A Bi-directional Residual Network for Image Expression Recognition

Daihong Jiang[1], Sanyou Zhang[2(✉)], Cheng Yu[1], and Chuangeng Tian[1]

[1] Xuzhou University of Technology, Xuzhou 221000, Jiangsu, China
[2] China University of Mining and Technology, Xuzhou 221000, Jiangsu, China

Abstract. In this paper, an improved model based on the combination of residual and inverted residual blocks is proposed for image expression recognition, named as bi-directional residual network. The main objective of the proposed method is to alleviate the problem of feature dispersion due to the deep network level in traditional expression recognition research. In this case, residual block is a good solution. However, residual network with small scale of training data can easily lead to over-fitting, which is often the case for image expression recognition. To improve the robustness of the network during training, inverted residual blocks are therefore adopted. Depending on the organization sequence of residual blocks and inverted residual blocks, three network structures are proposed and studied. Fer2013 and CK+ datasets in facial field are adopted for experiment. The experimental results show that the optimized algorithm improves the accuracy by 2.79% on Fer2013 dataset compared with ResNet-50 models.

Keywords: Residual network · Deep learning · Image recognition · Feature expression

1 Introduction

Image expression recognition plays an important role in social life. It has also become an important research topic in the field of artificial intelligence. For example, according to some previous study [14], there are six basic expressions, including happy, angry, surprise, fear, disgust, and sad. Therefore, the goal of image expression recognition is to design algorithms [21] for machines to recognize image expressions, especially the six basic expressions.

So far, image expression recognition has been widely studied over the past several decades. The recent work mainly adopts deep learning to address the task. Liu et al. [17] proposed an expression recognition network based on deep belief network. With Convolutional Neural Networks (CNN) achieving great performance improvement in image recognition tasks, Yu and Zhang [26] proposed an expression recognition network based on deep convolutional network integration and came up with two different network integration strategies. Based on their work, Mollahosseini et al. [15] further proposed a deeper network by introducing the inception layer to the network structure. Recently,

H. Song and D. Jiang (Eds.): SIMUtools 2020, LNICST 369, pp. 173–182, 2021.
https://doi.org/10.1007/978-3-030-72792-5_16

Yang et al. [25] proposed to generate an expressionless face image through conditional generative adversarial network. Since the expression information is recorded in the middle layer of the network in this process, the article proposed a method of performing expression recognition based on residues in the middle layer of the network. Specifically, during the training process, the gradient of a derivative close to 0 will continue to decrease after multiple successive products (in a back propagation process), which makes the networks' training ability poor. However, according to some previous research, the deeper the network, the stronger the fitting ability. Therefore, there is a problem of how to make the deep network easy to train. Additionally, according to Orhan and Pitkow [16], the degradation of the weight matrix causes much worse problem, that is to say, only a small number of hidden units can output valid activation values. As the network growing deeper, this effect becomes worse. The deep residual network (ResNet) [11] can solve these problems. The residual unit used by ResNet is to add a shortcut outside the ordinary multi-layer convolutional layer. In this way, the gradient can be effectively transmitted back to the shallow layers.

However, directly using ResNet for image expression recognition may result in slow calculation speed and poor recognition performance, due to its deep network structure and strong fitting ability. When only relatively small dataset is available, it may become easily overfitting, leading to degraded performance. To address these problems, this paper proposes a bi-directional residual network (Bi-ResNet) which combines the inverted residual blocks [19] and residual blocks. Inverted residual block was proposed in MobileNetV2 [19], whose main idea is to replace the ordinary convolution of the residual block with DepthWise (DW) convolution and PointWise (PW) convolution. In DW convolution, a convolution kernel is responsible for the convolution of a channel, that is, convolution in a two-dimensional plane. DW convolution is followed by PW convolution, which combines the weights of the individual channels of the DW convolution into one new feature. These calculations effectively improve the calculation speed. In this paper, the network structure combining residual block and inverted residual block is studied. Specifically, three kinds of network structures, alternated residual connection, IR residual connection and RI residual connection, are proposed. Compared with traditional residual network structures, it is verified that the network structure of this paper performs better in expression recognition.

The main contributions of the paper are as follows:

In the proposed Bi-directional residual and the inverted residual network, alternated residual connection, IR residual connection and RI residual connection are proposed respectively in terms of the network structure design, with RI residual network finally chosen as the best proposal.

Through the results of multiple sets of experiments, the proposed algorithm is more advantageous than the traditional methods (baseline models). Specifically, RI achieves the best results on the tested two datasets.

2 Related Work

2.1 Traditional Feature Based Methods

Traditional expression recognition methods relies heavily on manual feature extraction. At the same time, it requires designing appropriate classifiers. There are many traditional feature based methods. Some famous manual feature have been applied to expression recognition, including Gabor wavelet transform [8, 23], Histogram of Oriented Gradient (HOG) [23], local binary patterns (LBP) [20], feature extraction based on manifold learning [1, 2, 5, 10, 22, 24]. For classifiers, in [6, 9], support vector machine is used as a classifier. Since the recognition algorithm directly depends on the features extracted manually, and the expression recognition is affected by many factors such as imaging posture, object occlusion, illumination change, etc., the robustness and recognition accuracy of traditional feature based methods still have much room for improvement.

2.2 Deep Learning Based Methods

Liu et al. [17] proposed an expression recognition network based on deep belief network. With Convolutional Neural Networks (CNN) presenting growing prominent advantages in image recognition, Yu and Zhang, in [18, 26], proposed an expression recognition network based on deep convolutional network integration and came up with two different network integration strategies. Based on this, Mollahosseini et al. [15] probed further and deeper by introducing the Inception layer in the network structure. Based on Generative Adversarial Networks (GANs), Yang et al. [25] proposed to generate an expressionless face image through cGAN. Since the expression information is recorded in the middle layer of the network in this process, the article proposed a method of performing expression recognition based on residuals in a network middle layer.

3 Facial Expression Recognition Model

3.1 The Overall Network Structure

The overall structure of the proposed model is shown in Fig. 1. First, the face image is aligned by a pre-processing module and we get normalized face images as input. Before training, the images are augmented by data augmentation methods, such as, rotation, cropping and so on. Then it will be input into one of the three networks proposed in this paper. Finally, a fully connected layer is used. Classification results are obtained by the Sigmoid classification layer. Categorical cross entropy is chosen as the loss function. In the following, the residual block unit structure and the inverted residual block unit structure will be introduced first. Then the three network structures proposed in this paper will be presented.

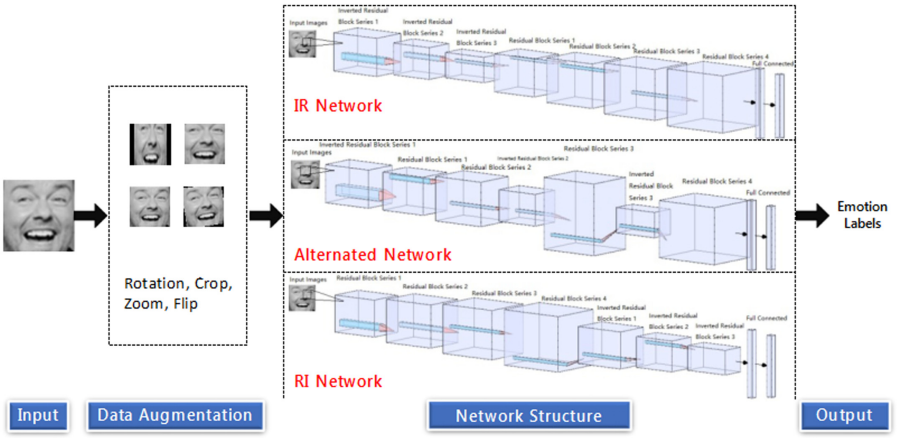

Fig. 1. Framework of the proposed method. With a normalized face input, data augmentation is firstly performed. The augmented training data will be input into one of the three proposed network structures for training. The final output is given by a sigmoid layer for all the three network structures.

3.2 Residual Block Unit

The residual network is a breakthrough of CNN network. It significantly improves the back-propagation ability of network, alleviating problems of gradient dispersion and gradient explosion. Also, it accelerates convergence and effectively improves deep learning ability. The basic unit of the residual network is the residual block [7, 11]. Each residual block contains a short-circuited parameter transfer path. This unique parameter transfer method can transmit information across layers, thereby improving the effect of back-propagation. A residual block is defined by the following function:

$$y = F(x, W_i) + x \tag{1}$$

Where x is the input vector, y is the output vector, and $F(x, W_i)$ represents the residual map to be learned. $F(x, W_i)$ consists of multiple convolutional layers, where x is a shortcut.

For the residual block unit in the above figure, if the x and F dimensions are different and the shortcut changes, the linear projection $h(x_l) = W_l' x$ is used to correspond to the dimension. The function of the residual block unit at this time is as follows:

$$x_{l+1} = h(x_l) + \mathcal{F}(x_l, W_l) \tag{2}$$

In the case where the residual block is stacked multiple times and the input dimension is maintained, the recursive expression can be obtained as follows:

$$x_L = x_l + \sum_{i=1}^{L-1} \mathcal{F}(x_i, W_i) \tag{3}$$

The gradient of the partial differential of a residual block can be obtained by the chain rule:

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\sum_{i=1}^{L-1}\mathcal{F}(x_i, W_i)\right) = \frac{\partial \varepsilon}{\partial x_L} + \frac{\partial \varepsilon}{\partial x_L}\frac{\partial}{\partial x_l}\sum_{i=1}^{L-1}\mathcal{F}(x_i, W_i) \quad (4)$$

In the above formula, $\left(1 + \frac{\partial}{\partial x_l}\sum_{i=1}^{L-1}\mathcal{F}(x_i, W_i)\right)$ is never going to be 0, which means the gradient does not disappear, and it can be reversely inferred from the process of backpropagation that any shallower layers can receive the influence of the previous layer.

3.3 Inverted Residual Block Unit

The inverted residual block unit is proposed in MobileNetV2 [19] with the aim to reduce computational cost. To fulfil this objective, it adopts the DepthWise (DW) separable convolution and linear bottlenecks. Bottleneck and the DW separation convolution is used to raise and decrease dimensions respectively. Its operation consists of three steps: $F(x) = [A \circ N \circ B]x$. Among them, A and B are linear transformations, and N is a nonlinear layer-by-layer transformation. In actual use, $N = ReLU\ 6 \circ DW \circ ReLU\ 6$. Details can be found in [19].

With such a design to replace the convolutional operations, the inverted residual block can largely reduce the computational cost. However, we found that when combined with residual blocks, it can also improve the facial recognition performances under many cases.

3.4 Bi-directional Residual Block Network

Based on the above introduction of residual block and inverted residual block and considering that deep network close to the pyramid shape can improve accuracy, this paper introduces to combine residual and inverted residual blocks to form new network structures. Based on how we organize the sequence of residual and inverted residual blocks, three kinds of network structures are proposed and studied, which are as follows [4, 12].

(1) *Alternated residual and inverted residual network:* The alternated residual and inverted residual network first passes the data through the inverted residual block, then flows it through the residual block. Looping back and forth over multiple such connected inverted residual and residual blocks. Finally Relu(6) is used to activate and fully connected layer is used to produce output. Details of this network structure is shown in Fig. 1.

(2) *RI residual and inverted residual network:* The RI residual and inverted residual network first passes data through multiple consecutive residual blocks, then flows it through multiple consecutive inverted residual blocks.
Finally same structure as the alternated residual and inverted residual network is used to produce output. Details given in Fig. 1. Among all the proposed three networks, in RI network structure, since the residual block is first compressed and then expanded, and the inverted residual block is first expanded and then compressed, the shape of the RI residual and inverted residual network is closer to the shape of a

pyramid, which makes it performs best. The main reason may be because for structures of inverted pyramid shape, it is hard for network to complement information when expanding the network. In experiments, we verified RI network performs best among all the three network structures and it is also better compared with ResNet, MobileNet and MobileNet V2.

(3) IR residual and inverted residual network: The IR residual and inverted residual network adopts an inverse structure of the second network. Data flows through multiple consecutive inverted residual blocks first, then flows through multiple consecutive residual blocks. Details are given in Fig. 2.

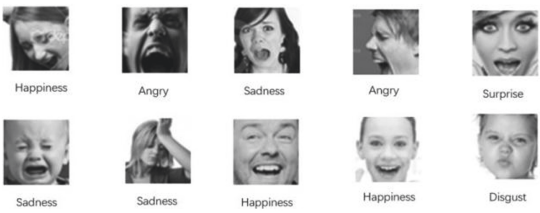

Fig. 2. Example images from Fer2013+

3.5 Network Training Loss and Training Details

The fully connected layer adopts the Dense (2048)-BN-Dense (1024)-BN-Dense (10) structure. All networks in this paper use the classification categorical cross entropy as the loss function.

4 Experiment and Analysis

4.1 Dataset

The size of the facial expression dataset is mostly small. Most of the datasets only label smiling and not smiling. This paper uses the Fer2013 dataset. The Fer2013 dataset contains 35,886 images, including 28,708 images in the training set. There are 3,589 images in the validation set and 3,589 images in the test set. The FER+ [3] tag is used for multi-classification learning. Fer2013+ dataset tags consist of neutral, happiness, surprise, sadness, anger, disgust, fear, contempt, unknown, and NF, of which NF means the image is not human face. The image resolution of the Fer2013 dataset is 48 48 1. The images are collected in the wild with various variations than other facial expression datasets. Example images are given in Fig. 2

On the other hand, this paper also uses the CK+ dataset [13] to further verify the effectiveness of the three models introduced. The CK+ dataset has 593 sequences of images, of which 327 have expression tags, and the expression tags consist of anger, contempt, disgust, fear, happiness, sadness, and surprise.

4.2 Experimental Environment and Parameter Settings

The experiment is run on a computer with CPU of Intel(R) Xeon(R) E5-2698 v4 @ 2.20 GHz. The graphics card is Tesla V100 32G, the graphics card driver version is NVIDIA-SMI 384.125, and the total memory is 528275840 kB. The environment is Ubuntu 16.04.4 with Python version 2.7.12, Keras version 2.2.4, and tensorflow-gpu version 1.8.0.

The batch size is set to 64, the optimization function uses Adam (0.001, 0.5), and the learning rate is set as 0.0001. This paper uses two dropouts, one being Dropout (0.5) and the other being Dropout (0.3)-BN layer - Dropout (0.3). The residual and inverted residual block parameters adopt are embedded in ResNet-50 and MobileNetV2 models.

4.3 Comparison of the Proposed Method and State-of-the-Art Methods on FER2013

In the following, we mainly compare our three versions of methods with the current state-of-the-art methods. Details of the accuracy and loss are summarized in Table 1. We mainly compared our method with ResNet-50, MobileNet and MobileNet V2. Besides, to fully study the state-of-the- art methods, we also combine them with transfer learning, leading to six methods for comparison.

Table 1. Comparison of accuracy and loss of different networks on the FER2013 test set

Network	Accuracy	Loss
ResNet-50	79.19	1.0435
Migration Learning ResNet-50	78.76	1.0395
MobileNet	61.54	1.3921
Migration Learning MobileNet	81.12	1.0004
MobileNetV2	66.68	1.342
Migration Learning MobileNetV2	79.9	1.0209
Interspersed residual and inverted residual network	79.36	1.0313
MR residual and inverted residual network	75.72	1.1
RM residual and inverted residual network	**81.98**	**0.9803**

(1) ResNet-50 is the baseline model in this paper. ResNet-50, MobileNetV1, and MobileNetV2 that used transfer learning were upgraded by - 0.43%, 19.58%, and 13.22% respectively on the Fer2013 dataset compared with results without using transfer learning. Due to the serious over-fitting of ResNet-50, the accuracy is reduced.

(2) Transfer learning adopted by MobileNet and MobileNetV2 witness a 2.36% and 1.14% improvement respectively compared with ResNet-50 in recognition accuracy, which proves that series network like MobileNet has great demand for transfer

learning. MobileNet based network has the advantage of smaller scale hashes and faster convergence.

(3) In the absence of transfer learning, the accuracy of the alternated residual and inverted residual network is 0.17%, 17.82%, and 12.68% higher than that of ResNet-50, MobileNetV1, and MobileNetV2. Compared to MobileNet based network, its demand for transfer learning is significantly reduced, and its over-fitting problem is well suppressed compared to the ResNet-50 network.

(4) The accuracy of the proposed alternated network compared with the transfer learning based MobileNet and MoblineNetV2 is a little lower. However, the stability of our method is far better than that of MobileNet and MoblineNetV2. As the proposed method does not adopt transfer learning, it still got quite comparable result.

(5) Another finding is that, for the IR and RI residual and inverted residual networks, the accuracy of IR is reduced by 3.64% compared with the alternated type, while the RI network takes a leading position. RI network requires smaller number of iterations for fitting, indicating wonderful stability and lowest loss. Besides, it got the best accuracy with a nearly 82% accuracy. The accuracy is improved by 2.62% compared with the alternated network. Also, it compensates for the problem of overfitting which is severe for ResNet.

4.4 Results on CK+ Dataset

In order to further verify the effectiveness of the proposed residual and inverted residual network, the three residual and inverted residual block networks are further verified on the CK+ dataset. The following data is the test results of the validation optimal network after 1200 epoch training, given in Table 2.

Table 2. Comparison of accuracy and loss of different networks on CK+ test set.

Network	test_acc	test_loss
Interspersed residual and inverted residual network	96.70%	0.13781
RM residual and inverted residual network	**97.41%**	**0.09756**
MR residual and inverted residual network	95.92%	0.15331

In a laboratory environment where the image is clearer, the RI residual and inverted residual network still achieves the highest accuracy and lowest loss. Since the residual block is first compressed and then expanded, and the inverted residual block is first expanded and then compressed, the shape of the RI residual and inverted residual network is closer to the shape of a pyramid. When the number of layers being observed increases, the collected features by CNN network change from a lower level to a higher level. By using as many parameters as possible on a high level fitting, better expression recognition performance and fitting ability can be obtained, which also explains why the RI residual and inverted residual network can achieve better results.

5 Conclusion

In this paper, the residual and inverted residual blocks are combined and applied to the research of facial expression recognition. It adopts the two kinds of residual blocks iteratively in the designed network structure, which jointly promotes the feature representation ability of the neural network and alleviates the problem of feature dispersion caused by deep network shown in the traditional facial expression recognition research. The experimental results show that the RI residual network has achieved the best results in the three Bi-directional residual networks proposed in this paper, whose accuracy has increased by 2.62% compared with the alternated bi-directional residual network. It also compensates for the problem of over-fitting of ResNet. Meanwhile, the algorithm inherits the excellent features of ResNet and MobileNetV2 with high accuracy, robust to over-fitting, and impressive recognition performance. The next step of this research is to take a step further and accomplish the study on network confusion matrix data based on the transfer of the RI residual and inverted residual network. Considering that the network structure should be further optimized, it is also possible to try to improve and fine-tune the network structure.

Acknowledgement. The study was supported by the Major Project of Natural Science Research of the Jiangsu Higher Education Institutions of China (18KJA520012), and the Xuzhou Science and Technology Plan Project (KC19197).

References

1. Akiba, T., Suzuki, S., Fukuda, K.: Extremely large minibatch SGD: training ResNet-50 on ImageNet in 15 minutes (2017)
2. Balduzzi, D., Frean, M., Leary, L., Lewis, J.P., Mcwilliams, B.: The shattered gradients problem: if ResNets are the answer, then what is the question? (2017)
3. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction (2016)
4. Bengio, Y.: Knowledge matters: importance of prior information for optimization (2016)
5. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks (2017)
6. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision Pattern Recognition (2017)
7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, pp. 249–256 (2010)
8. Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D., Lin, H.: Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recogn. **45**(1), 80–91 (2012)
9. Guan-Ming, L.U., Guo, M., Xiao-Nan, L.I., Hai-Bo, L.I.: Recognition for expression of pain in neonate using support vector machine. J. Nanjing Univ. Posts Telecommun. **25**(3), 582–587 (2008)
10. Guan-Ming, L.U., Zuo, J.K.: Feature extraction based on two-dimensional locality preserving discriminant analysis. J. Nanjing Univ. Posts Telecommun. (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)

12. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift (2015)
13. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Computer Vision Pattern Recognition Workshops (2010)
14. Mehrabian, A.: Communication without words. Commun. Theory **6**, 193–200 (2008)
15. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: IEEE Winter Conference on Applications of Computer Vision (2016)
16. Orhan, A.E., Pitkow, X.: Skip connections eliminate singularities (2017)
17. Ping, L., Han, S., Meng, Z., Yan, T.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision Pattern Recognition (2014)
18. Rusiecki, A.: Trimmed categorical cross-entropy for deep learning with label noise. Electron. Lett. **55**(6), 319–320 (2019)
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks (2018)
20. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
21. Shan, L., Deng, W.: Deep facial expression recognition: a survey (2018)
22. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-ResNet and the impact of residual connections on learning (2016)
23. Wang, X., Chao, J., Wei, L., Min, H., Ren, F.: Feature fusion of HOG and WLD for facial expression recognition. In: IEEE/SICE International Symposium on System Integration (2013)
24. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the ResNet model for visual recognition, vol. 90 (2016)
25. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. Int. J. Comput. Sci. Eng. (2018)
26. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: ACM on International Conference on Multimodal Interaction (2015)