Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

A Review of Facial Expression Recognition

Jianghai Lan[1,2], Guojun Lin [1,2] *

[1] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Zigong 643000, China
[2] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong 643000, China

[*] **Corresponding author**: Guojun Lin (Email: 386988463@qq.com)

Abstract: With the development of deep learning, deep learning is becoming more and more common for facial recognition. We summarize some widely used public data sets for facial expression recognition; The basic flow of facial expression recognition is briefly introduced. This paper mainly analyzes some existing deep learning methods, especially deep convolutional neural network. The structure analysis and performance comparison of four classical convolutional neural networks (AlexNet, GoogleNet, VGGNet and ResNet) are carried out. Finally, the present research on expression recognition is summarized and prospected.

Keywords: Expression recognition; Deep learning; Feature extraction.

1. Introduction

People can convey a lot of rich information through facial expressions, which play an important role in the communication between people. For example, a look in the teacher's eyes in class can make naughty students correct attitude; A lot of situations in movies also use people's expressions to convey information; There is the annual Spring Festival Gala to see the sketch, the actors on the stage through its exaggerated facial expressions let us smile from ear to ear. There are many applications, in short, the study of human facial expression has a lot of value, can be applied to many aspects. This paper mainly analyzes facial expression recognition through the deep learning method.

Common data sets are shown in Table 1.

Table 1. Common data sets

The data set	Release time	Sample size	The expression distribution
JAFFE	1998	213	6 basic expressions +1 neutral
CK+	2012	981	6 basic expressions +1 neutral
Fer2013	2013	35886	6 basic expressions +1 neutral
RaFD	2010	8040	7 emoticons plus 1 neutral

2. Expression recognition method

Facial expression recognition is mainly divided into four parts: data acquisition, image preprocessing, feature extraction and expression classification. The flow chart is shown in Figure 1.

Feature extraction is the most critical link, and the effect of feature extraction directly affects the accuracy of the final expression recognition. With the rise of deep learning, facial expression feature extraction through deep learning is becoming more and more widespread. Some deep learning networks are introduced below.

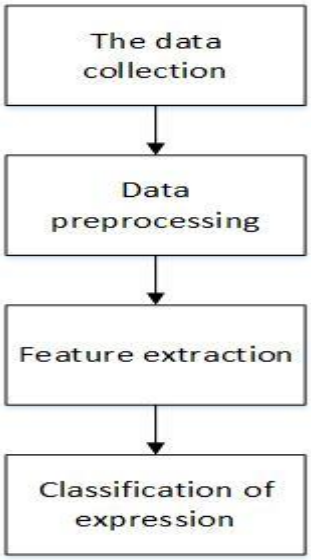

Figure 1. Facial expression Recognition Process

2.1. Deep belief network

The concept of Deep Belief Network (DBN) was put forward by Hinton et al. [1] in 2006. DBN is a special kind of neural network. Generally, it is composed of Restricted Boltamann Machine (RBM) connected in series and Back Propagation (BP) neural network. On the basis of DBN, Boosted Deep Belief Network (BDBN) was proposed by Ping Liu et al in 2014. Features are jointly fine-tuned and selected to form a strong classifier in a new enhanced top-down supervised feature enhancement (BTD-SFS) process, through which highly complex features can be learned from facial images. Literature [2] proposed an AU-inspired Deep Networks (AUDN) composed of three sequential modules, which achieved the best results in the experiments on CK +, MMI and SFEW databases.

2.2. Automatic encoder method

In 1986, Rumelhart proposed the concept of autoencoders, which can extract the implicit features of data and learn to reconstruct the data with these features. Early autoencoders were used in data compression and data processing, but the compression effect largely depended on the data compression

itself, and there would be data loss. Literature[3] describes the conversion of high-dimensional data into low-dimensional data by Deep Auto Encoder (DAE). The idea of DAE is to train the whole model layer by layer in pre-training. Compared with the automatic encoder, DAE is optimized to reconstruct its input with as low reconstruction error as possible. Literature[3] puts forward a new facial expression recognition method using DSAE, which combines geometric features and appearance features to recognize expressions automatically and accurately.

2.3. Deep convolutional neural network

Convolutional Neural Networks (CNN) are generally composed of three processing layers: convolutional layer, pooling layer and full-link layer. The function of the convolution layer is to extract the feature of the image. The pooling layer carries out sparse processing on the feature image to reduce the computational load (dimensionality reduction). In the fully linked layer, the neurons of each layer are connected with all the neurons in the subsequent layer, and whether the neurons are triggered is determined by the sum of the input weights of the connected neurons. The unique local connection and weight sharing of CNN are not found in other neural networks. This makes CNN network less parameters, higher efficiency, better regularization effect and so on. Classical networks based on CNN include AlexNet, GoogLeNet, VGGNet, ResNet,etc.

AlexNet[4] was proposed in 2012 by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton et al. AlexNet has 5 convolution layers and 3 fully connected layers. The structure of the first two AlexNet convolution layers is similar. First, a convolution layer is followed by a ReLU activation function layer, then a pooling layer, and finally an LRN layer. But the parameters of the two convolution layers are different. The last three convolution layers of AlexNet are followed by a ReLU activation function layer, and the fifth convolution layer is followed by a pooling layer. Finally, the output is delivered through three fully connected layers and the ReLU activation function layer and Dropout layer are added between the fully connected layers. This network can deal with the gradient diffusion problem when the network is deep, and has good anti-overfitting ability and generalization ability.

The structure is shown in Figure 2.

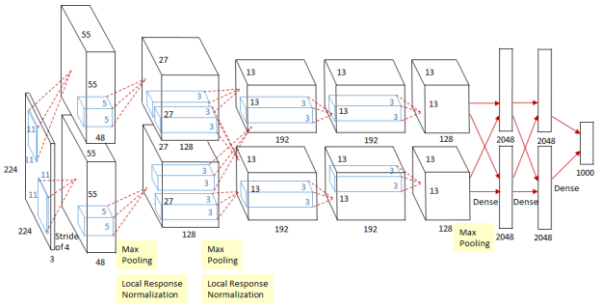

Figure 2. AlexNet structure drawing

The key to GoogLeNet[5] is the Inception module. Inception module is divided into four branches. By convolution of multiple scales at the same time, richer features of different scales can be extracted, which makes the final recognition effect of GoogLeNet more accurate. The first branch is a 1x1 convolution layer; The second branch goes first by a convolution of 1x1 and then by a convolution of 3x3; The third branch is also first convolved by 1x1 and then by a 5x5 convolution; And the fourth branch is first maximized by 3x3 and then convolved by 1x1. Inception module is to achieve multi-scale feature extraction through the convolution and maximum pooling of these three convolution kernels with different sizes, and the addition of three 1x1 convolution is because when the sensitivity field is the same, more abundant features can be extracted by adding convolution. GoogLeNet also designed three loss units and replaced the full connection layer in AlexNet with global averaging pooling to reduce the number of parameters. But the network as a whole is complex. The Inception module structure is shown in Figure 3.

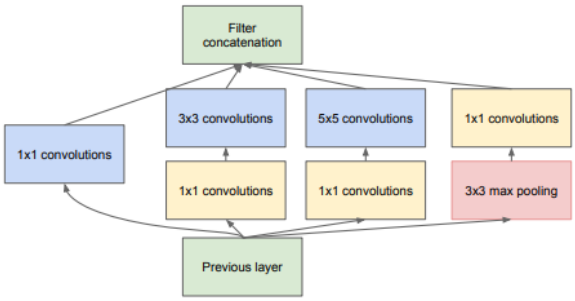

Figure 3. Inception Module structure diagram

VGGNet[6] has five sets of convolution kernels with a size of 3x3, two maximum pooling layers and three fully connected layers. VGGNet has 6 different structures according to different network setup methods. VGGNet has a deeper network structure, which deepens the network to 19 layers. The previous convolutional layers are replaced by convolutional blocks composed of different numbers of convolutional layers, which improves the receptive field of the network.

The key to ResNet[7] (residual neural network) is the residual block. Input x passes through two paths respectively: one first passes through 3x3 convolution to a BN layer and ReLU activation function layer, and then passes through a 3x3 convolution layer, BN layer and ReLU activation function layer; The other way is identity mapping. The results of these two branches are added together, and then a ReLU activation function layer is added to output. The residual module can solve the problem of gradient explosion and disappearance, and simplify the learning objective and difficulty. Through this module ResNet can realize ultra-deep network layer number. ResNet keeps stacking this basic module to get different ResNet models, common ResNet18, ResNet50, ResNet101 and so on. The residual structure is shown in Figure 4.

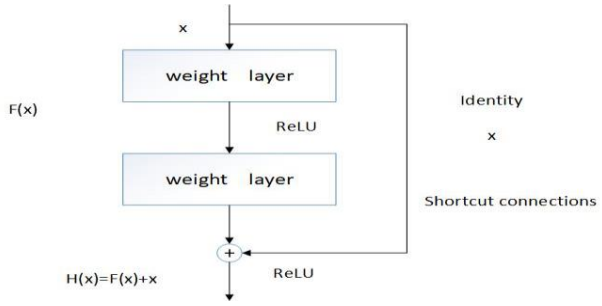

Figure 4. Residual structure diagram

In addition to these networks, there are a number of Pai Sheng frameworks based on them. A localized CNN[8] is used to extract facial expression features. Literature [9] proposed a facial expression recognition method based on Region of Interest (ROI). Wen Xian [10] combines the area of

interest and K-nearest Neighbor (KNN) algorithm to propose an improved ROI-KNN training method, which solves the problem of poor generalization ability of deep neural network models caused by insufficient facial expression training data, thus improving the robustness.

2.4. Generating adversarial network

In 2014, Ian Goodfellow proposed a Generative Adversarial Networks (GAN) using an unsupervised architecture. The generative adversarial network consists of Generator network and Discriminator network, which can achieve better output by making them compete with each other. In literature [11], a production Ad hoc network (DR-GAN) for unwrapping representation learning is proposed. A face with arbitrary posture or even extreme profile can be positively transformed or rotated through a codec structure generator, which has far-reaching significance for the study of facial expression recognition with low robustness in the field.

3. Conclusion

Facial expression recognition technology is becoming more and more mature, but there are still some problems: (1) The types of facial expressions are not rich enough, human expression is more than six basic emotions, which also leads to the poor recognition effect of facial expression recognition in natural and complex scenes. (2) Lack of data sets. Most of the current data sets are collected in the laboratory and other scenes, and these expressions are not natural enough. The types and quantity of data sets are not enough, and deep learning requires a lot of data for training. Rich data can make the trained network have better performance. (3) Although many methods based on deep learning have good recognition performance, their network structure is too complex, requires high hardware requirements, and requires too much computation, which requires a lot of training time.

In the future research, we should develop more new algorithms with higher recognition effect and less cost. The network is improved by using lightweight model, reducing network parameters and computation, greatly reducing the training time, so that the network can be recognized in a more complex environment.

Acknowledgment

This work was supported in part by the 2022 Graduate Innovation Fund Project of Sichuan University of Science and Engineering (Y2022146). The authors express their acknowledgement for the anonymous review.

References

[1] HINTON GE, OSINDE ROS, Teh YW. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[2] LIU M, LI S, SHAN S, et al. Au-inspired deep networks for facial expression feature learning [J]. Neurocomputing,2015, 159: 126-136.

[3] ZENG N, ZHANG H, SONG B, et al. Facial expression recognition via learning deep sparse autoencoders [J]. Neurocomputing, 2018, 273: 643-649.

[4] DHALL A, GOECKE R, Lucey S, et al. Collecting Large, Richly Annotated Facial-Expression Databases from Movies[J]. IEEE Multimedia, 2012, 19(3): 34-41.

[5] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 770-778.

[6] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014, 1409(15): 1-9.

[7] HUANG G, LIU Z, et al. Densely Connected Convolutional Networks[C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:4700-4708.

[8] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[9] CUI R, LIU M, LIU M. Facial expression recognition based on ensemble of mulitple CNNs [C]//Chinese Conference on Biometric Recognition. Springer, Cham, 2016: 511-518.

[10] XIAO S, DING P, FUJI REN, Facial Expression recognition based on ROI-KNN convolutional neural network [J] , 2016, 42(6): 883-891.

[11] TRAN L, YIN X, LIU X. Disentangled representation learning gan for pose-invariant face recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1415-1424.