Dynamic Expression Recognition Based on Hybrid Features and Optimized Extreme Learning Machine Model

Xiujuan Sui, Lei Xue and Dunyu Li

Department of Communication and Information System, School of Communication and Information Engineering, Shanghai University, 99 Shangda Road Baoshan District, Shanghai, 201900, China.
Email: suixiujuanas@163.com

Abstract. Facial expression recognition is an academic subject, and it can be applied not only in the field of computer vision but also in psychology and sociology. This paper proposed a robust dynamic facial expression recognition method which using hybrid features and the PSO-ELM (Particle Swarm Optimization-Extreme Learning Machine) model. At first, six PSO-ELM models were trained on the training set of six basic expressions, respectively. Then by extracting the hybrid features which fusing geometric features and texture features, the dynamic expressions were recognized by PSO-ELM models. These experiments were performed on the Cohn-Kanade+ face database. The results show that this method has a better recognition effect on dynamic facial expression recognition.

1. Introduction

In recent years, with the improvements of computer computing speed and performance, the relationship between humans and computers has become more closely, and intelligence has penetrated into every corner of life. With the development of artificial intelligence, facial expression recognition is receiving more and more attention. Facial expression recognition involves many researching fields such as pattern recognition, image processing, computer vision, cognitive psychology, etc. Facial expressions can present a lot of emotional and psychological information, and are the main carrier of human emotions. American psychologist Mehrabian [1] believes that emotions are composed of 7% language, 38% voice and 55% facial expression.

Expression recognition generally consists of three steps [2]: face detection, facial expression feature extraction and expression classification. And the facial expression feature extraction is the most important part of the entire recognition process.

Face detection technology, mainly detecting whether the images input include faces, has matured in recent years. Face detection methods are roughly divided into two categories: statistics-based methods and structure-based methods. Specifically, these methods include binary wavelet transforms, template matching, local features matching, etc.

Expression feature extraction is to extract useful facial expression information in face images. At present, the main facial expression feature extraction methods can be roughly divided into two categories: geometric features and texture features. The method about geometric features is to extract the face contour points and the local feature points which mainly around the eyes, eyebrow, nose and mouth. Then using the geometric relationship of feature points to describe facial expression features. The main geometric feature extraction algorithms are based on ASM (the Active Shape Model) [3] or AAM (the Active Appearance Model) [4]. The Local Binary Features (LBF) method was proposed by Shaoqing Ren, Jian Sun et al. [5] in 2014. It uses the cascade and shape regression to locate and track the feature points of human faces. Texture features can represent the subtle changes in facial parts such as eyes, eyebrows, corners of mouth, etc. Common algorithms of texture features include Gabor wavelet transform [6], local binary pattern (LBP)[7] and gray level co-occurrence matrix (GLCM). Overall, the geometric features can reflect the dynamic changes of facial expressions, while the texture features focus on representing local detail changes.

The current methods of expression classification mainly use machine learning methods, such as support vector machine (SVM) [8], dynamic Bayesian network (DBN) [9], artificial neural network [10], BP neural network [11] and so on.

In this paper, we performed experiments on the Cohn-Kanade+ facial expression database. Fusing geometric features and texture features, a dynamic expression recognition method based on PSO-ELM is proposed. The geometric features are the coordinate difference of facial feature points in the normalized image sequence of the same group. The texture features are the difference of the local texture information corresponding to the feature points. Given six different expression sequences, six PSO-ELM models are obtained through training. By comparing the similarity between the predicted sequence and the real sequence, the expression category of the given sequence can be discriminated. The complete facial expression recognition process is showed in figure 1.

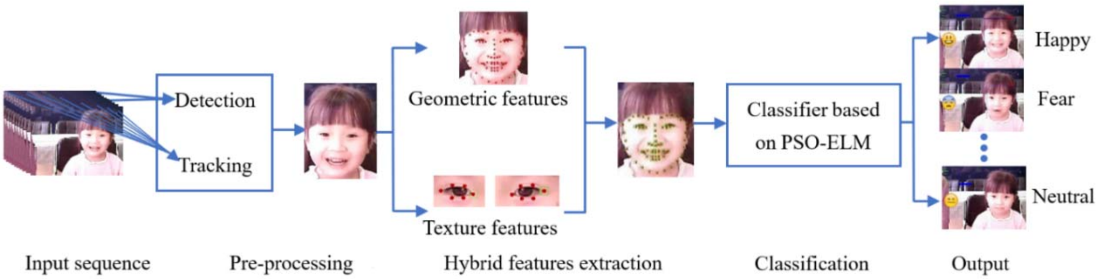

Figure 1. Expression recognition flowchart.

2. Feature Extraction

In this paper, LBF method is used to track and locate feature points of facial expression images in image sequences. Then between the normalized facial expression image and the neutral image, the geometric feature is obtained by the coordinate difference of the feature points, and the texture feature is obtained by the feature difference of the texture region corresponding the feature points.

2.1. Geometric Feature Extraction

As mentioned earlier, expression recognition generally consists of three steps, and the first is face detection. Face detection is a very active direction. Skin-based face detection [12] and Adaboost-based face detection [13] have achieved good results. In recent years, some methods based on deep learning also perform well, especially in complex scenes. These methods have high detection speed and accuracy. In this paper, we use the deep learning algorithm MTCNN [14] to detect human faces, then extract and track facial feature points by LBF method in the detected face frame and entire image sequence. The feature point positioning results in the sequence are shown in Figure 2.

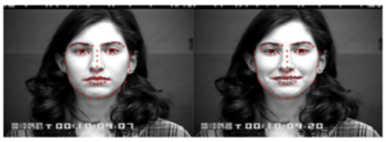

Figure 2. Geometric feature points positioning results in expression sequence.

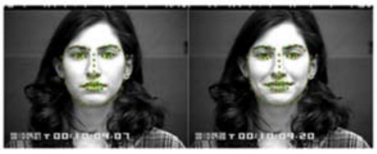

Figure 3. Texture features area corresponding to geometric feature points.

After finding the position of the feature points, the baseline connecting the centres of the two eyes is converted to 0° orientation, and its length is standardized to 1 unit. Then all the faces are normalized by conversion and standard. The x and y coordinates of 68 feature points are calculated as a 68×2-dimensional feature vector.

2.2. Texture Feature Extraction

The texture feature is the facial detail feature of the local area corresponding to the feature point. Here, the texture difference between the neutral expression image and six other expression images is taken as texture features. The calculation of texture features is greatly affected by environmental factors, like uneven or too strong illumination. Gray level co-occurrence matrix is defined as the joint probability distribution of pixel pairs. It is a symmetric matrix. It not only reflects the comprehensive information of image grayscale in directions, adjacent intervals, and amplitude, but also reflects the position distribution characteristics between the same gray level pixels. It is not sensitive to light. The texture analysis method based on GLCM studies the spatial dependence of gray level in image texture with low computational complexity. Due to the robustness of the normalized cross-correlation coefficients, the gradient value difference between the neighborhood of the same feature point corresponding to the neutral image and the expression image can be defined as correlation:

$$\mathrm{Corr} = \left[\sum_{f_a}\sum_{f_b}\big((a,b)P(\mathrm{a},b)\big)-\mu_x\mu_y\right]\Big/_{\sigma_x\sigma_y} \tag{1}$$

where $\mu_x,\ \mu_y$ are variances, $\sigma_x,\ \sigma_y$ are standard deviations. The size of the correlation value reflects the local gray correlation in the image. When the values of matrix elements are evenly equal, the correlation value is large. Corresponding to the 68 feature points extracted by the geometric features, 68 texture features can be obtained for each image. The 68x2-dimensional geometric features and the 68-dimensional texture features are fusing to obtain a total of 204-dimensional hybrid features. Each image can be represented by the 204-dimensional features. Figure 3 shows the texture feature area corresponding to geometric feature point.

2.3. PCA Dimensionality Reduction

Hybrid features have a total of 204 dimensions. The higher the dimension, the more the feature quantity. With large amount of calculation, it is easy to fall into “curse of dimensionality”. In addition, as the dimension increases, the sparseness of the data becomes higher and higher. It is more difficult to explore the same data set in a high-dimensional vector space than to explore in the same sparse data set. So that we need educe the redundant information in the original data by PCA dimension reduction, and retain the most representative feature vector. The feature vector after dimension reduction can better represent the facial expression image.

3. Expression Classification

In this section we use a PSO-ELM (Particle Swarm Optimization-Extreme Learning Machine) model to perform the final expression classification.

3.1. PSO-ELM model

The Extreme Learning Machine (ELM) [15] is an algorithm proposed by Guangbin Huang for solving a single-layer neural network. The algorithm will randomly generate the deviation b_j of the hidden neuron and the linking weight ω_j between the input layer and the hidden layer. And the number of neurons in the hidden layer needs to be set. The output of ELM is:

$$\mathrm{Output} = \sum_{j=1}^{L}\beta_j g\big(w_j.x_i + b_j\big) = O_j \tag{2}$$

where, β_j is the output weight of the jth hidden layer neuron; g ( ) is the output function of the jth implicit neuron; L is the number of implicit nodes. Learn the extreme learning machine by minimizing $\sum_j^N ||O_j - y_j||$. The traditional ELM method has fast learning speed and good generalization performance, but the input weights and deviation are set randomly. This may cause the parameter allocation of the hidden nodes to be uneven, which leads to poor robustness on classification problems.

To solve the problem of uneven allocation of hidden node parameters in the input layer matrix and the implicit layer matrix, we use Particle Swarm Optimization algorithm to optimize them. Then we can obtain the optimal network structure. The learning process of PSO-ELM model is as follows:

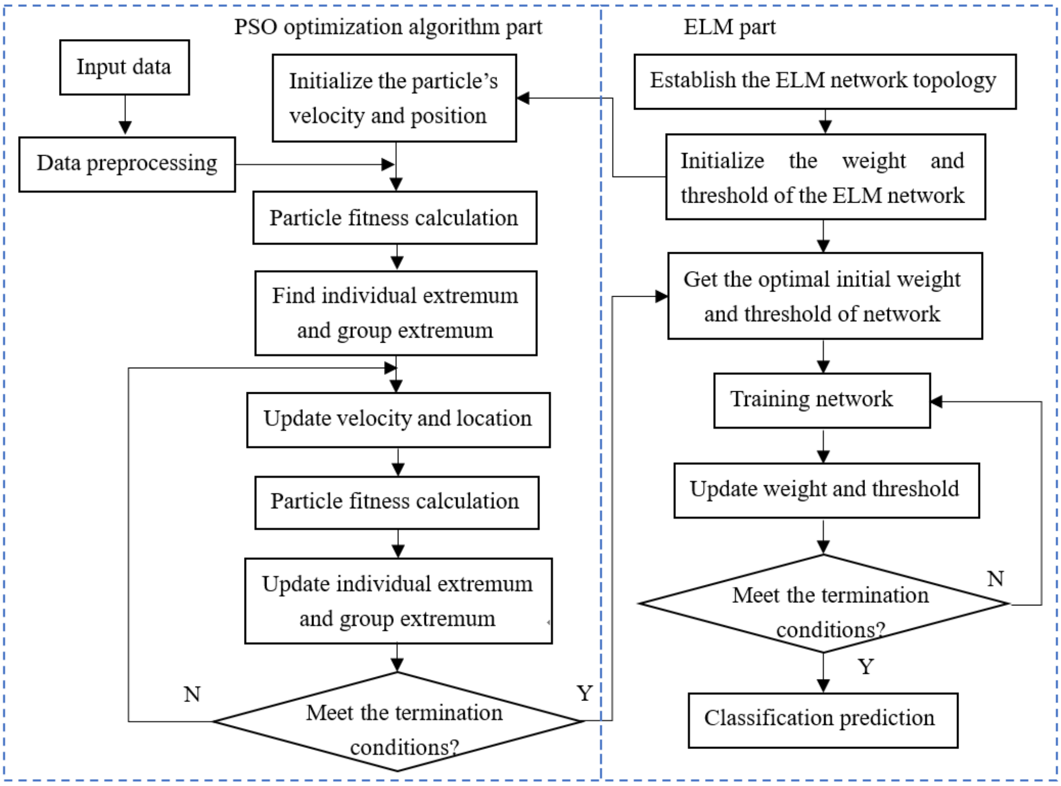
PSO optimization algorithm part
Input data
Data preprocessing
Initialize the particle's velocity and position
Particle fitness calculation
Find individual extremum and group extremum
Update velocity and location
Particle fitness calculation
Update individual extremum and group extremum
Meet the termination conditions?
N
Y
ELM part
Establish the ELM network topology
Initialize the weight and threshold of the ELM network
Get the optimal initial weight and threshold of network
Training network
Update weight and threshold
Meet the termination conditions?
N
Y
Classification prediction

Figure 4. PSO-ELM model flowchart.

3.2. Dynamic Expression Recognition

Facial expression is a dynamic process, all expressions are from the first calm state to the most expressive state and then returning to calm state. This process sometimes is fast, sometimes slow. The expression at each moment is related to the facial muscle movement at the adjacent moment. Six PSO-ELM models correspond to six basic facial expressions. Given a set of facial expression sequences, N frames of images are selected and normalized to make up a Sequence I. Sequence I includes vectors $V_1, V_2, \ldots, V_N$, where V_i, $i = 1,2,\ldots,N$ and V_i is a feature vector of each frame image.

4. Experimental Results

These experiments were performed on the Cohn-Kanade+ (CK+) database [16], which is one of famous expression databases that are widely used to verify facial expression algorithms. The database consists of 593 image sequences and 123 human objects. Each sequence contains a series of dynamic facial expression images. In this database, we randomly select three-fourths of the expression sequence samples as the training set, and the remaining one-fourth as the test set. Using the cross-validation to evaluate the proposed algorithm. Figure 5 shows some samples of CK+ database.

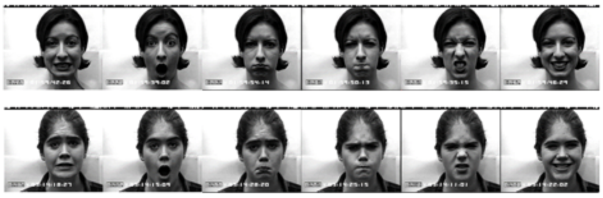

Figure 5. Samples of CK+ database.

After numerous experiments we have found that when the number of hidden nodes is between 50 and 300, the average recognition rate is constantly increasing; from 300 to 400, the average recognition rate shows a downward trend; between 400 and 500, the average recognition rate rises and tends to be stabilized, but the recognition accuracy did not reach the highest. If we continue to increase the number of nodes in the hidden layer, the recognition rate might not increase much, but the structure of the hidden layer will become more complicated, and the time consumption of the algorithm will increase. The results are shown in Figure 6. Therefore, in this paper we have set hidden node number as H=300.

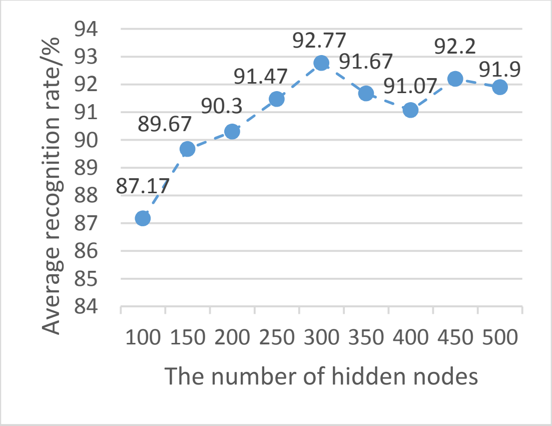

Figure 6. Relationship between average recognition rate and hidden layer nodes.

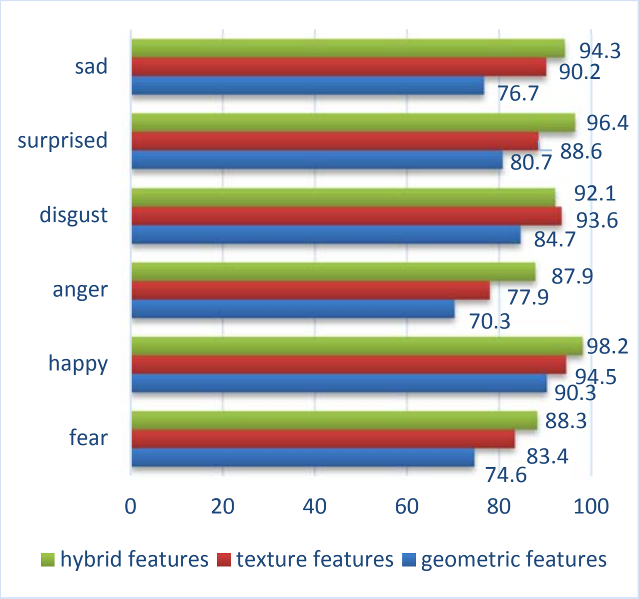

Figure 7. Experiment results using different features based on PSO-ELM (%).

Under the PSO-ELM model with hidden layer node number H=300, we performed the expression recognition experiments by extracting the hybrid features. At the same time, the experiments using other feature extraction methods were also carried out. The recognition results are shown in Table 1. Comparing the three groups of experiments: geometric features, texture features, and hybrid features, the results are shown in Table 2. And Figure7 intuitively reflects the different results of different feature vectors in expression classification.

Table 1. Test accuracy confusion matrix of our PSO-ELM model (%).

	Fear	Happy	Anger	Disgust	Surprised	Sad
Fear	88.3	1.0	2.0	2.0	4.3	2.4
Happy	0.0	98.2	0.0	0.0	1.4	0.4
Anger	2.5	1.1	87.9	3.2	3.0	2.3
Disgust	1.3	0.0	3.2	92.1	1.8	1.4
Surprise	0.3	1.2	1.7	0.4	96.4	0.0
Sad	2.3	0.4	1.5	1.5	0.0	94.3

Table 2. Experiment results using different features based on PSO-ELM (%).

	Fear	Happy	Anger	Disgust	Surprised	Sad	Average
Geometric features	74.6	90.3	70.3	84.7	80.7	76.7	79.5
Texture features	83.4	94.5	77.9	93.6	88.6	90.2	88.0
Hybrid features	88.3	98.2	87.9	92.1	96.4	94.3	92.8

According to Table 1 and Table 2, we can find that only using geometric features, the highest expression recognition rate is 84.7%, and the average recognition rate is 79.5%; only using of texture features, the highest expression recognition rate is 94.5%, and the average recognition rate is 88.0%. However, based on hybrid features, the highest expression recognition rate is up to 98.2%, and the average recognition rate is 92.8%, which is better than the other two. The experimental results according to Table 2 show that the recognition effect of the hybrid features is better than that using only geometric features or texture features.

5. Conclusion

This paper proposes a robust dynamic expression recognition method based on the hybrid features of geometric features and texture features and the PSO-ELM model. In this paper, hybrid features are obtained by fusing the geometric information and the texture information of the feature points. Six PSO-ELM classifiers are trained using six different expression sequences. The dynamic expression sequence can provide more expression information and can eliminate the influence of appearance differences. Through the trained six PSO-ELM models, comparing the similarity between the given sequence and the actual sequence, the expression of the given image sequence can be determined. The method of this paper is tested on CK+ database. The experimental results show that the proposed method has higher average recognition rate of 92.8%, and its highest expression recognition rate is up to 98.2%.

6. References

[1] Argyriou A, Evgeniou T and Pontil M 2008 *Machine Learning* **73(3)** pp 243-272
[2] Abate A F, Nappi M, Riccio D and Gabriele S 2007 P*attern Recognition Letters* **28(14)** pp 1885-1906
[3] Cootes T F, Taylor C J, Cooper D H and Graham J 1995 *Computer Vision and Image Understanding* **61(1)** pp 38-59
[4] Jizheng Y, Xia M and Yuli X 2013 *Journal of Electronics & Information Technology* **35(10)** pp 2403-2410
[5] Ren S, Cao X, Wei Y and Sun J 2014 *Proc. Int. Conf. Computer Vision and Pattern Recognition (Columbus)* vol 25 (USA: Institute of Electrical and Electronics Engineers) pp 1685-1692
[6] Yuan L, Ye C and Yan, W 2014 *Semiconductor Optoelectronics* pp 330-333
[7] Ying Z, Tang J, Li J and Zhang Y 2008 *Acta Electronica Sinica* **36(4)** pp 725-730
[8] Tang J, Zhang Y and Ying Z 2008 *Computer Engineering and Applications* **44(8)** pp 220-222
[9] Saudagare P V and Chaudhari D S 2012 *International Journal of Soft Computing and Engineering* **2 (1)** pp 238-241
[10] Rumelhart D E and Hinton G E 1986 *Nature* **323(6088)** pp 533-536
[11] Kanade T, Cohn J F, Tian and Y L 2000 *Proc. Int. Conf. Automatic Face and Gesture Recognition* (Los Alamitos) pp 46-53
[12] Wang Z and Li S 2011 *Advanced Materials Research* **271-273** pp 165-170
[13] Viola P and Jones M J 2004 *International Journal of Computer Vision* **57(2)** pp 137-154
[14] Zhang K, Zhang Z, Li Z and Qiao Y 2016 *IEEE Signal Processing Letters* **23(10)** pp 1499-1503
[15] Huang G, Zhu Q and Siew C 2004 *Proc. Int. Joint Conf. Neural Networks* (Budapest) vol 2 (USA: Institute of Electrical and Electronics Engineers) pp 985–990
[16] Lucey P, Cohn J F, Kanade T and Saragih J 2010 *Proc. Int. Conf. Computer Vision and Pattern Recognition Workshops* (Los Alamitos) pp 94-1017