symmetry

Article

Facial Expression Recognition Based on Dual-Channel Fusion with Edge Features

Xiaoyu Tang [1,2,*], Sirui Liu [1], Qiuchi Xiang [2], Jintao Cheng [1], Huifang He [3] and Bohuan Xue [4]

[1] School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China
[2] Xingzhi College, South China Normal University, Shanwei 516600, China
[3] Guangdong Engineering Polytechnic, Guangzhou 510520, China
[4] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
* Correspondence: tangxy@scnu.edu.cn; Tel.: +86-139-2898-6468

Abstract: In the era of artificial intelligence, accomplishing emotion recognition in human–computer interaction is a key work. Expressions contain plentiful information about human emotion. We found that the canny edge detector can significantly help improve facial expression recognition performance. A canny edge detector based dual-channel network using the OI-network and EI-Net is proposed, which does not add an additional redundant network layer and training. We discussed the fusion parameters of α and β using ablation experiments. The method was verified in CK+, Fer2013, and RafDb datasets and achieved a good result.

Keywords: facial expression recognition; channel weighting; feature fusion; edge detection

check for updates

Citation: Tang, X.; Liu, S.; Xiang, Q.; Cheng, J.; He, H.; Xue, B. Facial Expression Recognition Based on Dual-Channel Fusion with Edge Features. *Symmetry* **2022**, *14*, 2651. https://doi.org/10.3390/sym14122651

Academic Editors: Chengxi Zhang, Jin Wu and Chong Li

Received: 10 November 2022
Accepted: 2 December 2022
Published: 15 December 2022

1. Introduction

Facial expression recognition (FER) is an important research direction in the affective computing field [1]. The psychologist Mehrabian's research shows that emotional expression = 7% language + 38% voice + 55% facial expressions [2]. This research shows that facial expressions play an important role in human emotional judgment. The accurate recognition of facial expression helps to improve the effect of human–computer interaction. At present, FER has been applied in many fields, such as intelligent teaching, medical facilities, security monitoring, psychological warning, and driver fatigue monitoring.

Feature extraction is a crucial step of FER. Early feature extraction is mainly based on handcrafted methods, such as HOG [3], SIFT [4], and LBP [5]. Among them, HOG and SIFT are calculated by the local gradient of the image. Up to now, these methods have still been common in the FER task, because they can extract the local information of the image in a targeted manner. Both SIFT and HOG have a certain degree of robustness on the impact of illumination, but a common problem with them is a large amount of calculation.

With the development of deep learning, features extracted by deep neural network, an end-to-end method, has become popular. Typical neural network models are AlexNet [6], VGGNet [7], ResNet [8], and GAN [9]. However, there is much redundant information in the extracted features when using the methods of convolutional neural networks. The redundant information is hardly helpful for FER tasks, and some features can be classified as noise. These problems affect the recognition accuracy of FER, which cannot fulfill the current FER needs well.

The gradient information of the image contains much information about the shape of the object, and the edge provides critical information in FER. Although the original images contain the edge information, the deep network trained by original image will lose this information. Aiming at solving the above problem, this paper proposes a dual-channel FER method based on edge feature fusion. The purpose is to effectively focus on the edge information of facial expressions while maintaining high-level semantic features. Adding the network channel for extracting edge features can remove confounding factors in the image. Moreover, the problem domain is simplified, effectively reducing the amount of data and the number of layers required by the deep CNN model. Experiments show that the method proposed in this paper is feasible and can improve the accuracy and robustness of each benchmark dataset. The main contribution of this paper are:

1. This paper proposes a dual-channel FER method based on edge feature fusion to enhance edges, discuss the weight of two channels, and analyze the contribution of edges.

2. This paper proves in the experiment that more than just extracting the edge feature is needed to provide all the information needed for FER. Facts have proved that it can only be used as a supplement to the original image feature, and more information that determines FER is included in the original image.

3. The FER method proposed in this paper performs more robustly on three datasets, including CK+, Fer2013, and RafDb.

2. Related Work

In the area of FER, handcrafted features have been frequently used. Appearance-based features, one of the traditional handcrafted features methods, focus on extracting low-level features, such as edges and corners. Hu et al. [10] proposed a new local feature recognition center-symmetric local octonary pattern (CS-LOP), which improved the LBP algorithm and the CS-LBP algorithm. Meena et al. [11] proposed using graph signal processing (GSP) to solve the problem of HOG high-dimensional feature vectors and computational complexity. In 2021, Shanthi and Nickolas [12] combined LBP features and LNEP features to encode the relationship between pixels, realizing an effective texture representation. These feature extraction methods all focus on low-level features. The handcrafted feature-based methods above have the disadvantages of two points. Firstly, they are useful in datasets with small samples. On the contrary, they are useless in others, such as wild datasets. Secondly, they usually only consider a single feature.

In recent years, convolutional neural networks (CNNs), proposed for image classification tasks, have achieved better recognition performance. Usually, networks with deep layers can extract high-level features but bring more noise and network parameters with excessive redundant information. So, researchers began to find methods to solve these problems. Xie et al. [13] made more targeted improvements to CNNs, mainly including an attention-based salient expression region descriptor (SERD) and a multipath mutation suppression network (MPVS-Net). Minaee et al. [14] proposed an FER method based on an attention-based convolutional network, focusing on critical face parts. Wang, Kai et al. [15] proposed Region Attention Networks (RAN) to solve the obstacles of occlusion and posture changes in FER. They used the attention mechanism and improved CNNs to emphasize the learning of key regions, thereby improving the effectiveness of FER. Actually, low-level features are easily lost. In recent years, some researchers tried to combine handcrafted features with CNNs. G Levi and T Hassner [16] proposed using LBP to preprocess an image and perform deep neural network learning with the original image. H Zhang, B Huang, and G Tian [17] proposed to use the LBP for preprocessing and then weighted fusion with the original image through dual-channel training and added time series, using LSTM to achieve image-sequence-based FER. F. Bougourzi [18] and others proposed the FTDS method, which combined shallow features and in-depth features to identify six basic facial expressions in static images. The paper used HOG, LPQ, and the BSIF to extract low-level features, while using l-PML and VGG-Face networks to extract high-level features. Yu et al. [19] proposed a multitask global–local FER method, using global facial models and part-based models to learn global spatial information features and key dynamic features.

CNNs easily lose low-level features, such as edges. Thought handcrafted feature-based methods can obtain low-level features, these features are sensitive to illumination conditions. However, the edges are stable and contain critical information. This paper proposes a dual-channel FER method to extract features from the original image and edge image, using a shallow network to focus on edge features. The method determined the contribution of the edge in the FER.

3. Proposed Method

This paper uses a dual-channel network model, as shown in Figure 1. We use two channels to extract features from an original image and an edge image, respectively. The two networks are based on VggNet [7]. The channel extracts original image features, called the original image network (OI-Net), which consists of 12 layers of convolution. Another one is called the edge image network (EI-Net), consisting of 8 layers of convolution. The feature fusion of the two channels is performed by the given original image feature parameter α and edge feature parameter β, and Softmax is used for classification.

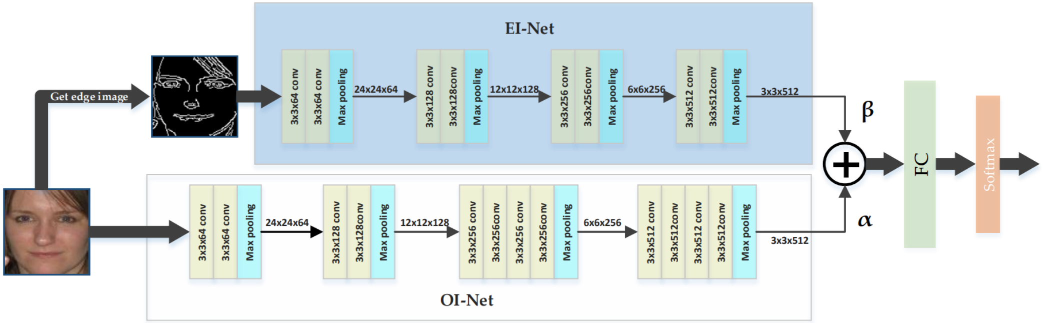


Figure 1. Our proposed network architectur.

3.1. *Edge Image Feature Extraction*

The edge contains critical information about the face, including three essential senses needed for FER. It contains information such as facial muscle texture and wrinkles corresponding to different expressions, which improve the accuracy of the recognition. Extracting the edge features can effectively reduce redundant information and, meanwhile, distinguish the information that the original image is focused on.

Edge information can be extracted by the shallow network; this low-level feature is important and easily lost in the deeper network. To avoid the lack of edge information, we consider extracting the edge feature as a supplement to the OI-Net and discuss its effect.

When performing edge feature extraction, the gradient information obtained by the solution is very sensitive to noise. Therefore, this paper chooses canny edge detection [20] that can reduce noise interference to extract edge images. Canny edge detection can remove noise while introducing two thresholds, T1 and T2, to better preserve edges.

The specific edge detection steps are as follows. The first is Gaussian filtering. The purpose is to remove noise using formula (1);

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}} f(x,y) \tag{1}$$

Among them is the gray value of the image for a position, and it is the gray value of the image after Gaussian filtering.

The second step is to calculate the image gradient value and gradient direction; see formulas (2)–(7);

$$G\_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{2}$$

$$G\_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & +1 \end{bmatrix} \tag{3}$$

$$G_x = G(x,y) \times G\_x \tag{4}$$

$$G_y = G(x,y) \times G\_y \tag{5}$$

$$G = \sqrt{(G_x^2 + G_y^2)} \tag{6}$$

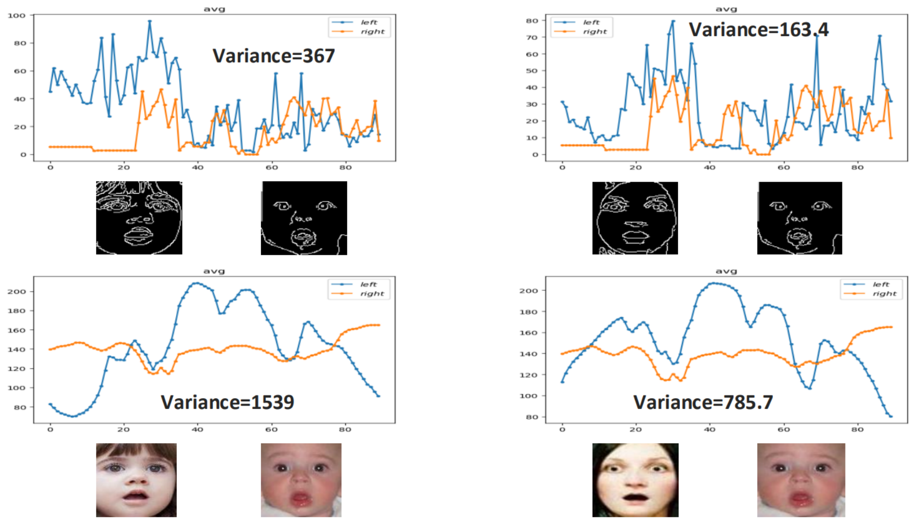
$$\theta = arctan\left(\frac{G_y}{G_x}\right) \tag{7}$$

G_x and G_y are the convolution factors needed to calculate the x-direction and the y-direction, respectively. By convolving them with $G(x,y)$ in a plane, the horizontal and vertical brightness difference approximate values G_x and G_y can be obtained. G is the gradient value, and θ is the gradient direction.

The third step is to perform non-maximum suppression on the gradient image. In the process of Gaussian filtering, the edge may be amplified. Use non-maximum suppression to filter non-edge points. The main idea is first to determine the edge, then compare the gradient direction of the edge with the gradient of neighboring points to determine whether to keep or discard.

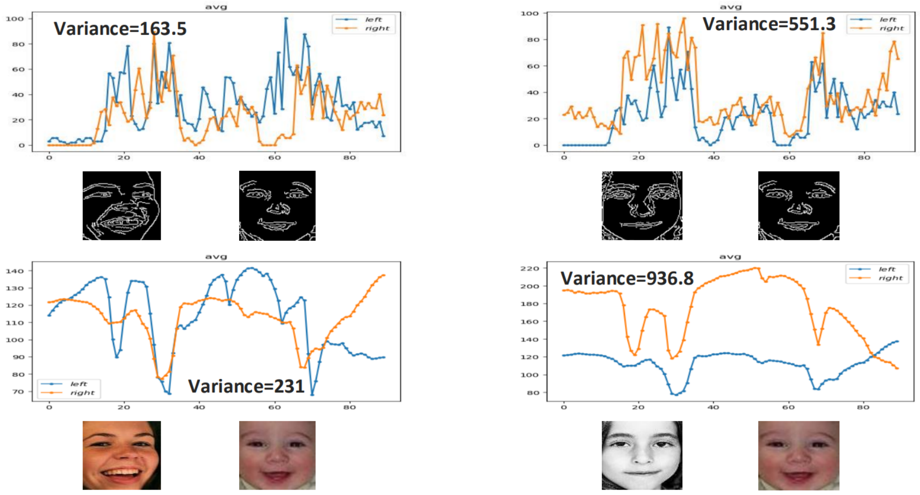
The fourth step is to use dual thresholds for edge connection. First, a higher threshold is used to detect the edges with a higher degree of certainty, called strong edges, and then a smaller threshold is used to reveal more edges, called weak edges, and choose to keep those edges connected with the strong edges and, finally, form the edges that close the entire image.

In this part, we discuss the image similarity of the same category of facial expression from, respectively, the original image and edge image. Taking the RafDb dataset as the example, the images were cropped to a size of 90*90. The variance indicates the similarity of the image. The smaller the variance, the higher the similarity. We calculated the variance of the pixels and compared the similarity between the original image and the edge image with the same category. Figure 2 shows the average value of pixels in each row of the image. Among the same category, the edge information can show more obvious consistency. When used as a supplement to the features of the original image, an edge can better emphasize the commonality of similar expressions.



(**a**) Comparison of frightened expressions

Figure 2. *Cont.*



(**b**) Comparison of happy expressions

Figure 2. Image similarity comparison.

3.2. Feature Fusion

The different networks can extract different feature information. The fusion of two dissimilar nets makes the model complementary. To discover and discuss the usefulness of facial edges, we perform a weighted fusion of the features extracted from the two channels.

We use the two parameters of α and β to denote the parameters of OI-Net and EI-Net and calculate the weighted feature of them; see Equations (8) and (9).

$$F_1 = \alpha f_1 \tag{8}$$

$$F_2 = \beta f_2 \tag{9}$$

where f_1 and f_2 are the feature maps of OI-Net and EI-Net, and F_1 and F_2 are the weighted feature maps.

When $\alpha = 1$ and $\beta = 0$, it means that only the feature map from OI-Net is used for softmax classification. Additionally, it is converse when $\alpha = 0$ and $\beta = 1$.

After obtaining the weighted feature map, we use weight fusion to aggregate them.

$$F = Add(F_1, F_2) \tag{10}$$

The size of F_1 and F_2 are 512 dimensions. The size of F is 512 dimensions and is the same. It denotes more supplementary information.

In the next section, we discuss the contribution of F_1 and F_2 with an ablation experiment and explore the important proportion of edges in FER.

4. Experiments

In this section, we conduct a detailed experimental analysis and verify the designed model on three different facial expression datasets, namely CK+ [21], Fer2013 [22], and RafDB [23]. Additionally, the effectiveness of this method is demonstrated through experiments.

4.1. Datasets

- CK+

The CK+ dataset is a relatively extensive laboratory control dataset used for FER. The dataset contains 593 video sequences of 123 subjects, and each sequence contains changes from neutral to peak expressions. According to the facial motion coding system, 327 sequences are labeled with seven basic expression tags (anger, contempt, disgust, fear, happiness, sadness, and surprise).

- Fer2013

This dataset contains 28,709 training images, 3589 verification images, and 3589 test images. Each image has a pixel size of 48*48. It contains seven facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral.

- RafDb

The RafDb dataset is a facial expression dataset containing basic expressions or compound expressions annotated by 40 well-trained human annotators on facial expression images. The dataset contains 30,000 facial expression images. In the experiment, we only use 12,271 face images as the training set and 3068 face images as the test set, which contains seven basic expressions: surprised, fear, disgust, happiness, sadness, anger, and neutral expression.

4.2. Experimental Details Settings

The experimental platform configuration is as follows: Ubuntu18.04 system, Intel Xeon Gold 5218 with a CPU frequency of 2.3 GHz, and NVIDIA RTX2080Ti graphics card using Pytorch1.2 learning framework and CUDA framework 10.2.

In the experiment, the same hyperparameters are used for the experiment. Using the SGD optimizer, the weight attenuation coefficient is 5×10^{-4} the momentum is set to 0.9, and the initial learning rate is set to 0.01. The number of iterations on the Fer2013 dataset and RafDB dataset is 150. After 50 iterations, the learning rate is attenuated every five iterations. Each attenuation is 0.9 times the original, and the batch size is set to 128. In order to avoid overfitting, for the Fer2013 dataset, we randomly crop the 48*48 images into the 44*44 size and perform random flips for data enhancement; for the RafDB dataset, we randomly select 100*100 images, cut them into a size of 90*90, and perform random flips for data enhancement. The CK+ dataset uses a 10-fold cross-validation method. The data are randomly divided into ten parts. Each time, nine parts are taken as the training set, and the other part is used as the test set. Then, the accuracy of the ten tests is averaged as the final accurate result of the dataset. The batch size is set to 32 and the number of iterations is set to 40 times; after 15 iterations, the learning rate is attenuated to 0.9 times the original every five iterations, and the 48*48 size image is also randomly cropped to the 44*44 size and randomly flipped for data enhancement.

4.3. Ablation Experiment

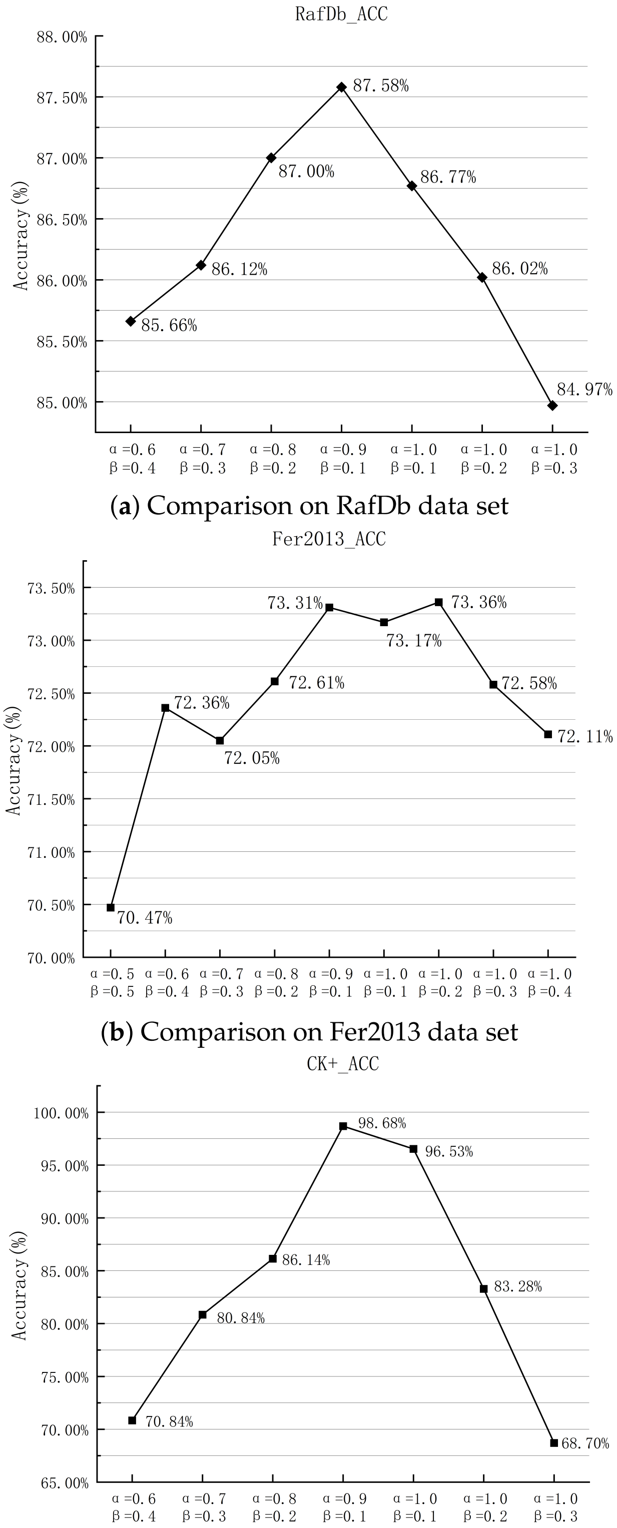
4.3.1. Discussion of α and β

In this part, we discuss the parameters of α and β on the RafDb dataset, Fer2013 dataset, and CK+ dataset. In the RafDB dataset, we show the different αs from 0.6 to 1 and the different βs from 0.1 to 0.4. Figure 3a is a broken line chart of the accuracy of the RafDB corresponding to different α and β parameters. We find that when α = 0.9 and β = 0.1, the accuracy rate is as high as 87.58%.

In the Fer2013 dataset, we showed different αs from 0.5 to 1 and different βs from 0.1 to 0.5. Figure 3b is a broken line graph of the Fer2013 accuracy rate when different α and β parameters are selected. We could find that when α = 1 and β = 0.2, the accuracy rate is as high as 73.36%.

In the CK+ dataset, we showed different αs from 0.6 to 1 and βs from 0.1 to 0.4. Figure 3c is a broken line chart of CK+ accuracy when selecting different α and β parameters. When α = 0.9 and β = 0.1, the accuracy rate reaches 98.68%.

From the experiment results between RafDb, Fer2013, and CK+, we found that α = [0.9, 1] and β = [0.1, 0.2] can achieve the best recognition effect when performing feature fusion on OI-Net and EI-Net, which shows that features extracted by EI-Net can indeed be used to supplement the features extracted by OI-Net, but not as the primary characterization information. In the FER task, the primary characterization information is still the original image feature; the original image often loses some critical information during feature extraction, and edge features can ameliorate this problem.



(**a**) Comparison on RafDb data set

(**b**) Comparison on Fer2013 data set

(**c**) Comparison on CK+ data set

Figure 3. Image similarity comparison.

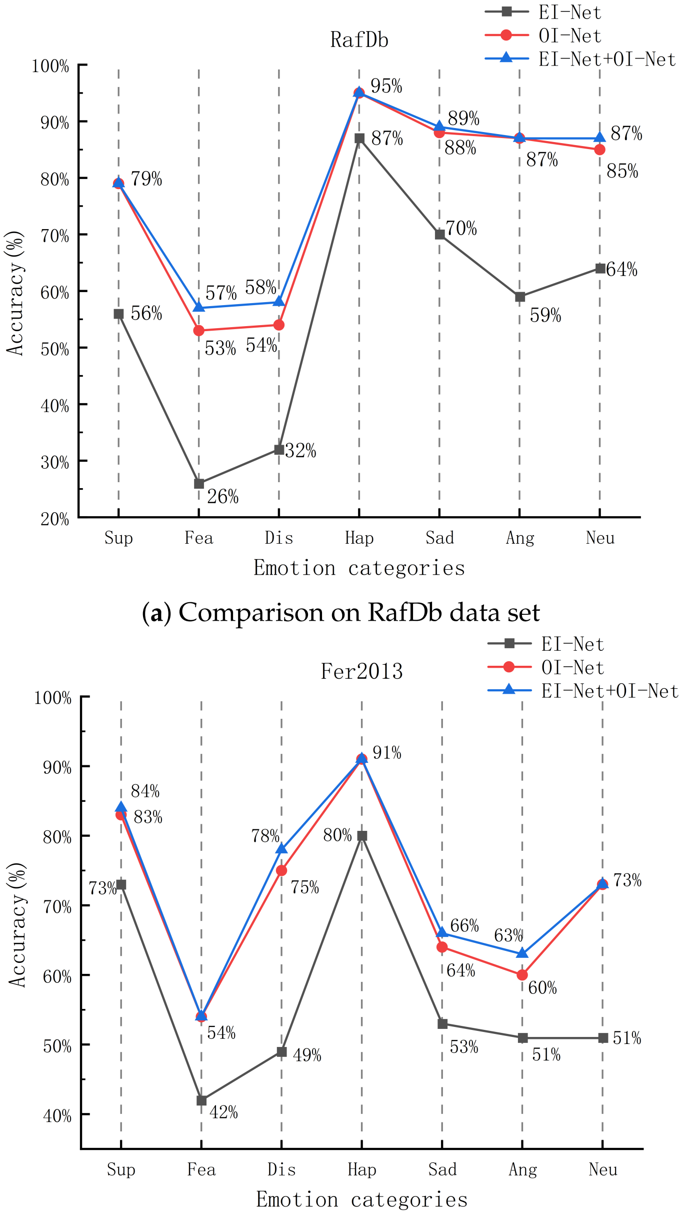
4.3.2. Comparison of Proposed Method with Single Channel

In this part, three methods were compared, respectively, the proposed method (OI-Net+EI-Net), only OI-Net, and only EI-Net. By comparing the recognition rates of the three methods on each expression category, we find that the method proposed in this paper can effectively supplement the facial expression information needed for FER tasks. Figure 4a–c are the comparisons of the three methods on the RafDb, Fer2013, and CK+, respectively.

In RafDb, the recognition rates of fear and disgust are usually low. In Figure 4a, in the absence of edge supplementary features, the recognition rate of fear was 53%, and the recognition rate of disgust was 54%. After adding edge features, the recognition rate of both categories increased by 4%.

Similarly, on the Fer2013 dataset, our method effectively improved the recognition rate of the two categories of anger and disgust. In the absence of edge features, the anger recognition rate is 60%. After adding edge features, the recognition rate reached 63%, an increase of 3%. Similarly, the accuracy rate of disgust also increased by 3%.

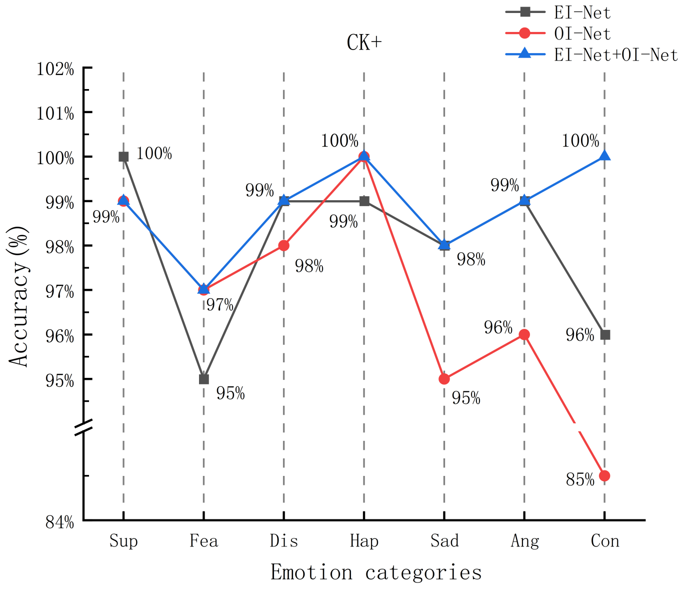
On the CK+ dataset, after adding edge features, the recognition rates of the three expression categories of anger, sadness, and contempt were significantly improved, and contempt increased by 15%.



(**a**) Comparison on RafDb data set

(**b**) Comparison on Fer2013 data set

Figure 4. *Cont.*



(**c**) Comparison on CK+ data set

Figure 4. Comparison of OI-Net+EI-Net, OI-Net, and EI-Net.

4.4. Confusion Matrices and Comparison with State-of-the-Art Methods

4.4.1. Confusion Matrices

The confusion matrices of the method proposed by the paper on the RafDb, Fer2103, and CK+ datasets can be seen in the Table 1a–c. We achieve outstanding results in both three datasets. However, we should take note of the categories of fear and disgust in RafDb, anger and fear in Fer2013, and sadness in CK+. Additionally, happiness always achieves the best recognition accuracy.

Table 1. Confusion matrices for the CK+, Fer2103, and RafDb datasets.

(**a**) **RafDb Confusion matrices**

	Sur	**Fea**	**Dis**	**Hap**	**Sad**	**Ang**	**Neu**
Sur	79%	3%	2%	7%	6%	0	3%
Fea	8%	57%	1%	10%	11%	9%	4%
Dis	5%	0	58%	9%	4%	8%	16%
Hap	0	1%	0	95%	3%	0	1%
Sad	0	1%	0	4%	89%	5%	2%
Ang	0	2%	0	4%	7%	87%	0
Neu	1%	2%	2%	3%	5%	0	87%

(**b**) **Fer2013 Confusion matrices**

	Ang	**Dis**	**Fea**	**Hap**	**Sad**	**Sur**	**Neu**
Ang	99%	1%	0	0	0	0	0
Dis	1%	99%	0	0	0	0	0
Fea	0	0	100%	0	0	0	0
Hap	0	0	0	100%	0	0	0
Sad	0	0	0	0	98%	2%	0
Sur	0	0	0	0	0	98%	2%
Con	0	0	0	0	0	0	100%

Table 1. *Cont.*

(c) CK+ Confusion matrices							
	Ang	**Dis**	**Fea**	**Hap**	**Sad**	**Sur**	**Neu**
Ang	63%	1%	8%	3%	14%	2%	9%
Dis	9%	78%	4%	2%	5%	0	2%
Fea	10%	0	54%	2%	18%	7%	9%
Hap	1%	0	1%	91%	3%	1%	3%
Sad	6%	0	6%	5%	66%	0	17%
Sur	1%	0	7%	3%	2%	84%	3%
Neu	4%	0	4%	4%	14%	1%	73%

4.4.2. Comparison with State-of-the-Art Methods

To verify that the edge features extracted by EI-Net can really supplement the OI-Net, we compare our method with state-of-the-art methods including ReCNN, CNN-SIFT, and so on. Table 2 illustrates the comparison of accuracies between different methods. Our method shows superior performance in RafDb, Fer2013, and CK+. It can be seen that some methods used pretraining, and we have not.

Table 2. Compared with the accuracy of existing methods

Method	**Pretraining**	**RafDb**	**Fer2013**	**CK+**
[24] Gan et al.	✓	85.69%	-	96.28%
[25] ACNN	✓	85.07%	-	-
[26] SHCNN	-	-	69.10%	-
[27] SCN	✓	87.03%	-	-
[15] Wang et al.	✓	86.90%	-	-
[28] Gao H	-	-	65.2%	-
[14] Minaee et al.	-	-	70.02%	98.0%
[29] MBCC-CNN	-	-	71.52%	98.48%
[30] Multiple CNN	-	-	70.1%	94.9%
[31] Xie et al.	-	-	72.67%	97.11%
[32] CNN+ SIFT	-	-	72.85%	93.46%
[33] DCNN+RLPS	-	72.84%	72.35%	-
[34] ReCNN	✓	87.06%	-	-
[35] LBAN-IL	✓	77.80%	73.11%	-
Ours	-	87.58%	73.36%	98.68%

5. Conclusions

In order to find a simple method to reserve edges and discuss their contribution, we proposed a dual-channel facial expression recognition method to fuse the edge image features and original image features by EI-Net and OI-Net. The weighted fusion method is selected to merge the two network channels, and the fusion parameters are discussed. Through ablation experiments, it is determined that the recognition effect is best when α = [0.9, 1] and β = [0.1, 0.2], which also shows that the primary characterization information is still the OI-Net channel.

This paper verifies the proposed method on the three datasets Fer2013, CK+, and RafDb. From the experimental results, the accuracy rate reaches 87.58% on RAFDB, 73.36% on Fer2013, and up to 98.68% on CK+. The experiment demonstrates the effectiveness of the method proposed in this paper.

In the future, we will try to discuss the importance of more low-level features and find a way to achieve feature fusion adaptive parameters.

Author Contributions: Conceptualization, X.T.; methodology, S.L.; software, J.C.; validation J.C. and B.X.; data curation, Q.X. and H.H.; writing—original draft preparation, S.L.; writing—review and editing, Q.X., J.C., H.H. and B.X.; visualization, B.X.; supervision, X.T.; project administration, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Project of Special Funds for the Cultivation of Guangdong CollegeStudents' Scientific and Technological Innovation ("Climbing Program" Special Funds) grant number pdjh2021a0126.

Data Availability Statement: Not applicable

Acknowledgments: This work was supported by the Project of Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (Grant No. pdjh2021a0126), any opinions expressed in this work are those of the authors and do not necessarily represent those of the funding agencies.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kumari, J.; Rajesh, R.; Pooja, K. Facial expression recognition: A survey. *Procedia Comput. Sci.* **2015**, *58*, 486–491. [CrossRef]
2. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; The MIT Press: Cambridge, MA, USA, 1974.
3. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
5. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
7. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Caramihale, T.; Popescu, D.; Ichim, L. Emotion classification using a tensorflow generative adversarial network implementation. *Symmetry* **2018**, *10*, 414. [CrossRef]
10. Hu, M.; Zheng, Y.; Yang, C.; Wang, X.; He, L.; Ren, F. Facial expression recognition using fusion features based on center-symmetric local octonary pattern. *IEEE Access* **2019**, *7*, 29882–29890. [CrossRef]
11. Meena, H.K.; Joshi, S.D.; Sharma, K.K. Facial expression recognition using graph signal processing on HOG. *IETE J. Res.* **2021**, *67*, 667–673. [CrossRef]
12. Shanthi, P.; Nickolas, S. An efficient automatic facial expression recognition using local neighborhood feature fusion. *Multimed. Tools Appl.* **2021**, *80*, 10187–10212. [CrossRef]
13. Xie, S.; Hu, H.; Wu, Y. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recognit.* **2019**, *92*, 177–191. [CrossRef]
14. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [CrossRef] [PubMed]
15. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef] [PubMed]
16. 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, Seattle, WA, USA, 9–13 November 2015; pp. 503–510.
17. Zhang, H.; Huang, B.; Tian, G. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognit. Lett.* **2020**, *131*, 128–134. [CrossRef]
18. Bougourzi, F.; Dornaika, F.; Mokrani, K.; Taleb-Ahmed, A.; Ruichek, Y. Fusing Transformed Deep and Shallow features (FTDS) for image-based facial expression recognition. *Expert Syst. Appl.* **2020**, *156*, 113459. [CrossRef]
19. Yu, M.; Zheng, H.; Peng, Z.; Dong, J.; Du, H. Facial expression recognition based on a multi-task global-local network. *Pattern Recognit. Lett.* **2020**, *131*, 166–171. [CrossRef]
20. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, 679–698. [CrossRef]
21. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
22. Carrier, P.L.; Courville, A.; Goodfellow, I.J.; Mirza, M.; Bengio, Y. *FER-2013 Face Database*; Universit de Montral: Montral, QC, Canada, 2013.

23. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
24. Gan, Y.; Chen, J.; Yang, Z.; Xu, L. Multiple attention network for facial expression recognition. *IEEE Access* **2020**, *8*, 7383–7393. [CrossRef]
25. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef]
26. Miao, S.; Xu, H.; Han, Z.; Zhu, Y. Recognizing facial expressions using a shallow convolutional neural network. *IEEE Access* **2019**, *7*, 78000–78011. [CrossRef]
27. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6897–6906.
28. Gao, H.; Ma, B. A robust improved network for facial expression recognition. *Front. Signal Process.* **2020**, *4*, 4. [CrossRef]
29. Shi, C.; Tan, C.; Wang, L. A facial expression recognition method based on a multibranch cross-connection convolutional neural network. *IEEE Access* **2021**, *9*, 39255–39274. [CrossRef]
30. Chuanjie, Z.; Changming, Z. Facial Expression Recognition Integrating Multiple CNN Models. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 1410–1414.
31. Xie, W.; Shen, L.; Duan, J. Adaptive weighting of handcrafted feature losses for facial expression recognition. *IEEE Trans. Cybern.* **2019**, *51*, 2787–2800. [CrossRef] [PubMed]
32. Wang, H.; Hou, S. Facial expression recognition based on the fusion of CNN and SIFT features. In Proceedings of the 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 17–19 July 2020; pp. 190–194.
33. Li, H.; Xu, H. Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowl.-Based Syst.* **2020**, *204*, 106172. [CrossRef]
34. Xia, Y.; Yu, H.; Wang, X.; Jian, M.; Wang, F.Y. Relation-aware facial expression recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 1143–1154. [CrossRef]
35. Li, H.; Wang, N.; Yu, Y.; Yang, X.; Gao, X. LBAN-IL: A novel method of high discriminative representation for facial expression recognition. *Neurocomputing* **2021**, *432*, 159–169. [CrossRef]