Research on Facial Expression Recognition Based on Voting Model

IOP Conf. Series: Materials Science and Engineering **646** (2019) 012054 doi:10.1088/1757-899X/646/1/012054

Research on Facial Expression Recognition Based on Voting Model

Yang Fei* **and Guo Jiao**

Hubei University of Technology

*E-mail address: yishanyishui@163.com

Abstract. In order to improve the recognition rate of real-time classification of facial expressions,we proposed a method of facial expression recognition based on voting mechanism. Firstly, different neural network models are constructed to learn facial features. Then, the extracted features are fed into the classifier to obtain the posterior probability of various features. Finally, through the voting mechanism, the optimal decision-making level fusion is achieved to complete the facial expression classification.Experiments show that the average recognition rate of fer2013,CK+ and JAFFE database is 74.58%, 100% and 100% respectively. Compared with other recognition methods, experiental data show that this method has superior performance, improves the recognition rate and robustness of the algorithm, and ensures the universality of the algorithm.

1. Introduction

Facial expression is a common way to convey emotional information in social activities. The human face can be broken down into thousands of different expressions, but the mainstream approach divides the basic facial expressions into seven categories: happy, sad, fear, disgust, surprise, anger and neutral. Facial expressions contain a variety of information, including personal emotional status, personality characteristics and cognitive activities[1]. Moreover, many factors such as personal emotional state, psychological state and physical state are closely related to the information hidden in facial expressions. Mehrabian, a psychologist, suggests that only 7% of human communication is verbal, compared with 55% for facial expressions, which account for the most information. It can be seen that the study of facial expressions can help us get a lot of valuable and diverse information.

With the development of computer vision deep learning technology, great progress has been made in feature extraction of facial expression images[2]. At present, the main feature extraction methods are divided into two categories: feature extraction based on manual method and feature extraction based on machine learning. The main methods of manual feature extraction are LBP(Local Binary Pattern), Gabor and Hog[3]. The manual feature extraction method has good effects for specific data sets. However, the manual extraction method is not universal. When facing different data sets or data sets with a large number of pictures, its accuracy will decline. Feature extraction based on machine learning has greater generality, but it is often not accurate and takes a long time to calculate, which causes a great waste of computing resources and time[4]. When dealing with the data of large samples, the deeper network can achieve a better fitting effect. When dealing with the data of small samples, the shallow network often has a higher precision, while the deep network has a lower precision due to the lack of sufficient sample quantity support.

This paper proposes a voting mechanism model based on neural network, which integrates deep neural network, VGG19 and ResNet18, and retains the advantages of different neural networks for different sample sets, so as to further improve the original network model on the basis of retaining the original accuracy, and achieve good experimental results.

2. Related work

Traditional facial expression recognition methods generally adopt manual method to extract facial expression features and use traditional machine learning methods to determine facial expressions. In literature [5], Local Binary Pattern (LBP) was used to extract expression features to reduce the impact of facial illumination imbalance on expression recognition, and Support Vector Machine was used for expression determination. Zhao et al. [6] further enhanced the robustness of expression feature extraction by fusing LBP on three orthogonal planes, reduced the influence of facial posture and light and other factors, and determined expressions by combining k-nearest neighbors (KNN) and Hidden Markov Models (HMM).Based on the theory of non-negative matrix factorization (NMF), Zhi et al. [7] proposed the graph-preserving non-negative matrix factorization (GSNMF) method to highlight facial features and improve the recognition rate of facial expression features, and the final expression determination was also realized by KNN. In traditional expression recognition research, the efficiency of expression recognition is not considered and the accuracy of expression recognition is low due to the limitations of traditional machine learning classification methods.

As Convolutional Neural Network (CNN) shows excellent performance in computer vision fields such as image segmentation and image classification, more and more researches tend to use Convolutional Neural Network to extract facial features and improve the robustness of facial expression judgment. Kim et al. [8] determined the static expression type by integrating the results of multiple CNN in an exponential weighted decision fusion. Li et al. [9] proposed a new CNN method to maintain depth and locality, aiming to enhance discrimination between expression categories by maintaining local tightness and maximizing the gap between classes. Kample et al. [10], after analyzing algorithm differences and performance effects in multiple literatures, improved the accuracy of expression recognition by constructing cascading CNN. Although the above research improves the recognition accuracy to some extent, the complex network structure and connection mode make the training process quite tedious and difficult to achieve real-time recognition. In order to improve the recognition efficiency, Arriaga et al. [11] greatly simplified the network structure by combining the residual module and the deep separable convolution layer, so that the constructed CNN could realize the real-time recognition effect, but the model could only reach the benchmark accuracy. To sum up, the current research on facial expression recognition cannot take into account the recognition accuracy and efficiency.

3. Framework for the proposed facial expression recognition system

3.1. VGG19 and Resnet18

3.1.1 Model design. Deep convolutional neural network is used to integrate facial expression feature extraction and expression classification into an end-to-end network. VGG19 and Resnet18 were used to identify and classify facial expressions.

VGG network is a kind of classical image classification network[12]. Because it can extract image features, it is also applied to the loss function in the style transfer network. The other is used for Edge Detection of HED (Holistically - Nested Edge Detection) network[13] is developed based on VGG network. The name VGG is derived from the Visual Geometry Group at Oxford University, where the author works.

The Table 1 shows the structure of VGG19 image classification network:

Table 1.VGG19 network

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

VGG19 advantages:

The structure of VGGNet is very simple. The entire network USES the same size convolution kernel size (3x3) and maximum pooling size (2x2).A combination of small filters (3x3) convolution layers is better than a large filter (5x5 or 7x7) convolution layer.It is proved that the performance can be improved by deepening the network structure.

ResNet network[14] is based on VGG19 network, which is modified, and residual element is added through short-circuit mechanism. The change is mainly reflected in ResNet directly using the convolution of stride=2 for down-sampling, and replacing the full connection layer with the global average pool layer. An important design principle of ResNet is: when the size of feature map decreases by half, the number of feature maps increases by twice, which keeps the complexity of network layer.

The Table 2 shows the structure of resnet-18 image classification network:

Table 2. ResNet18 network

Layer name	Output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7.64, stride 2				
conv3_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv4_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv5_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv3_x	7×7	$\begin{bmatrix}3\times3.512\\3\times3.512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3.512\\3\times3.512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	Average pool, 1000-d fc, softmax				
Flops		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

ResNet - 18 advantages:

Compared with the traditional convolutional neural network (VGG), ResNet has the advantages of lower complexity, deeper parameter reduction, simple optimization and deeper classification accuracy.

3.1.2 Loss function design. In the design, we explore two calculation methods of loss function, the first is cross entropy loss function. The model obtained the output probability of each category after the full connection layer, but at this time, the probability was not normalized. We normalized the probability to 1 through a softmax layer, which was easier for data processing. The calculation formula of cross entropy loss function is as follows:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))] \tag{1}$$

In the softmax regression, we solved the multi-classification problem by normalizing the probability size, and class index y could take k different values (instead of 2).

The second kind of loss function is the linear SVM (Support Vector Machine) classification loss, which is abstract as the hinge loss function, and the call of hinge loss is derived from the graph of the loss function, which is a broken line.

$$J(i) = \max(0, 1-h_\theta(x^i)) \tag{2}$$

If properly classified, the loss is 0, otherwise the loss is

$$1-h_\theta(x^i) \tag{3}$$

SVM and Softmax classifiers are the two most commonly used classifiers. The difference is that the SVM outputs each classification score and then selects the class with the highest score. Different from SVM, Softmax classifier is a logistic classifier that is faced with the induction of multiple classifications. Its normalized classification probability is more intuitive, and its probability sum is 1.

3.2 Deep Neural Networks

DNN is deep neural network[15]. The difference between RNN (Recurrent Neural Network) circular neural network and CNN convolutional neural network is that DNN refers to the fully connected neuron structure, which does not include convolution unit or temporal correlation. This paper adopts a two-layer neural network, as shown in the Figure 1 below.

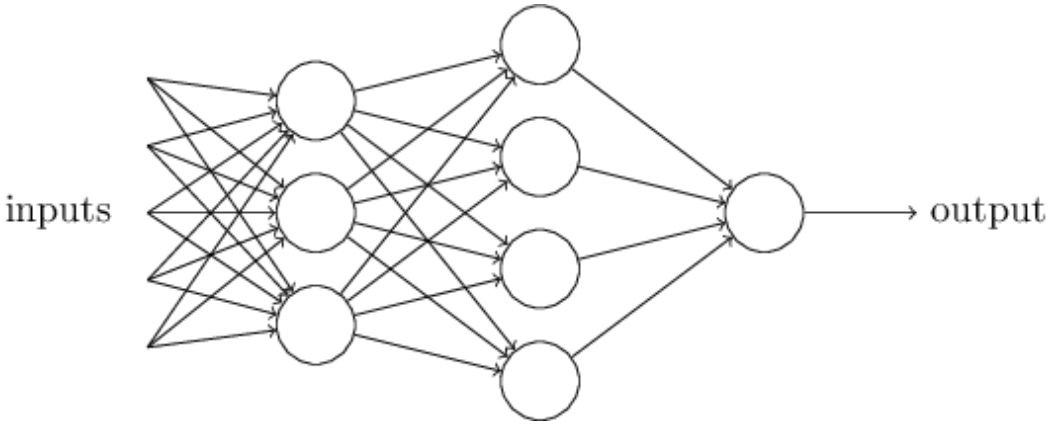

Figure 1. DNN network

3.3 Voting model

We introduce a voting mechanism model(as Figure 2) to integrate the three networks. When the same photo is input into different models, it is possible to get the same or different classification results. We

select three different models mentioned above and choose the classifier which can achieve the highest accuracy. When the classification results of the three models are identical, the classification results are output. When the classification results of two models are different from the other, the classification results of most models are output. When the classification results of the three models are different, the classification results of the model with the highest accuracy are output when a single model is tested. This model retains the advantages of the original model and makes the model universal on different data sets.

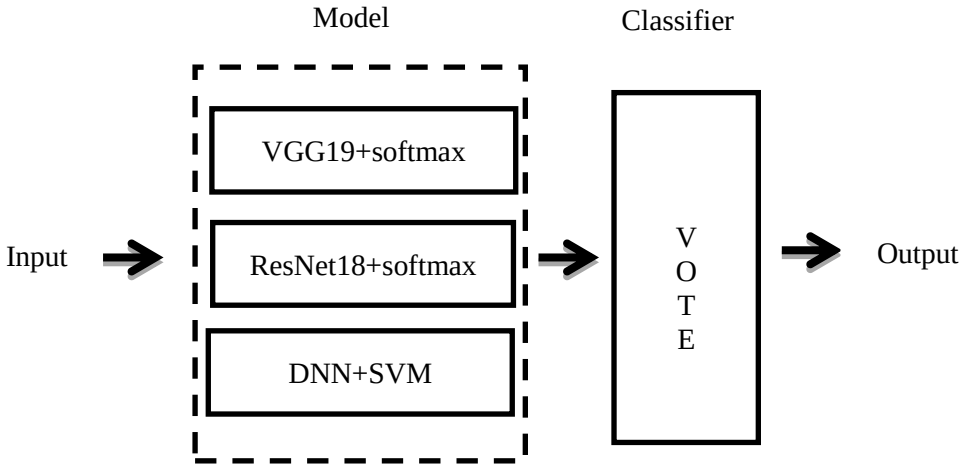

Figure 2. voting model

4. Experiment

4.1 Database

FER2013 data set consists of 28,709 training graphs, 3589 public test graphs and 3589 private test graphs. Each image is a grayscale image with 48 by 48 pixels. FER2013 has seven emojis in its database: angry, disgusted, fearful, happy, sad, surprised and neutral. This database is the data of Kaggle competition in 2013. Since it is mostly downloaded from web crawlers, there are some errors. The artificial accuracy of this database is about 65%.

The CK+ database was released in 2010 as an extension of the cohn-kanade Dataset. This database includes 123 subjects, 593 image sequences, and the last Frame of each image sequence has the label of action units, while 327 of the 593 image sequences contain the label of emotion. This database is obtained under laboratory conditions and is rigorous and reliable. CK+ is a database of comparative standards in facial expression recognition, and many articles use this data for testing.

The JAFFE data set had 213 images. Ten Japanese female students were asked to make seven facial expressions. Seven emotions include: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. (anger, disgust, fear, happiness, sadness, surprise, neutral)

4.2 The experimental results

In order to prevent the network from over-fitting too fast, some artificial image transformations, such as rollover, rotation and cutting, can be done. This is called data enhancement. Another benefit of data manipulation is to expand the amount of data in the database, making the trained network more robust. In this experiment, at the training stage, we randomly cut images of 44*44, and randomly mirror the images, and then send them to the training. In the test phase, this paper takes an integrated approach to reduce outliers. We cut and mirror the images in the upper left, lower left, upper right, lower right and center, which makes the database 10 times larger, and then send these 10 images into the model. Then the probability is averaged, and the maximum output is the corresponding expression, which effectively reduces classification errors.

We first carried out experiments on a single model. To ensure the effectiveness of the experiment, we used 10 times of cross-validation in the experiment. The experimental results are as follows Table 3:

Table 3. experimental results

Model	classifier	Fer2013	CK+	JAFFE
VGG19	SVM	70.772	93.285	91.265
VGG19	softmax	**72.112**	94.646	92.735
ResNet18	SVM	70.549	93.343	89.285
ResNet18	softmax	71.19	94.04	90.707
Deep NN	SVM	53	**100**	**100**
Deep NN	softmax	25	22	**100**
VOTE	VOTE	**74.58**	**100**	**100**

4.3 Discussion

The depth models of VGG and Resnet18 can achieve good classification results on the FER database. And VGG's method is better than Resnet18's. Dropout can effectively reduce overfitting and improve accuracy. The Dropout method is equivalent to randomly deactivating some connections during training and then supplementing them during testing, which is equivalent to integrating several good models to make comprehensive predictions.The 10-fold cutting scheme can further reduce the error rate of identification. For the training phase, random cutting increases the data volume, which is equivalent to directly expanding the data set and slowing down the effect of overfitting. For the prediction stage, 10 times of data can predict the result at the same time, which is equivalent to the integration operation to reduce the network misjudgment.

The classification method of Softmax is better than that of SVM. Softmax takes all the categories into account, and the result of the classification, is the result of the classification in all K classes. However, the SVM method only USES binary classification each time, so it can only learn whether it is the sample, but not necessarily the correct sample, which further increases the difficulty of classification

In the effect of single model, we reached a very high level of VGG19+dropout+10crop+softmax. In the Public test set, it reaches 71.496%, and in the Private test set, it reaches 73.112%, which is also the state-of-the-art level under a single model.Deep network works better for large data sets, while shallow network works better for small data sets.The voting mechanism can retain the advantages of each model and integrate them, increasing the robustness and applicability of the model.Deep network works better for large data sets, while shallow network works better for small data sets.

5. Conclusion

This paper presents a multiple model based on voting mechanism of facial expression recognition method, the neural network model is built with expression classification experiment, respectively by using the voting mechanism to obtain all kinds of model of the time, realize the model of decision fusion, through the principle of the minority is subordinate to the majority, making multiple classifier accuracy maximization. Experimental results in FER2013, CK+ and JAFFE databases show that the method in this paper not only highlights the differences between models, but also optimizes the classification recognition rate and robustness by integrating the advantages of multiple models and optimizing the weight of each feature through voting mechanism. According to the experimental data, the recognition rate of this algorithm is improved compared with other algorithms, and the recognition universality is guaranteed. However, the time performance of the algorithm in this paper is not as good as that of the single-feature construction method because it needs multithreading to conduct three sub-experiments at the same time. In addition, in the process of expression feature fusion, in order to

highlight different features and reduce the uncertainty of decision results, the next step is to introduce parallel processing technology and improve the performance of recognition methods.

References

[1] Gu, W. , Xiang, C. , Venkatesh, Y. V. , Huang, D. , & Lin, H. . (2012). Facial expression recognition using radial encoding of local gabor features and classifier synthesis. Pattern Recognition, 45(1), 80-91.
[2] Carcagn ì, Pierluigi, Del Coco, M. , Leo, M. , & Distante, C. . (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus, 4(1), 645.
[3] Chao, W. L. , Ding, J. J. , & Liu, J. Z. . (2015). Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. Signal Processing,117, 1-10.
[4] Sajjad, M. . (2017). Facial appearance and texture feature-based robust facial expression recognition framework for sentiment knowledge discovery. Cluster Computing, 21(5), 1-19.
[5] Shan C, Gong S, McOwan P W. Facial expression recognition based on local binary patterns: A comprehensive study[J]. Image and vision Computing, 2009, 27(6): 803-816.
[6] Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007 (6): 915-928.
[7] Zhi R, Flierl M, Ruan Q, et al. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2011, 41(1): 38-52.
[8] Kim B K, Lee H, Roh J, et al. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 427-434.
[9] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2852- 2861.
[10] Pramerdorfer C, Kampel M. Facial expression recognition using convolutional neural networks: state of the art[J]. arXiv preprint arXiv:1612.02903, 2016.
[11] Arriaga O, Valdenegro-Toro M, Pl öger P. Real-time convolutional neural networks for emotion and gender classification[J]. arXiv preprint arXiv:1710.07557, 2017.
[12] Goodfellow I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[C]//International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2013: 117-124.
[13] Bulat A, Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem(and a dataset of 230,000 3d facial landmarks)[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1021-1030.
[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
[15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic Segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.