactly the same size, all layers have added padding.

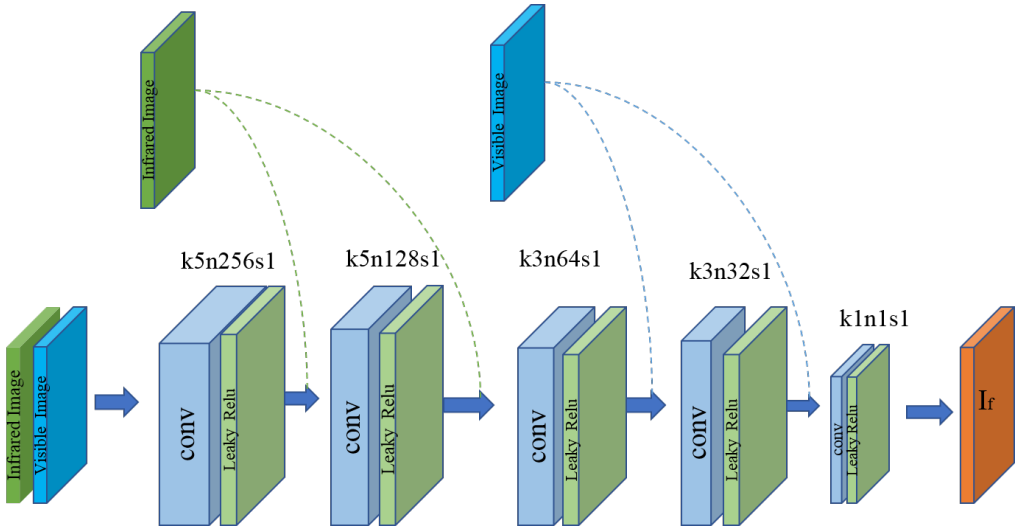


Figure 3: The structure of generator network

3.1.2 Structure of Discriminator

As shown in Fig. 4. The discriminator network consists of six convolution layers and two fully connected layers. The input image of the network is a visible image and a fused image. The size of

convolution kernels is 3 × 3. The number of filters in the convolution layers is set to 64, 128 and 256, respectively. No padding is set for all convolutional layers. the stride in the first, third and fifth layers is set to 1, and other layers are set to 2. Leaky ReLu [16] activation function is used in each convolution layer. Due to the introduction of WGAN, we can judge the similarity between the distribution of the fused image and the distribution of the source image by calculating the Wasserstein distance between the visible image and the fused image. Thus, we give up the sigmiod cross entropy layer.

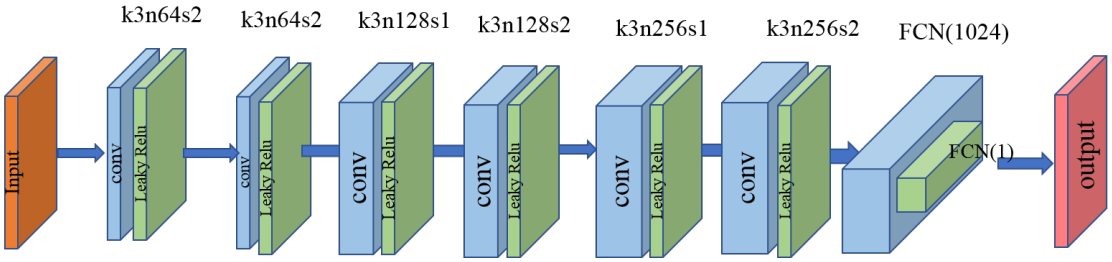


Figure 4: The structure of discriminator network

3.2 Loss Function

In our network structure, the loss function includes the loss of generator and the loss of discriminator.

3.2.1 Loss Function of Generator

The loss function of the generator is defined as Eq. (2), which includes two parts: confrontation loss and content loss.

$$L = L_{adv}(G) + \lambda L_c \tag{2}$$

where, L represents the total loss of the generator, the first term is the confrontation loss between the generator G and the discriminator D_{vis}, and the second term represents the content loss. is the equilibrium coefficient. It consists of the following two parts, see Eq. (3).

$$L_c = \frac{1}{HW}[\partial(\left\|I_f - I_r\right\|_F^2 + (1 - SSIM(I_f, I_r))) + \beta(\left\|LBP(I_f) - LBP(I_{vi})\right\|_F^2 + Gradient(I_f, I_{vi}))] \tag{3}$$

where h and W represent the height and width of the input image respectively and $\|\square\|_F$ represents the Frobenius norm of the matrix. ∂ and β are constant values that control the balance of the two terms. The first item forces the fused image I_f to retain pixel intensity information from the infrared image. In addition, $SSIM_f$ is added to make I_f obtain more structure information from I_{ir}.

The similarity calculation algorithm $SSIM(x, y)$ is used to calculate the brightness, contrast and structure differences between two image x and y. The calculation Eq. (4) is as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x^2 + \sigma_x^2 + c_2)} \tag{4}$$

where μ represents the mean and σ represents the standard deviation. The larger the SSIM, and the higher the structural similarity between two images. Thus, in the L_c, we use $1 - SSIM(I_f, I_r)$ as a part of the L_c to ensure the structural similarity between the fused image and the infrared image [17].

Since a part of the texture information in the visible image can be characterized by gradient information [18], a gradient operation item is added in the content loss to make the fused image retain more gradient information in the visible image. Eq. (5) is as follows:

$$Gradient(x,y)=\frac{1}{M}\sum_{n=1}^{M}(\nabla x-\nabla y)^2 \tag{5}$$

where M is the number of pixels of the image and ∇ is the gradient calculation operation. However, it is obviously not enough to only use the gradient information to preserve the texture information in the visible image for the fused image. Therefore, in the second item, LBP is applied to measure the texture similarity between the fused image and the visible image [19]. As shown in Eq. (6).

$$LBP(xc,yc)=\sum_{p=0}^{p-1}2^p s(i_p-i_c) \tag{6}$$

where (xc,yc) is the center pixel i_c with intensity value, and i_p represents the intensity value of adjacent pixels, P represents the $p\text{-}th$ pixel of adjacent pixels, and s represents the symbol function. The above is the content loss of generator. the antagonistic loss between G and D_{vis} is described below. $L(G)$ is defined as [20], as shown in Eq. (7), where Z represents generated data and Pg represents generated data distribution (generated data $g(z)\square Pg$).

$$L_{advers}(G)=-E_{z\square p_g}[D(z)] \tag{7}$$

3.2.2 Loss Function of Discriminator

The loss function of the discriminator $L_{D_{vi}}$ is defined as shown in Eq. (8).

$$L_{D_{vi}}=-E_{x\square p_{vi}}[D_{vi}(x)]+E_{z\square p_g}[D_{vi}(z)]+\lambda_1 E_{\hat{x}\square}[\left\|\nabla_{\hat{x}\square} D_{vi}(\hat{x})\right\|_2-1] \tag{8}$$

where $L_{D_{vi}}$ represents the loss of D_{vi}. The first two terms represent Wasserstein distance estimation. The last item is the gradient penalty of network regularization. Pg represents the generated data distribution and P_{vi} represents the visible image data distribution, λ_1 is constant weighting parameter.

4 Facial Expression Recognition Model

4.1 PyConv-SE18 Network Model

First, a 7×7 convolution is used to extract image feature information. Then, referring to the network structure of ResNet18, 3×3 convolution in the residual network is replaced by pyramid convolution and the other 3×3 convolution is replaced with 1×1 convolution [21]. It is used to extract multi-scale feature information of face image. At the same time, after 1×1 convolution, SE module is added to distribute the attention weight in the channel dimension, so that the extracted feature information can gather with the key parts of facial expression and obtain more discriminative local information, so as to improve the accuracy of facial expression recognition. Finally, the full connection layer is used to get the classification results of expressions. The specific structure of PyConv-SE18 model is shown in Fig. 5.

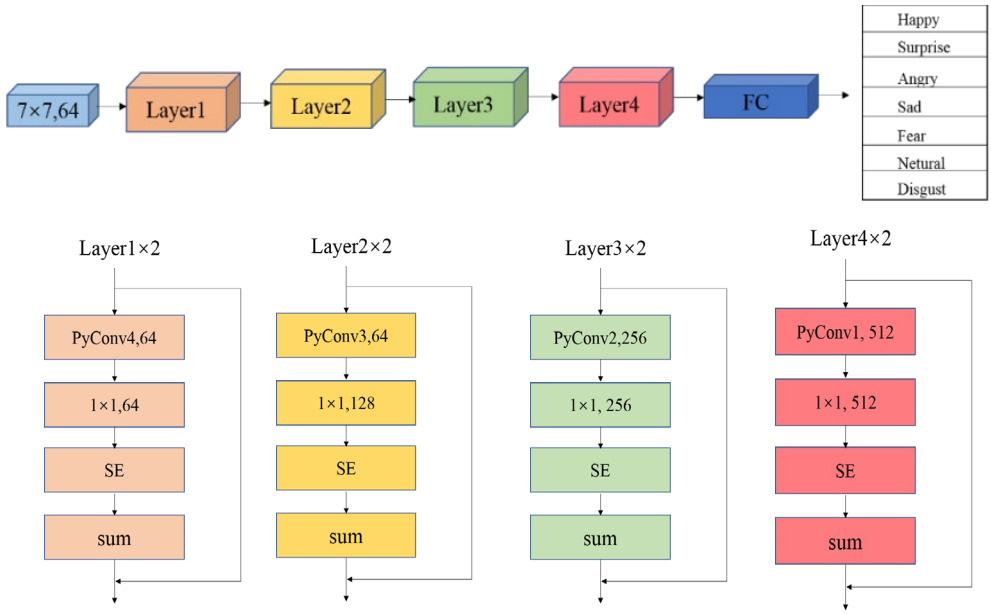


Figure 5: PyConv-SE18 network model

4.2 Loss Function

In the expression recognition task, cross entropy (CE) is a common loss measurement function, and its formula is expressed as Eq. (9).

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{9}$$

where $p(x)$ represents the real distribution of sample. $q(x)$ represents the distribution predicted by the model. It can be approximated $q(x)$ by repeated training. The cross-entropy loss function is defined as follows:

$$L_{cross} = -\sum_{i=1}^{C} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{j=1}^{n} e^{W_{y_i}^T x_i}} \tag{10}$$

where, x_i is the input of the last full connection layer of the sample, and its category is y_i, w_j is the weight parameter of the $j\text{-}th$ full connection layer, and w_{y_i} is the weight parameter of the $y_i\text{-}th$ full connection layer. C is the batch size in one training, and N is the number of categories. In order to reduce the difference of intra-class features in the feature space and increase the inter-class feature distribution, the cosine distance loss function [22] is added to the cross entropy function as the loss function of this network model, so as to improve the accuracy of recognition effect, the formula is shown as Eq. (11).

$$L_{lmc} = \frac{1}{C}\sum_{i=1}^{C} -\log \frac{e^{s(\cos(\theta_{yi},i)-m)}}{e^{s(\cos(\theta_{yi},i)-m)} + \sum_{j\neq yi} e^{s\cos(\theta_j,i)}} \tag{11}$$

$$W_j = \frac{W_j}{\|W_j\|}, x_i = \frac{x_i}{\|x_i\|}, \cos(\theta_j, i) = W_j{}^T x_i$$

In the cosine distance loss function, where m is the decision margin, which can change the cosine distance of network weight W_j and feature vector x_i into $\cos\theta - m$, and adjust the distance between

features through M, S is the scaling factor, usually an integer greater than 1, is the angle between W_j and x_i. The total loss function is defined as cross entropy loss plus cosine distance loss, the formula is shown as Eq. (12), where, λ is the equilibrium coefficient. Set λ to 0.1 in the experiment.

$$L = L_{cross} + \lambda L_{lmc} \tag{12}$$

5. Results and Analysis

5.1 Comparative Experiment

In this paper, FER2013 and CK+ facial expression data sets are selected for the experiment. Fer2013 data set has a total of 35,887 pictures, including 28,709 pictures in the training set, 3589 in the verification set and the test set, respectively. The expressions are divided into seven kinds: neutral, happy, sad, surprised, disgusted, angry and afraid.

The CK+ dataset includes 593 video sequences extracted from 123 objects. This paper selects the last three frames of 327 video sequences, a total of 981 pictures. In the training stage, in order to prevent the network from over fitting, the input image is randomly cropped, and the clipping size is 44 × 44, and the data set is expanded by rotation, translation, mirroring and other operations. The batch size was 128 and the initial learning rate was 0.1. A total of 100 epochs were performed. The algorithm in this paper is compared with other mainstream algorithms in FER2013 and CK+ data set. The experimental results are shown in Table 1 and Table 2.

The algorithm in this paper is compared with other mainstream algorithms in Fer2013 and CK+ data sets. According to the comparison results in Table 1 and Table 2, the accuracy of facial expression recognition of Pyconv-se18 network model proposed in this paper on Fer2013 and CK+ data set is 69.3% and 94.6%, respectively. Compared with other algorithms, the recognition rate of the algorithm in this paper has obvious advantages. The main reason is that the pyramid convolution and attention mechanism module is added on the basis of the original Resnet18, and the cosine distance loss function is superimposed on the loss function, so that the feature information of human face can be extracted more effectively and the extracted features can be classified correctly, so as to improve the accuracy of facial expression recognition. At the same time, because Pyconv-SE18 in this paper is a facial expression recognition model based on infrared optical fusion image, because the fused image has the characteristics of infrared image and has good robustness to environmental changes, it can effectively avoid the disadvantage that the mainstream facial expression recognition model is vulnerable to illumination.

Table 1: Accuracy of different algorithms on Fer2013

Network	Accuracy (%)
Parallel CNN [23]	65.6
Resnet18 [24]	66.4
Resnet18+CBAM [25]	67.1
Xception [26]	68.1
Ours	69.3

Table 2: Accuracy of different algorithms on CK+

Network	Accuracy (%)
Improved Le-Net-5 [27]	83.74
AlexNet [28]	87.03
Zou et al. [29]	91.5
Fei et al. [30]	93.5
Ours	94.6

5.2 Test Examples

In order to further test the performance of the facial expression recognition system based on optical infrared image fusion proposed in this paper, by processing the video collected by the equipment, we collected the relevant videos of 7 kinds of expressions of the tester at 3 m and 3.5 m away from the lens when the light is sufficient and slightly dark, and obtained 100 consecutive images of the same kind of expression for testing. The Experimental equipment is shown as Fig. 6, The resolution of video collected by the equipment is 640 × 480.

Figure 6: Experimental equipment

The process of face recognition is shown in Fig. 7, and the recognition accuracy is shown in Table 3 and Table 4.

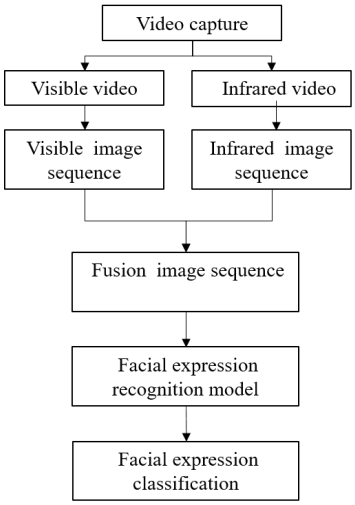


Figure 7: Facial expression recognition based on the fusion of infrared and visible image

Table 3: Facial expression recognition accuracy with a distance of 3 m

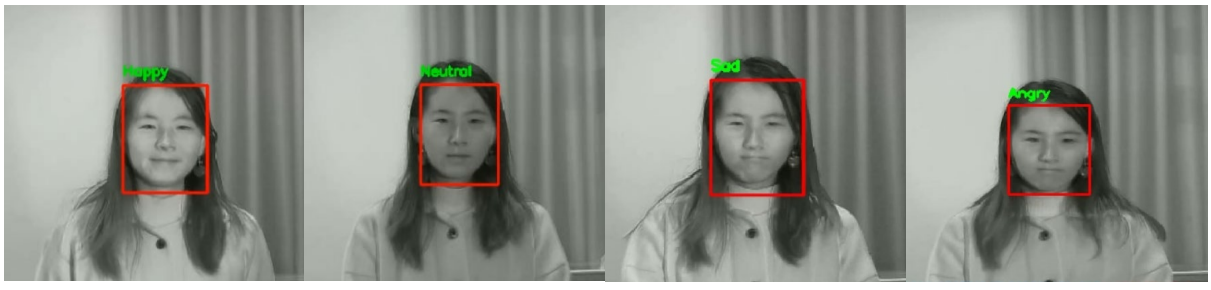
Image sequence	Expression	Number of test frames		Error recognition frames		Recognition rate		Average recognition rate	
		light	dark	light	dark	light	dark	light	dark
Visible image	Neutral	100		0	0	100	100	91%	88.2%
	Happy	100		9	12	91	88		
	Sad	100		11	15	89	85		
	Angry	100		17	20	83	80		
Infrared image	Neutral	100		5	6	95	94	76%	76.7%
	Happy	100		17	16	83	84		

	Sad	100	45	46	55	54		
	Angry	100	29	25	71	75		
	Neutral	100	0	0	100	100		
	Happy	100	8	10	92	90		
Fusion image	Sad	100	15	13	85	87	91.2%	91.5%
	Angry	100	12	11	88	89		

Table 4: Expression recognition accuracy with a distance of 3.5 m

Image sequence	Expression	Number of test frames		Error recognition frames		Recognition rate		Average recognition rate	
		light	dark	light	dark	light	dark	light	dark
	Neutral	100		0	0	100	100		
Visible image	Happy	100		7	10	93	90	92.75%	90.5%
	Sad	100		8	9	92	89		
	Angry	100		14	17	86	83		
	Neutral	100		4	5	96	95		
Infrared image	Happy	100		13	15	87	85	79%	78.75%
	Sad	100		42	44	58	56		
	Angry	100		25	21	75	79		
	Neutral	100		0	0	100	100		
Fusion image	Happy	100		7	8	93	94	93.25%	93%
	Sad	100		9	7	91	93		
	Angry	100		11	15	89	85		

It can be concluded from the above that under the same illumination and distance, the expression recognition accuracy of infrared image is the lowest, because the infrared image is fuzzy and the image texture is not clear, so it is impossible to accurately judge the facial expression. The accuracy of visible image is higher, and the accuracy of fused image is slightly higher than that of visible image. At the same distance, when the illumination changes, the accuracy of facial expression recognition of visible image decreases, because visible image is sensitive to illumination and environment. The recognition accuracy of the fused image is the highest, because the fused image generated based on WGAN not only have the texture of the visible image, but also has high contrast and brightness information of the infrared image, which can avoid the disadvantage that the visible light is sensitive to light. Thus, the recognition accuracy is the highest. The test results at 3.5 m are shown as Fig. 8.



(a) Neutral (b) Happy (c) Angry (d) Sad

Figure 8: Test results of different facial expressions

6 Conclusion

The current facial expression recognition model is easily affected by environmental changes such as illumination and distance, We propose a facial expression recognition model based on the fusion of infrared and visible image, The collected infrared and visible images are fused through the self-designed equipment with array infrared and visible lens, By using the fused image with complementary information features for facial expression recognition, the robustness of the expression recognition algorithm to illumination is improved. Our fusion model based on the improved WGAN, by adding SSIM and LBP loss functions, we ensure the structural similarity between the fusion image and infrared image and the texture similarity between the fusion image and visible image. At the same time, a facial expression recognition model PyConv-SE18 with pyramid convolution and attention mechanism module is proposed. The important feature information of facial expression is extracted through multi-scale, and the cosine distance loss function is added to reduce the feature difference within the class and improve the accuracy of facial expression recognition. The accuracy rates of 69.3% and 94.6% are obtained on the public data sets FER2013 and CK+, which are better than many facial recognition models. However, the model in this paper still has many short comings. For example, when the distance between the tester and the lens is too large or the light is too dark, the results of facial expression recognition are still inaccurate. In the follow-up work, the model will be further optimized to achieve better results.

Funding Statement: The work of this paper is supported by the Innovation Capability Improvement Project of Science and Technology Service for the Elderly by Beijing Municipal Science & Technology Commission under Grant No. Z1811000009218012, and the Innovation Project Foundation NCUT.

Conflicts of Interest: We have no conflicts of interest to report regarding the present study.

References

[1] A. Mehrabian and J. A. Russell, "An approach to environmental psychology," in *American Psychological Association*, Cambridge, MA, ENG: MIT Press, pp. 222–253, 1980.

[2] P. E. Ekman and W. Friesen, "Pictures of facial affect," in *Psychology*. Palo Alto, CA, Consulting Psychologists Press, 1976.

[3] T. Baltrušaitis, P. Robinson and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. WACV*, Santa Rosa, CA, USA, pp. 1–10, 2016.

[4] M. Taini, G. Zhao, S. Z. Li and M. Pietikainen, "Facial expression recognition from near-infrared video sequences," in *Proc. ICPR*, Tampa Florida, FL, USA, pp. 1–4, 2008.

[5] G. Zhao, X. Huang, M. Taini, S. Z. Li and M. Pietikainen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[6] M. M. Khan, R. D. Ward and M. Ingleby, "Automated classification and recognition of facial expressions using infrared thermal imaging," in *Proc. CIS*, Singapore, SG, 2005.

[7] P. Ahmad, Y. Svetlana and G. Maria, "An efficient facial expression recognition system in infrared images," in *Proc. EST*, Cambridge, MA, UK, pp. 239–244, 2013.

[8] X. Ju, "An overview of face manipulation detection," *Journal of Cyber Security*, vol. 2, no. 4, pp. 197–207, 2020.

[9] S. M. Mostafa, "Clustering algorithms: Taxonomy, comparison, and empirical analysis in 2D datasets," *Journal on Artificial Intelligence*, vol. 2, no.4, pp. 189–215, 2020.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial nets," in *Proc. NIPS*, Montreal, MTL, Canada, pp. 2672–2680, 2014.

[11] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, Sydney, SYD, Australia, pp. 214–223, 2017.

[12] J. Xu and W. Chen, "Convolutional neural network-based identity recognition using ECG at different water temperatures during bathing," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1807–1819, 2022.

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS,* California, CA, USA, pp. 5767–5777, 2017.

[14] I. C. Duta, L. Liu, F. Zhu and L. Shao, "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition," *arXiv preprint arXiv*: 2006.11538, 2020.

[15] J. Huang, S. Li, S. Gang and S. Albanie, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2013, 2019.

[16] A. L. Maas, A. Y. Hannun and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, Georgia, GA, USA, 2013.

[17] L. Palaniappan and K. Selvaraj, "Profile and rating similarity analysis for recommendation systems using deep learning," *Computer Systems Science and Engineering*, vol. 41, no. 3, pp. 903–917, 2022.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham *et al.,* "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR,* Hawaii*,* HI, USA, pp. 4681–4690, 2017.

[19] J. Ma, C. Chen, C. Li and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[20] T. Ojala, M. Pietikainen and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proc. ICPR,* Jerusalem, Israel, pp. 582–585, 1994.

[21] G. Abosamra and H. Oqaibi, "An optimized deep residual network with a depth concatenated block for handwritten characters classification," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 1–28, 2021.

[22] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong *et al.,* "Cosface: large margin cosine loss for deep face recognition," in *Proc. CVPR*, Utah, UT, USA, 2018.

[23] L. L. Xu, S. M. Zhang and J. L. Zhao, "Expression recognition algorithm for parallel convolutional neural network," *Journal of Image and Graphics*, vol. 24, no. 2, pp. 227–236, 2019.

[24] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR,* Las Vegas, LV, USA, pp. 770–778, 2016.

[25] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, MUC, German, pp. 3–19, 2018.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, Hawaii, HI, USA, pp. 1800–1807, 2017.

[27] Y. Li, X. Z. Lin and M. Y. Jiang, "Facial expression recognition with cross-connect LeNet-5 network," *Automatica Sinica,* vol. 44, no. 1, pp. 176–182, 2018.

[28] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Nevada, NV, USA, 2012.

[29] J. C. Zou and H, Deng, "An automatic facial expression recognition method based on convolution neural network," *Journal of North China University of Technology*, vol. 31, no. 5, pp. 51–56, 2019.

[30] Z. Fei, E. Yang, D. Li, S. Butler and H. Zhou, "Combining deep neural network with traditional classifier to recognize facial expressions," in *Proc. ICAC*, State of New Jersey, NJ, pp. 1–6, 2019.