International Core Journal of Engineering Volume 7 Issue 7, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

Research on Expression Recognition Algorithm based on Deep Learning

Yao Qin*

College of Southwest Minzu University, Chengdu 610000, China.

Abstract

In order to improve the accuracy of expression recognition and reduce the computational load of network, a deep learning-based expression recognition algorithm was proposed. The algorithm mainly preprocesses the image through the spatial transformation network (STN) in the early stage, and completes the image transformation processing required by the network. Then the super resolution (SR) algorithm is used to improve the overall image quality, so as to improve the overall feature information of the input image. Experimental results show that the proposed algorithm has a good performance in the accuracy of facial expression recognition, and can meet the requirements of accuracy and robustness of facial expression recognition algorithm.

Keywords

Facial Expression Recognition; Spatial Transformation Network; Super Resolution.

1. Introduction

With the rapid development of artificial intelligence and the gradual improvement of face recognition system, the research and application of facial expression recognition has become an important field of social application and scientific and technological research. Mollahosscini et al. [1] based on the GoogleNet model, a single-component network architecture is proposed. The improved GoogleNet method is used to recognize facial expressions and achieves good results. Yang Feng et al. [2] proposed a facial expression recognition method based on small-scale kernel convolution, using multi-layer small-scale convolution blocks instead of large convolution blocks to extract facial expression features to achieve expression classification.

The facial expression recognition system proposed in this paper is a classification algorithm based on a deep learning network, with an emphasis on image preprocessing at the front end of the network. First, use the Spatial Transformation Network (STN) to align the input image to improve the various problems of the input image, such as different sizes, different object shapes, and fusion with the background. Then the improved image is improved through the super-resolution (SR) algorithm to improve the overall quality and the overall information of the image. Finally, the classification result of facial expression recognition is finally obtained through the traditional classification network.

2. Methodology

2.1 Spatial transformation network(STN)

The Spatial Transformation Network (STN) is mainly composed of three parts: parameter prediction (Localisation net), coordinate mapping (Grid generator), and pixel acquisition (Sampler). The main task is to complete the transformation of various spaces, expressed in formulas as follows:

$$\text{General layer}: a_{nm}^{l} = \sum_{i=1}^{3}\sum_{j=1}^{3} w_{nm,ij}^{l} a_{ij}^{l-1} \quad (1)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

Localisation net is mainly to generate a 2*3 matrix, through which the image can be translated, rotated, zoomed, etc. to achieve the effect of alignment. For example, from a_{11}^{l-1} translation to a_{21}^{l}, the expression is:

$$a_{21}^{l} = w_{2111}^{l} a_{11}^{l-1} + w_{2112}^{l} a_{12}^{l-1} + w_{2113}^{l} a_{13}^{l} + \ldots + w_{2133}^{l} a_{33}^{l-1} \quad (2)$$

The coordinate mapping (Grid generator) marks the corresponding position points between the original image and the target image according to the generated mapping matrix, and finally the pixel sampling module maps the original pixel to the target pixel according to the corresponding relationship, and finally realizes the spatial conversion of the image. Among them, the transformation relationship of the mapping is:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathrm{T}_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (3)$$

Among them (x_i^t, y_i^t) are the coordinates of the output target image, and (x_i^s, y_i^s) are the coordinates of the original image, A_θ representing the affine relationship. The spatial transformation network is an end-to-end network that can be added to any position of the CNN. At the same time, the 2*3 matrix parameters can automatically learn the transformation of features without causing excessive calculations, which helps to reduce the overall cost of network training.

2.2 Super Resolution Algorithm (SR)

Super-resolution algorithm refers to the process of using software or hardware to increase the resolution of the original image, and obtaining a high-resolution image through a series of low-resolution images. The super-resolution algorithm essentially uses the known image information to predict the required pixels [3]. In order to obtain more accurate prediction results, the prediction model of this model is much more complicated than traditional algorithms. Generally, there are multiple convolutional layers and activation layers, which use image information of a large area around the target pixel, including thousands of model parameters.

The EDSR[4] used in this algorithm uses Generative Adversarial NetWork (GAN)[5] to solve the problem of super-resolution. On the basis of the original ResNet [6] network, adjustments and optimizations were made, and some unnecessary batch normalization (BN) layers in the residual structure were removed. The location of the BN layer saves 40% of the memory usage of the model during training. Therefore, a larger-scale network with better performance can be constructed with limited computing resources.

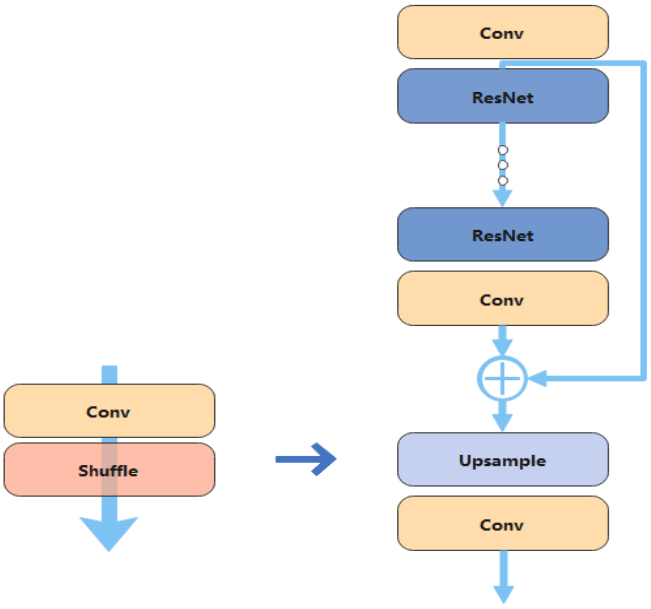

Figure 1. Upsample (left) and ESDR (right) structure

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

In the training process, the L1 norm style loss function is used to optimize the supervision network. The number of residual blocks is set to B=16, and the number of features is F=128. When training, train the low-multiple upsampling model first, and then use the parameters obtained by training the low-multiple upsampling model to initialize the high-multiple upsampling model, so that it can Reduce the training time of high-multiple up-sampling models, and at the same time get better training results[7].

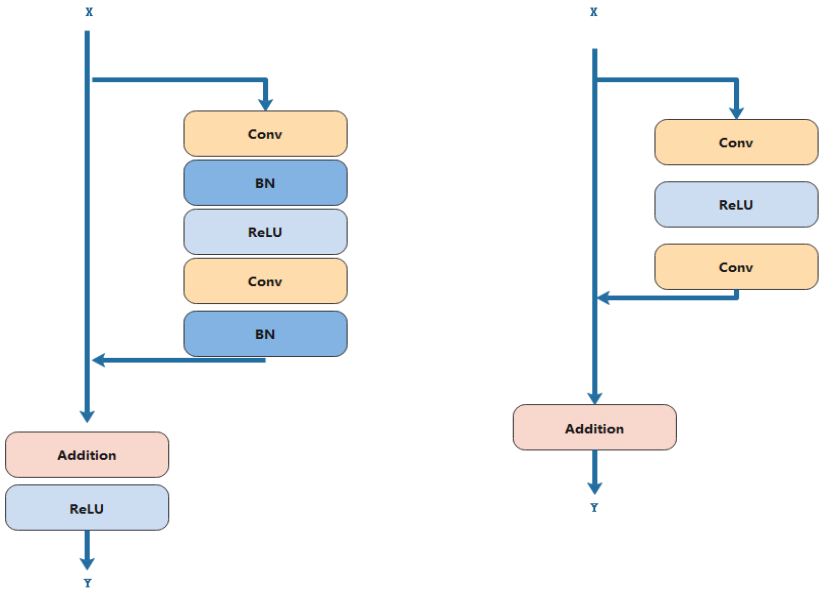

Figure 2. ResNet(right) and SR(left) structure

2.3 Algorithm specific process

In order to improve the recognition accuracy and classification robustness of the entire network, this paper proposes to improve the performance from the source through image preprocessing and reduce the amount of calculation. First, use the spatial transformation network (STN) to align the input images. Since most of the current data sets have inconsistent image sizes and the face position of each image is also different, we first need to align. The network structure of STN is as follows:

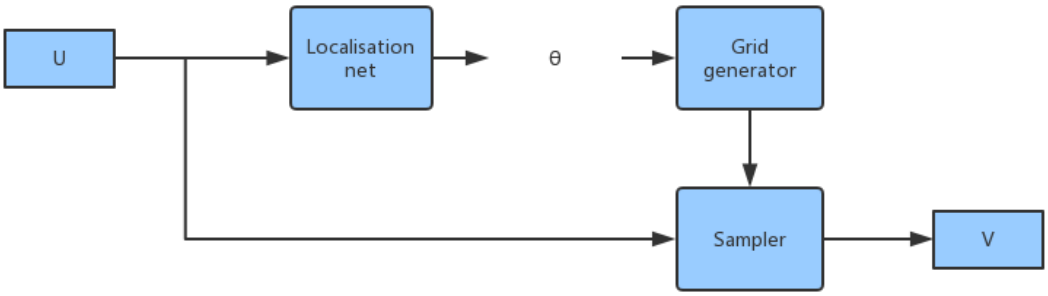

Figure 3. STN structure diagram

Secondly, use SR technology to expand the pixels of the original image to obtain relatively more effective information, and then extract the corresponding features through the feature extraction network. Finally, the extracted features are input to the classifier to obtain the final expression classification. The flow chart of the entire algorithm scheme is as follows:

Figure 4. Algorithm flow

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

3. Results and discussion

3.1 Experimental data

The data sets used in the experiment are CK+[9] and FER2013[10], and Table 1 describes the distribution of the number of pictures in the training set and the test set. The CK+ dataset contains a total of 123 objects and 593 image sequences, of which only 327 have emoticons. The FER2013 data set contains 35886 facial expressions, including 28708 test images (Training), 3586 public verification (Public Test) and non-public verification (Private Test) each, and each image is fixed to 48×48 grayscale. The image is composed of seven expressions, corresponding to the number labels 0-6.

Table 1. Quantity distribution of CK+ and FER2013

	CK+		FER2013	
Emotion	train	test	train	test
anger	45	45	2054	271
disgust	59	89	107	15
fear	25	25	510	78
happiness	69	69	7335	895
neutral	-	-	9030	1102
sadness	28	28	3047	384
surprise	83	83	3173	402
contempt	18	18	115	13
Total	327	327	25371	3160

3.2 Analysis of results

First, the accuracy of the STN network is verified. In order to see the conversion effect more intuitively, it is verified on the MINIST data set. The result is shown in Figure 6, it can be seen that the STN network can obtain higher accuracy under the premise of reducing the amount of calculation, and the recognition accuracy of the network after adding the STN module can reach more than 99%. At the same time, the results in Figure 5 show that the network can achieve image space conversion, can reduce image deflection, so that the image can be extracted more easily.

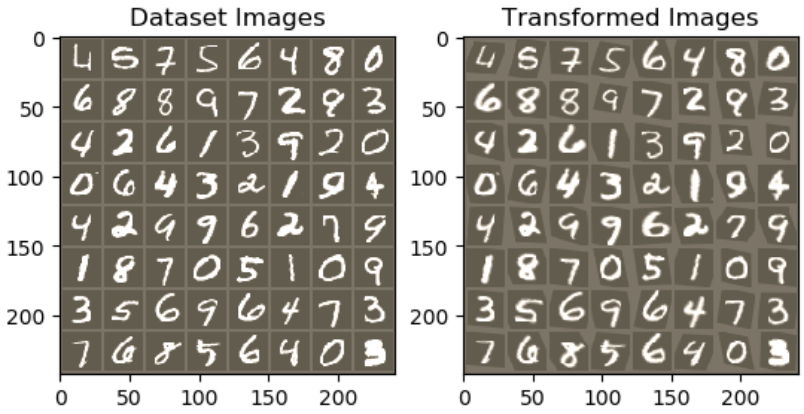

Figure 5. Test results of STN on MINIST

```
Test set: Average loss: 0.0395, Accuracy: 9881/10000 (99%)
```

Figure 6. Testing accuracy and loss of STN against MINIST

In order to verify the effectiveness of the algorithm proposed in this paper, a comparative test of multiple modules was carried out. The input is 48×48 RGB images. This model uses the ADAM

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

optimizer, among them $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$. The minibatch size is 16, and the learning rate is initially set to 10^{-4} update the learning rate after each 2×10^5 minibatch.

In the training process, this article uses L1 and L2 loss functions for comparison. According to the results of the comparison experiment, it can be seen that the L1 loss function is better than the L2 loss function in the network. Through analysis, the reason for this result may be because the images in this experiment contain a variety of features, corresponding to the true distribution of the relevant image composition. The underlying assumption of using L2 as the loss function is that the collected samples all belong to the same Gaussian distribution, that is, the peak value is single, but in fact, most data distributions have more than one peak value. This situation results in a low probability that the trained data set is really distributed, and the performance of the network is not good enough.

Table 2. Precision results of each network frame

Network	CK+	FER2013
Base	87.2	66.5
Base+STN	94.4	71.4
Base+SR	95.1	72.7
Base+STN+SR	97.8	73.4

According to the accuracy analysis of the training results of the data set, it can be seen that the accuracy of the network with the STN module is 7.2 and 4.9 percentage points higher than the accuracy of the network without the STN module in the two data sets, respectively. It can be seen that the pre-processing of the image can be Improve the performance of the network. In contrast, the accuracy of the network with the SR module is 7.9% and 6.2% higher than that of the network without the basic module, and 0.7% and 1.4% higher than the accuracy of the network with the STN module. The accuracy of the algorithm proposed in this paper is the highest in the entire experimental results, with an accuracy of 97.8 in the CK+ data set and an accuracy of 73.4 in the FER2013 data set. This shows that the algorithm proposed in this paper has high recognition performance and robustness.

4. Conclusion

This paper studies the performance and advantages of the spatial transformation network (STN) and the super-resolution algorithm (SR), and integrates them with the facial expression recognition feature extraction network. At present, most algorithms only aim at the modification of the feature extraction network to improve performance. The method proposed in this paper can improve the overall feature information through image preprocessing in the early stage of the recognition network, improve the robustness of the network, and reduce the calculation of the entire network. the amount. The experimental results show that the algorithm in this paper has higher performance and lower computational cost, which provides a new idea for improving the accuracy of facial expression recognition.

Acknowledgments

This work was financially supported by the Southwest Minzu University Graduate Innovation Research Project (Grant No:CX2021SP120).A special acknowledgement should give to Southwest Minzu University for its experimental conditions and technical support.

References

[1] M. Tan, and Q.V. Li (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.International Conference on Machine Learning, vol.24, no.1, p. 6015-6114.

[2] Y Feng, R Liu, T Lu (2020). Facial expression recognition based on small-scale kernel convolution. computer engineering, vol.18, no.1, p.1-8.

[3] T Li(2018). Super-resolution reconstruction algorithm of gray image based on sparse representation.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202107_7(7).0035

[4] B. Lim, S.Son, H.Kim, S.Nah, and K.M.Le e(2017). Enhanced deep residual networks for single image super-resolution. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), vol.16, no.4, p.1132–1140.

[5] Lan J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al(2014). Generative Adversarial Networks. NIPS, vol.47, no.351, p.23-45.

[6] Goodfellow I J, Pouget-Abadie J, Mirza M, et al (2014). Generative Adversarial Networkds[J]. Advances in Neural Information Processing Systems, vol.32, no.3, p.2672-2680.

[7] X.H Hu, J.G Zhang (2020). Research on Image Super-resolution Algorithm Based on Improved Convolutional Neural Network. Application Research of Computers, vol.37, no.341(03), p.313-316+322.

[8] J Ma, X.D Zhang, L Yang(2018). Gesture recognition method combining dense convolution and spatial conversion network. Journal of Electronics and Information, vol.40, no.4, p.951-956.

[9] P. Lucey, J.F. Cohn, T Kanade,et al (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. Proceddings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops.San Francisco, USA.vol235, no.422, p.94-101.

[10] I.J. Goodfellow, A.D. Erh, P.L. Carrier, et al(2013). Challenges in representation learning: A report om three machine learning contests.Proceedings of the 20th International Conference on Neural Information Processing. vol.256, no.86, p.117-124.

[11] S. Gao, Q. Ye, N. Ye(2011). 1-Norm least squares twin support vector machines. Neurocomputing, vol.74, no.17, p.3590-1680.

[12] T.Y. Wei, J Wang, H. Chen (2021). L2-norm prototypical networks for tacking the data shift problem classification. International Journal of Remote Sensing vol. 42, no.9, p.3326-3352.