Multilayer Convolution Sparse Coding for Expression Recognition

Shuda Chen*, Yan Wu
The College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, No.100, Haisi Road, Shanghai 200234, China

*chenshuda2020@foxmail.com

Abstract. Facial expression recognition is widely used in various research fields. For facial expression recognition problems, deep neural network methods have a complex structure and poor interpretability, while traditional machine learning methods have less plentiful diverse features and low recognition rates. Therefore, a new Multilayer Convolution Sparse Coding (MCSC) method is proposed for facial expression recognition. The MCSC method deeply extracts the salient features of the human face through a convolutional neural network. Furthermore, it uses a multilayer sparse coding to learn layer by layer to recognize different facial expression features based on sparse coding, which improves the recognition accuracy of facial expressions. Finally, the MCSC method was validated on three public facial expression datasets, i.e. JAFFE, CK+, and Fer2013. We also compared and analyzed 5 feature extraction approaches. The results show that MCSC has the best facial expression recognition performance in the comparison algorithm. Its accuracies of the three data sets reach to 90.8%, 98.2%, and 72.4%, respectively.

Keywords: Facial expression recognition, Multilayer Sparse coding, Feature extraction, Convolutional neural network.

1. Introduction

Facial expressions can decently express human psychological activities [1], and have become one of the important research topics in the field of computer. Among them, the common facial feature extraction methods include Local Binary Pattern [2,3], Histogram of Oriented Gradient [4], and scale-invariant feature transform [5]. Traditional machine learning methods, feature extraction, to a certain extent, destroy the integrity of the original image and lack the overall image diversity. In addition, feature extraction is relatively complex and the recognition effect is not as good as that based on the convolutional neural network method in the case of occlusion and shadow [6].

In order to overcome the shortcomings of deep neural networks and traditional machine learning methods, a decision tree-based integration method Deep Forest [7] and Deep Cascade Model [8] were proposed in recent years. A deep neural network requires more overshoot parameters and relies on...
large samples. Compared with DForest, it has fewer parameters and is suitable for small sample learning [7], such as gesture recognition [9], image analysis [10], and gene sequence analysis [11,12]. Therefore, this paper proposes a multi-layer Convolution Sparse Classification (MCSC). This method draws on the representation learning of deep neural networks and uses a sparse matrix as the basis for the structure recognition feature, which mainly relies on the layer-by-layer processing of original features to improve facial expression recognition accuracy.

In this paper, the methods of facial expression recognition are reviewed. First, the principle of multi-layer convolution sparse coding is described. Then, three facial expression open data sets are analyzed, and five feature extraction methods and the algorithm proposed are compared in this paper. Finally, the experimental results, future challenges, and opportunities have been summarized in this field.

2. Multilayer convolution sparse coding method

Most facial expression recognition methods based on convolutional neural networks use softmax layer targets to learn lower-level parameters [13,14]. In this paper, multi-layer sparse coding is used instead of the softmax layer in a deep neural network for classification. Stochastic gradient descent and multi-classification cross-entropy loss function are used to train the network, and the extracted feature vectors are input into the multilayer sparse coding to obtain the final result. Therefore, this method is divided into two parts. First, CNN extracts facial expression features; second, it adopts a multi-layer sparse classification method. The overall block diagram is shown in Figure 1.

Multilayer convolution sparse coding method is a kind of end-to-end structure, set the input images \( I_{x,y} \) ( \( x, y \) for the height and width of the picture, \( N \) for training set number) after the facial cutting form \( (N,W,H,L) \) input format, \( W \) as the width, \( H \) as the height, \( L \) as the channel number, through the CNN network to get a characteristic vector \( f(P_1,\ldots,P_c) \), then through the \( n \) hierarchy structure of sparse coding has been, finally \( \text{FinalP} = \text{Max}(\text{Ave}(f^*)) \) get the final forecast \( P \).

![Figure 1 Multilayer convolution sparse coding structure diagram](image)

2.1. Convolutional neural network

The structure of CNN combines the idea of Compact CNN[15] and Inception network [16] in this paper. The overall structure consists of 5 convolutional layers, 2 maximum pooling layers, and 2 full...
connection layers. First of all, the first convolutional layer has a convolution kernel of 32 filters with 1x1. The convolution kernel with 1x1 is mainly used to reduce dimension, add nonlinear operations, and increase network depth. Secondly, there are 2048 and 1024 output neurons in the full connection layer respectively, and dropout is set as 0.5, which is helpful to solve the problem of overfitting. The model consists of two convolution blocks and two pooling layers, followed by two full connection layers. PReLU is used as the activation function, and since PReLU only increases a very small number of parameters, the risk of computation and overfitting is reduced. The specific structure of CNN is shown in Figure 2.

The input image in this model first passes through the convolutional layer formula as follows.

\[ F_I^k = (I_{xy} \ast K^k_I) \]  

(1)

Where \( K^k_I \) is the convolution kernel, which \( F_I^k \) is obtained through three convolutional layers. In addition, the selection of activation function in the convolutional layer is also very critical. The selection of appropriate activation function can accelerate the learning process. The definition of the activation function is as follows. Where \( f_A(\cdot) \) the nonlinear transformation of the added K layer returns the output \( T_I^k \), the PReLU activation function is used in this article.

\[ T_I^k = f_A(F_I^k) \]  

(2)

Secondly, the pooling layer is given by the following formula.

\[ Z_I = f_p(F_{x,y}^l) \]  

(3)

The above formula shows the pooling operation, Where \( Z_I \) represents the output characteristics of the layer \( I \), \( F_{x,y}^l \) represents the input characteristics of the layer \( I \), and \( f_p(\cdot) \) defines the type of pooled operation. In this paper, the maximum pooling is selected, followed by two convolutional layers and the maximum pooling layer, and finally two full connection layers. In order to prevent overfitting, the Dropout rate is selected as 0.5. The two-layer fully connected layer will convert the final pooled output two-dimensional feature map \( Z_I \) into a \( f_s(P_1, ..., P_n) \) one-dimensional vector.

![Fig.2 CNN network structure](image)

2.2. Multilayer sparse coding structure

2.2.1. Sparse matrix
We let \( X = \{ x_n \}_{n=1}^N \) represent the training image, where \( x_n \in \mathbb{R}^{p \times q} \). We let \( y \in \mathbb{R}^{p \times q} \) represent the image size. The number of classes is \( C \), and the number of training images is \( N \). We use \( H^l_X \) to represent the training images, and \( H^l_y \) represent the testing image in level \( l \). \( W^l_y \) represents the coding coefficient.

According to the sparse matrix, we describe the sparse coding problem of samples of \( l \)-level query as follows:

\[
\min_{W^l_y} \| H^l_y - H^l_X W^l_y \|_2^2 + \lambda \| W^l_y \|_1
\]  

(4)

Where \( \lambda \) is the regularization parameter. In recent years, different solving algorithms have been proposed for sparse representation. In particular, the Alternating Direction Multiplier Method proposed in the 1970s [17] has attracted widespread attention. When Yang and Zhang [18] solved the \( l_1 \)-norm minimization problem, the near-end method was integrated into the ADMM. In this paper, the ADMM method was also used to solve the sparse representation problem.

In general, based on ADMM, we introduce an auxiliary variable \( Z \), which is \( Z^l_y = W^l_y \). Then, the augmented Lagrange function of problem (4) can be expressed as

\[
L_\mu(W^l_y, Z^l_y, \Lambda^l_y) = \min_{W^l_y, Z^l_y, \Lambda^l_y} \| H^l_y - H^l_X W^l_y \|_2^2 + \lambda \| Z^l_y \|_1 + \langle \Lambda^l_y, W^l_y - Z^l_y \rangle + \frac{\mu}{2} \| W^l_y - Z^l_y \|_2^2
\]

(5)

Where \( < P, Q > = tr(P^T Q) \), \( \Lambda^l_y \) is a Lagrange multiplier, and \( \mu \) is a scalar constant. At each iteration, the augmented Lagrange function minimizes only one coordinate direction. Specifically, the ADMM consists of the following iterations.

(i) Given \( Z^l_y = Z^{l(i)}_y, \Lambda^l_y = \Lambda^{l(i)}_y \), update \( W^l_y \) by

\[
W^{l(i+1)}_y = \arg \min_{W^l_y} L_\mu(W^l_y, Z^{l(i)}_y, \Lambda^{l(i)}_y)
\]

(6)

(ii) Given \( W^{l(i)}_y, Z^{l(i)}_y, \Lambda^{l(i)}_y \), update \( Z^l_y \) by

\[
Z^{l(i+1)}_y = \arg \min_{Z^l_y} L_\mu(W^{l(i)}_y, Z^l_y, \Lambda^{l(i)}_y)
\]

(7)

(iii) Given \( W^{l(i)}_y, Z^{l(i)}_y, \Lambda^{l(i)}_y \), update \( \Lambda^l_y \) by

\[
\Lambda^{l(i+1)}_y = \Lambda^{l(i)}_y + \mu(W^{l(i+1)}_y - Z^{l(i)}_y)
\]

(8)

The key steps are to solve the optimization problems in Eqs. (6) and (7). Based on the augmented Lagrangian function in the equation (5) (6) can be expressed as

\[
W^{l(i+1)}_y = (H^l_X^T H^l_X + \mu I)^{-1} (H^l_X^T H^l_y - \Lambda^{l(i)}_y + \mu Z^{l(i)}_y)
\]

(9)
Among them is an identity matrix. Based on the augmented Lagrangian function in Eqs. (5) and (7) is written as

$$Z_y^{l(t+1)} = \arg \min_{Z_y} (\lambda \|Z_y^l\|_1 + <A_y^l, W_y^l - Z_y^l> + \frac{\mu}{2} \|W_y^l - Z_y^l\|_2^2)$$ (10)

Since the 1-norm problem is convex but not differentiable at the zero point, the shrinkage technique [44] is used to solve this problem. Therefore, the optimal solution can be expressed as

$$Z_y^{l(t+1)} = \text{shrinkage}_{\mu} \left( \frac{A_y^l}{\mu} + \frac{W_y^{l(t+1)}}{\mu} \right)$$ (11)

After solving the representation coefficient $W_y^l$, the Sparse Matrix Coding (SMC) can be used to obtain the multi-layer Softmax vector $S_{y_{(l)}}^I \in R^{C \times l}$.

In the ADMM algorithm, the objective function will converge when certain optimal conditions and stopping criteria are met. This article sets the maximum number of iterations.

2.2.2. Multilayer sparse coding structure

Specifically, after CNN extracts the features, a feature vector $d_y^1$ is obtained as the input of the first layer in DCM. Then, input $d_y^1$ into the first layer sparse matrix the process, calculate the Softmax vector $S_y^1$, and then connect with $d_y^1$ to construct the input sample $d_y^2 = [(d_y^1)^T, (s_y^1)^T]^T$ of the second layer. Similarly, each column in $D_x^1$ is input to the first-level sparse matrix process to calculate the Softmax vector set $S_{(l)}^1$ connected to $D_x^1$, which to construct the second-level input dictionary $D_x^2 = [(D_x^1)^Y, (S_{(l)}^1)^Y]^T$. Similarly, the feature vector $d_y^K$ of the test sample $y$ and the training dictionary $D_x^K$ of $X$ in the $K^{th}(k = 1, ..., K)$ layer can be calculated. Finally, input $d_y^{(K-1)}$ and $D_x^{(K-1)}$ into $k - 1$ level sparse matrix to obtain the final Softmax vector $d_y^K$ of the test sample $y$ and the final Softmax vector set $D_y^K$ of the training set $K$. When identifying $y$, the label is determined to be the class related to the maximum value in the final Softmax vector, so there is

$$\text{label}(y) = \arg \max_i s_y^K$$ (12)

**Algorithm 1** The solving multi-layer sparse coding structure

| Input | $d_y^1$, $D_x^1$ | The trade-off parameters $\lambda_2$ and $\mu_2$ in sparse coding, the number of layers $K$ in DCM |
|-------|------------------|--------------------------------------------------|
| Output | Predict the label of the test sample $y$ |
| 1. Initialize | $k = 1$ |
2: repeat
3: Compute $S^k_y = SMC_{D_y^k}(d^k_y)$ using Eq.4
4: Compute $S^k_X = [SMC_{D_y^k}(D^k_y), ..., SMC_{D_y^k}(D^k_y)]$ using Eq.4
5: Update $d^{k+1}_y = [(d^k_y)^T, (s^k_y)^T]^T$
6: Update $D^{k+1}_X = [(D^k_X)^T, (s^k_X)^T]^T$
7: $k = k + 1$
8: Until $k > K$
9: Finally, use the largest index using Eq.(12) to find the largest $s^{(K)}_y$, and the displayed label $y$

3. Experiments

3.1. Dataset

In order to verify the experiment, this article uses three public facial expression datasets. The Japanese Female Facial Expression (JAFFE) [19], this expression library has only 213 facial expression images, and each image has 256x256 pixels. The expression library contains a total of 10 Japanese women, each with 7 expressions, which are happiness, sadness, surprise, angry, disgusted, fear, and neutral.

The Cohn Kanade (CK+) [20] database is the most extensive laboratory control database used to evaluate FER systems. CK+ contains 593 video sequences from 123 subjects. The sequence change lasts from 10 frames to 60 frames and shows the transition from neutral facial expressions to peak expressions. Based on facial actions are divided into anger, contempt, disgust, fear, happiness, sadness, and surprise.

The Fer2013 database has a total of 35887 face images in Kaggle facial expression recognition challenge[21], which have 28709 images are used for training, 3589 images are used for verification set, and 3589 images are used for the test set. The pixel size of each face is 48x48 images. There are 7 facial expressions including anger, disgust, fear, happiness, sadness, surprise, and neutrality.

3.2. Evaluation standard

The evaluation standard use five indicators, accuracy for facial expression classification, sensitivity, specificity, negative predictive value, and accuracy to measure the classification results. First, True Positives (TP, the positive class is judged as positive), False Positives (FP, Negative category is judged as the positive category), False Negatives (FN, positive category is judged as the negative category), True Negatives (TN, negative category is judged as the negative category).

1. Accuracy: The proportion of all predictions that are correct.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}
\]
2. Precision (Precision/TPR): The probability of a sample that is actually positive among all samples that are predicted to be positive.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{14}
\]

3. Recall rate (Recall): which also known as recall rate, that is, the proportion of the correct prediction that is positive to all that is actually positive.

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{15}
\]

4. F1 value (F1-score): which is the arithmetic average divided by the geometric average, and the larger the better, when the F1 value is small. True Positive will increase relatively, False will decrease, and Recall will increase.

\[
F_{1} - \text{score} = \frac{2TP}{2TP + FP + FN} \tag{16}
\]

3.3. Analysis of results

In order to comprehensively compare the special methods of extracting expressions, the feature extraction method uses traditional feature extraction methods such as LBP, HOG, SIFT, and deep extraction feature methods. Multi-granularity scanning (Mg_scanning) and CNN extraction features (CNN_feature) are compared. The method suitable for facial expression recognition finally compares the deep forest (DForest) and CNN methods.

First, we analyzed the three data sets. JAFFE, CK+, and Fer2013 have successively increasing scales. JAFFE has 213 pictures, CK+ has 981 pictures, and Fer2013 has 10,000 pictures. Analyze facial expression problems from traditional machine learning methods and deep neural network methods respectively.

There are a total of 213 samples in the JAFFE data set, and 70 test samples are selected, including 10 test sets for each of 7 expressions. As shown in Figure 3, there are a total of 4 evaluation indicators, a total of 5 characteristics. First, The highest accuracy rate of CNN_feature is 90.8%, higher than Mg_scanning 88.6%, but the precision rate, recall rate, and F1 index of Mg_scanning are all higher than CNN_feature (89.9%,88.6%, and 88.5%, respectively). SIFT features performed worst with 77.14%, 69.03%, 67.50% and 67.95%, respectively.

Since CK+ did not provide the specified training, validation, and test sets, the algorithms evaluated in this database are inconsistent. For static-based methods, the most common data selection method is to extract the peak formation of the last one to three frames and the first frame. Select 45% of 442 pictures as the test set. The experimental results are shown in Figure 4. The four evaluation indicators of the CK+ data set are the highest for CNN_Feature. The accuracy rate is 98.2%, the precision rate is 86.6%, the recall rate is 86.2%, and the F1-score is 86.5%, followed by HOG features.

The Fer2013 data set has given the test set and training set. FER2013 contains 28709 training sets, 3589 verification sets, and 3589 test sets. The experimental results are shown in Figure 5. In the case of a poor overall recognition rate, CNN extracts features still have decent results. The accuracy of the four indicators is 72.4%, accuracy is 66.1%, recall rate is 64.8%, and F1-score is 65.3%. The second is the HOG feature, the worst is MG_scanning.
The multilayer convolutional sparse coding method (MCSC) is compared with the deep forest (DForest) and CNN. Among them, the accuracy of MCSC has been improved on the three public data sets, as shown in Table 1.

| Method | JAFFE Acc | Pre | Rec | F1 | CK+ Acc | Pre | Rec | F1 | Fer2013 Acc | Pre | Rec | F1 |
|--------|-----------|-----|-----|----|--------|-----|-----|----|------------|-----|-----|----|
| DForest | 88.6 | 89.9 | 88.6 | 88.5 | 97.7 | 84.8 | 84.6 | 84.6 | 48.8 | 47.8 | 37.7 | 39.1 |
| CNN | 81.4 | 73.4 | 71.3 | 71.4 | 92.8 | 79.6 | 77.2 | 78.2 | 61.0 | 68.6 | 67.9 | 68.1 |
| MCSC | 90.8 | 80.2 | 79.3 | 79.6 | 98.2 | 86.6 | 86.2 | 86.5 | 72.4 | 66.1 | 64.8 | 65.3 |

Table 1: Comparison of three classification results

Fig. 3 JAFFE dataset experimental results  Fig. 4 CK+ dataset experimental results
4. Conclusion

This paper proposes a new multilayer convolution sparse coding method for facial expression recognition. The experimental results on three public facial expression data sets have reached 90.8%, 98.2%, and 72.4% accuracy. The results show that MCSC has the best facial expression recognition performance among the compared algorithms. The five types of facial expression extraction features are compared and analyzed. In addition, we compared with traditional feature extraction and multi-granularity scanning, the CNN has an average image feature extraction of 83%, which is much higher than other feature extraction methods.

Facial expression recognition has made great progress in modern theory and technology, and there are still many problems. In future work, we will try to apply the method to video data for facial expression recognition, which will not only improve the accuracy, which can reduce model parameters and training complexity too.

References

[1] Cheng S, Zhou G. Facial Expression Recognition Method Based on Improved VGG Convolutional Neural Network[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(07):265-283.

[2] Yuan L, Ling Z, Yunhua C, Bi Z and Wenchao J. An expression recognition algorithm for extracting features of salient areas of reverse cooperation [J]. Small Microcomputer System, 2018, 39 (07): 1585-1589.

[3] 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

[4] Zhou H, Lin D J, Huang T S. Static Hand Gesture Recognition based on Local Orientation Histogram Feature Distribution Model[C]//Computer Vision and Pattern Recognition Workshop, 2004. CVPRW ’04. Conference on. IEEE Xplore,2004.

[5] David G. Lowe. Distinctive Image Features from Scale Invariant Keypoints[J]. International Journal of Computer Vision, 60(2):91-110.

[6] Samadiani N, Guangyan H, et al. A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data.[J]. Sensors (Basel, Switzerland), 2019, 19(8).

[7] Zhou Z H, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017.

[8] Zhang L, Liu J, Zhang B, et al. Deep Cascade Model-Based Face Recognition: When Deep-Layered Learning Meets Small Data[J]. IEEE Transactions on Image Processing, 2019, 29:1016-
1029.

[9] Zhao J, Mao J, Wang G, et al. A miniaturized wearable wireless hand gesture recognition system employing deep-forest classifier[C]// 2017 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, 2017.

[10] Cai; X Y W L. Deep Forest-Based Classification of Hyperspectral Images[C]// 37th Chinese Control Conference. 2018

[11] Guo Y, Liu S, Li Z, et al. Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data[C]// 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017.

[12] Guo Y, Liu S, Li Z, et al. BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data[J]. BMC Bioinformatics, 2018, 19(S5).

[13] Li S, Deng W. Deep Facial Expression Recognition: A Survey[J]. IEEE Transactions on Affective Computing, 2018, PP(99).

[14] Haritha C V, Thulasidharan P P. Multimodal Emotion Recognition using Deep Neural Network-A Survey[J]. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEER-ING, 2018, 06(6):95-98.

[15] Kuo C M, Lai S H, Sarkis M. A Compact Deep Learning Model for Robust Facial Expression Recognition[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018.

[16] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. 2014.

[17] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation”, Computers & Mathematics with Applications, vol.2, no.1, pp. 1740, 1976

[18] J. Yang and Y. Zhang, “Alternating direction algorithms for 1-problems in compressive sensing,” Siam Journal on Scientific Computing, vol. 33, no. 1, pp. 250–278, 2009.

[19] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998, pp. 200–205.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Embadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Con-ference on. IEEE, 2010, pp. 94–101.

[21] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al., “Challenges in representation learning: A report on three machine learning contests,” in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124