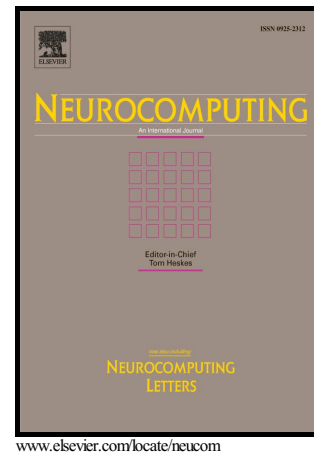


Author's Accepted Manuscript

Cognitive Facial Expression Recognition with Constrained Dimensionality Reduction

Yaxin Sun, Guihua Wen



PII: S0925-2312(16)31559-4
DOI: <http://dx.doi.org/10.1016/j.neucom.2016.12.043>
Reference: NEUCOM17862

To appear in: *Neurocomputing*

Received date: 18 March 2016
Revised date: 20 September 2016
Accepted date: 13 December 2016

Cite this article as: Yaxin Sun and Guihua Wen, Cognitive Facial Expression Recognition with Constrained Dimensionality Reduction, *Neurocomputing* <http://dx.doi.org/10.1016/j.neucom.2016.12.043>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Cognitive Facial Expression Recognition with Constrained Dimensionality Reduction

Yaxin Sun^a, Guihua Wen^b

Jiaxing University, Jiaxing 314001, China
South China University of Technology, Guangzhou 510006, China
sunyaxin2005@163.com
crghwen@scut.edu.cn

Abstract

Facial expression recognition(FER) is an important research area in human-computer interaction. In this paper, a new dimensionality reduction method together with a new classifier are proposed for FER. The goals of most dimensionality reduction contains minimizing the within-class distances. However, the within-class distances for some expressions could be very large, so that to minimize these distances could largely influence the optimization function. To overcome this defect, a new dimensionality reduction method is proposed by adding a penalty item, which is the sum of within distances that are far from each other. Through maximizing this item, the distances among faces with the same expression that are far from each other cannot be minimized to too small. Besides, this method can partly characterize the density information from training samples. To make full use of density information, a new classification method is developed that is based on the enhanced cognitive gravity model. The conducted experiments validate the proposed approach in term of the performance of facial expression recognition. The approach presents the excellent performance over previously available techniques.

Index Terms—facial expression recognition, dimensionality reduction, cognitive gravity model, manifold learning.

1. Introduction

Affective state plays a fundamental role in human interactions, influencing cognition, perception and even rational decision making. This fact has inspired the research field of affective computing which aims at enabling computers to recognize, interpret and simulate affects [1]. Automatic recognition of facial expressions is an important component in human-machine interfaces, human emotion analysis, and other applications such as learning environment, entertainment, customer service, computer games, security/surveillance, educational software, and driver monitoring [7]. Recently, most of facial expression recognition methods are proposed. Generally, they contain three steps: feature extraction, dimensionality reduction, and classification.

Currently, there are a lot of dimensionality reduction methods available for facial expression recognition [2-16, 52,62]. These methods have some defects, such as, class labels not used [4,7,11,12,13,14], the across-classes relationships not considered [2,3,6,9,10,13], or the manifold structure not preserved [5,8,13,52,62]. As facial expression recognition is a classification problem, the within class relationships, the across-classes relationships, and the class labels are all very important. Furthermore, it is investigated that the faces are often distributed on manifold[66]. Manifold learning should be applied to find the face manifold embedded in high dimensional space. To consider these problems together, some dimensionality reduction algorithms are proposed [17-21]. Among them, the semi-supervised long-term relevance feedback (RF) algorithm is much powerful [21], which can present the excellent representation for multimedia data. RF obtains the transformation matrix by maximizing the across-classes distances, minimizing the within-class distances and preserving the multimedia data distribution in low dimensional space.

However, RF cannot be directly used for facial expression recognition, as the extracted features by RF contain both face recognition information and facial expression recognition information. This makes that the distances among faces of different people with the same expression could be very large. Minimizing these distances could largely influence the overall optimization objective for facial expression recognition. To overcome the above disadvantage, this paper proposes a new dimensionality reduction method named enhanced relevance feedback(ERF), which is developed from RF. In the optimization objective of ERF, a penalty item is added to reduce the negative impact, which is the sum of within distances that are far from each other. Besides, ERF can extract some density information from training samples, as the larger the within-class distances among faces that are far from each other, the larger penalty effect of the item, where the density information is benefit for classification.

Recently, two methods can utilize the density information [22,23], which is ignored by many traditional classifiers used in facial expression recognition, such as support vector machine (SVM) [3,4,45], sparse representation classification (SRC) [46, 55] and K -nearest neighbors (KNN) [9,47]. Given a sample with the same distances from two classes, Lizhi Peng et. al. consider that a sample should be more likely classified to the class with larger density [23], whereas Guihua Wen et. al. consider that a sample should be more likely classified to the class with smaller density and then propose the cognitive gravity model (CGM) [22]. CGM is more consistent with human perceptions. However, the mass of a sample in CGM is estimated by self-information, which not only varies with the density of this sample, but also varies with the number of training samples. To overcome this disadvantage, we propose an enhanced cognitive gravity model (ECGM), which can utilize the density information better, as it designs a new algorithm to estimate the mass for the cognitive gravity model, which varies nearly only with the density.

Based on ERF and ECGM, a new hybrid approach is further presented, named as cognitive facial expression recognition (CFER). The main contributions are concluded as follow.

- (1) A new dimensionality reduction algorithm named ERF is proposed, which inherits the advantages of RF, overcomes the

defects of RF, and extracts the density information from training samples.

(2) A new classifier ECGM based on cognitive gravity model is presented for facial expression recognition, which can utilize the density information.

(3) A new algorithm is designed to estimate the mass for the cognitive gravity model, which varies nearly only with the density.

(4) A new hybrid facial expression recognition method (CFER) is presented, which extracts features by histogram of oriented gradients(HOG), performs dimensionality reduction by ERF, and recognizes the facial expression by ECGM.

The remainder of this paper is organized as follows. Section 2 introduces the preliminary. Section 3 presents CFER. Experimental results are presented and discussed in Section 4, while section 5 gives the concluding remarks.

2. Preliminary

The proposed facial expression recognition method is based on long-term relevance feedback (RF)[21] and cognitive gravitation model (CGM) [22]. In this section, they are simply introduced.

2.1 Long-term Relevance Feedback

RF is a constrained dimensionality reduction method which can be applied to many applications to infer a better multimedia representation for retrieval and classification. Given training samples $X = [x_1, \dots, x_n] \in \mathbb{R}^{Q \times n}$ and their labels $[l_1, \dots, l_n] \in \mathbb{R}^n$, RF aims to learn a linear transformation matrix $W = [w_1, w_2, \dots, w_q] \in \mathbb{R}^{Q \times q}$ ($q \ll Q$) to transform each x_i in X to the low dimensional space by $y_i = W^T x_i$. In the low dimensional space, the within-class distances can be minimized, the across-classes distances can be maximized, and the manifold can be preserved. Here W can be obtained by optimizing the following objective:

$$\max_{W^T W = I} \frac{\text{Tr}(W^T S^b W)}{\text{Tr}(A)} \quad (1)$$

$$A = \frac{W^T S^w W}{\text{Tr}(S^w)} + a_1 \times \frac{W^T X L^m X^T W}{\text{Tr}(X L^m X^T)}$$

The goal of $W^T S^w W$ is to minimize the within-class distances, which can be defined as:

$$W^T S^w W = \sum_{l_i = l_j} (W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j) \quad (2)$$

Where $S^w = \sum_{l_i = l_j} (x_i - x_j)^T (x_i - x_j)$.

The goal of $W^T S^b W$ is to maximize across-classes distances, which can be defined as:

$$W^T S^b W = \sum_{l_i \neq l_j} (W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j) \quad (3)$$

Where $S^b = \sum_{l_i \neq l_j} (x_i - x_j)^T (x_i - x_j)$.

The goal of $X L^m X^T$ is to minimize the difference between the manifold in high dimensional space and that in low dimensional space, where L^m can be defined as:

$$L^m = [B_1, B_2, \dots, B_n] \begin{bmatrix} L_1^m & & \\ & \dots & \\ & & L_n^m \end{bmatrix} [B_1, B_2, \dots, B_n]^T \quad (4)$$

$$L_i^m = H - H X_i^T (X_i H X_i^T + \lambda I)^{-1} X_i H$$

$$H = I - \frac{1}{u+1} \mathbf{1}_{u+1} \mathbf{1}_{u+1}^T \in \mathbb{R}^{(u+1) \times (u+1)}$$

Where $\mathbf{1}_{u+1} \in \mathbb{R}^{u+1}$ is a column vector with all ones, $B_i \in \mathbb{R}^{n \times (u+1)}$ if x_p is the q^{th} element of $N_u^+(x_i) = \{x_i, N_u(x_i)\}$, $(B_i)_{pq} = 1$, or else $(B_i)_{pq} = 0$, where $N_u(x_i)$ is u -nearest neighbor of x_i .

2.2 Cognitive Gravitation Model

The Newton's law of universal gravitation can be formally defined by:

$$F(x_i, x_j) = G \frac{m(x_i)m(x_j)}{\|x_i - x_j\|_2^2} \quad (5)$$

Where $F(x_i, x_j)$ is the gravity between two objects with the masses $m(x_i)$ and $m(x_j)$. G is the universal constant.

Obviously, the closer between x_i and x_j , the larger the $F(x_i, x_j)$. This law has been applied to classification by defining the masses with self-information, denoted as CGM, which assumes that $m(x_i)$ decreases with the increase of density of x_i [22]. This is more in line with human cognitive.

CGM is formally defined as:

$$\hat{F}(x_i, x_j) = G \frac{I(x_i)I(x_j)}{\|x_i - x_j\|_2^2} \quad (6)$$

Where $I(x)$ is the self-information of the point x defined through the density function.

The density function of a sample $x \in X$ is defined as the sum of the influence functions of all samples on that point [20]. Given n training samples $\{x_1, x_2, \dots, x_n\} \in X$ and Gaussian function as an influence function, the density function can be defined through Parzen window method:

$$f_d(x) = \sum_{i=1}^n \exp(-\|x - x_i\|_2^2 / 2\gamma^2) \quad (7)$$

According to Shannon's information theory, self-information is a measure of the information content associated with the outcome of a random variable. The smaller the probability of a given event is, the larger the self-information is. Self-information is formally defined through the density function:

$$I(x) = \log\left(\frac{1}{P(x)}\right) = -\log(P(x)) \quad (8)$$

$$P(x) = \frac{f_d(x)}{\sum_{i=1}^n f_d(x_i)}$$

3. Cognitive Facial Expression Recognition

CFER contains three steps: feature extraction, dimensionality reduction, and classification, where pyramid of histogram of oriented gradients (PHOG) is used for feature extraction, a new approach improved from RF is presented to perform the dimensionality reduction, and a new classifier based on the newly enhanced cognitive gravity model is used for classification.

3.1 Pyramid of Histogram of Oriented Gradients

PHOG is based on histogram of oriented gradients (HOG). As mentioned in [39, 40], there are many parameters needed to be considered for HOG: the size of spatial cells, orientation bins, and the size of spatial blocks. Setting different values for the above parameters, HOG can extract different features from the face. Therefore, we first normalize the size of images to 128×128 pixels. And then, we choose two kinds of spatial cell segmentations: (1) face images are divided into 256 cells, where the cell size is 8×8 and neighboring cells are not overlapped; (2) the face images are divided into 64 cells, where the cell size is 16×16 and neighboring cells are not overlapped. Another import parameter is the orientation bin number. We set the bin number as $H_n = 16$. The final features are combined by the features extracted by HOG with above two spatial cell segmentations. Obviously, the first spatial cell segmentation could be obtained by dividing each cell of the second spatial cell segmentation into 4 cells, which is a typical pyramid-based feature extraction method, and we name this method as PHOG.

3.2 Enhanced Relevance Feedback (ERF)

The number of features extracted by PHOG is very large, possibly leading to the curse of dimensionality when these features are directly used for classification, so that the dimensionality reduction is required, where the long-term relevance feedback (RF) is a recently proposed approach with good dimensionality reduction effect for multimedia data.

However, RF cannot be directly used for facial expression recognition. Because in Eq.(1), for a given a_1 , $W^T S^w W$ could be preferentially optimized, possibly leading to that $W^T X L^m X^T W$ cannot be optimized with the desired optimization goal. This is because different people may have the same facial expression whereas their faces are much different. Some of features extracted by PHOG that are much beneficial to face recognition could be unsuitable for facial expression recognition. As a result, minimizing the distances among these faces could largely influence for the final optimization objective for the facial expression recognition. On the other hand, if the manifold is preserved by increasing a_1 , the within-class distances among some faces could not be well

minimized, as the weight of optimizing $W^T S^w W$ is reduced. Therefore minimizing the within-class distances among these faces could have a too small influence for the optimization objective.

The second reason is that a_1 could be very sensitive when S^w and XLX^T are normalized by their traces. For example, in reference [21], a_1 is selected in the range $[10^{-2}, 10^2]$, which is instability and hard to be selected.

To overcome the first defect, a penalty item is added in minimizing the within-class distances among the samples that are far from each other. To overcome the second defect, a better normalization method is designed. As a result, the optimization objective can be defined as follow:

$$W^* = \arg \max_{W^T W = I} \frac{W^T XL^b XW + a_1 W^T XL^f XW}{W^T XL^w XW + a_2 W^T XL^m XW} \quad (9)$$

The goal of $W^T XL^b XW$ is to maximize the across-classes distances, which is defined by Eq.(10). Compared Eq.(9) with Eq.(1), $W^T XL^b XW$ in Eq.(9) are not normalized by trace. Alternatively, the across-classes distances of the c^{th} class are normalized by dividing n_c . Because n_c distances are computed for a sample in the c^{th} class, and dividing n_c can make that $W^T XL^w XW$ does not largely vary with the number of the training samples, where $n_c = n - n_c$, n_c is the number of training samples of the c^{th} class.

$$W^T XL^b XW = \sum_{c=1}^C \sum_{l=c \text{ and } l \neq c} \frac{(W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j)}{n_c} \quad (10)$$

$$\text{Where } L^b = \sum_{c=1}^C \sum_{l=c \text{ and } l \neq c} \frac{(x_i - x_j)^T (x_i - x_j)}{n_c}$$

The goal of $W^T XL^w XW$ is to minimize the across-classes distances, which is defined by Eq.(11). Similarly, $W^T XL^w XW$ is also not normalized by trace, but the within-class distances of c^{th} class are normalized by dividing n_c , as n_c distances are computed for a sample in c^{th} class.

$$W^T XL^w XW = \sum_{c=1}^C \sum_{l=c \text{ and } l = c} \frac{(W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j)}{n_c} \quad (11)$$

$$\text{Where } L^w = \sum_{c=1}^C \sum_{l=c \text{ and } l = c} \frac{(x_i - x_j)^T (x_i - x_j)}{n_c}$$

$W^T XL^f XW$ aims to penalize the within-class distances among faces that are far from each other, which is defined by Eq.(12), where $N_k(x_j)$ is k -nearest neighbor of x_j . Similarly, $W^T XL^f XW$ is also not normalized by trace, but each distance is normalized by dividing n_f , where $n_f = n_c - k$.

$$W^T XL^f XW = \sum_{c=1}^C \sum_{\substack{l=c \text{ and } l = c \\ \text{and } x_i \notin N_k(x_j) \\ \text{and } x_j \notin N_k(x_i)}} \frac{(W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j)}{n_f} \quad (12)$$

$$\text{Where } L^f = \sum_{c=1}^C \sum_{\substack{l=c \text{ and } l = c \\ \text{and } x_i \notin N_k(x_j) \\ \text{and } x_j \notin N_k(x_i)}} \frac{(x_i - x_j)^T (x_i - x_j)}{n_f}$$

The idea of constructing $W^T XL^f XW$ comes from our early work that local reconstruction and global preserving based semi-supervised dimensionality reduction (LRGPSSDR) [20]. However, LRGPSSDR is mainly used for face recognition where the number of classes is large and the number of samples in each class is small, and the goal of similar item is to keep non-neighborhood among across classes faces, which is much different with our method's.

After adding the item defined by Eq.(12) to the optimization objective, the within-class distances among faces that are far from each other cannot be minimized to too small, and then the first defect can be overcome.

The goal of $W^T XL^m XW$ is to preserve the manifold, which is defined by Eq.(4).

The projection vector W in Eq.(9) can be obtained by solving the generalized eigenvalues problem presented by Eq.(13). Let the column vector w_1, w_2, \dots, w_d be the solution of Eq.(13), ordered according to their eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, and the transformation matrix W can be obtained as $W = [w_1, w_2, \dots, w_d]$. Based on the obtained W by ERF, the features in low dimensional space can be obtained by $y = W^T x$.

$$X(L^b + a_1 L^f)X^T w = \lambda X(L^w + a_2 L^f)X^T w \quad (13)$$

ERF is summarized as algorithm 1. In Algorithm 1, the most time consuming operation is the eigenvalue decomposition in step 3, whose time complexity is $O(Q^3)$.

Algorithm1 Enhanced Relevance Feedback (ERF)

Input: labeled training samples $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{Q \times n}$, its labels $\{l_1, l_2, \dots, l_n\}$, neighborhood size u , trade-off parameters a_1, a_2 , the dimension of low-dimensional space q .

Output: transformation matrix W

- 1: Construct L^b , L^w , L^f , and L^m
 - 2: Construct the optimization objective by Eq. (9)
 - 3: Compute transformation matrix W according to Eq.(13)
-

ERF is improved from RF, which can solve the problem that different peoples may have the same facial expression whereas their faces are much different. Besides, the density information of training samples also can be extracted by ERF, as the larger the within-class distances among faces that are far from each other, the larger penalty effect of the item defined by Eq.(12) in minimizing Eq.(11).

Some other subspace learning methods also can solve the problem of RF. Jinghua Wang[63] proposed a new method named orthogonal discriminant vector for face recognition across pose. In this method, they found many independent vectors that are orthogonal to all the gallery face images of the i -th individual and not orthogonal to any gallery images of the other individuals. Obviously, regardless of how changes in density of i -th individual, the density of independent vectors is not changed. However, this method cannot handle the complex manifold structure, and the density information also cannot be extracted. Ref.[5] represented a local fisher discriminant analysis (LFDA) for facial expression recognition. In LFDA, only distances of nearest neighbors of within-class samples are minimized, and only distances of nearest neighbors of across-classes samples are maximized. This is another way to solve the problem of RF, however, no relationships among two far from samples are defined, and then two far from across samples could be near with each other in the low-dimensional space. Furthermore, LFDA only divides image samples in each class into multiple local classes, and then LFDA cannot extract the density information. Obviously, the first defect of LFDA can be overcome by ERF, as only the within-class distances among faces that are far from each other is penalized in ERF. Some kernel subspace learning, such as kernel Fisher discriminant analysis[62] also could be used to solve the above problem. However, the kernel and the corresponding kernel parameters are hardly chosen. Furthermore, the density information also cannot be extracted by these methods.

Obviously, compared with the above methods, ERF not only can solve the problem of RF, but also can overcome the defect of above methods and can extract more useful information.

To further illustrate the advantages of ERF, the visual results of samples in 2-dimensional space are presented on Fig.1, where the 2-dimensional space are obtained respectively by linear discriminant analysis (LDA), RF and ERF are presented from the first row to the third row respectively. The left two columns of Fig.1 are the visual results of training samples, the right two columns of Fig.1 are the visual results of testing samples, and the digital on the title are the selected dimensions. Different colors on Fig.1 represent the different classes. The samples of first 30 people on CK+ database are used as testing samples, while the rest samples are used as training samples.

The features in low dimensional space obtained by ERF have better effectiveness for classification than that obtained by LDA and RF. It can be seen from the first row of Fig.1 for LDA that the training samples are nearly projected on one point, while the testing samples are projected on the whole plane that are totally confused. Again, the second row of Fig.1 for RF shows that the distribution range of training samples are increased and the testing samples are formed to some clear clusters, which is more beneficial to the classification. It also can be seen from the third row that the training samples are further widely distributed and the testing samples are formed to more clear clusters, where the low dimensional space is obtained by ERF. For example, in Fig.1 (f)(1,3), there are two clusters that do not coincide with the other clusters, whereas in Fig.1 (d)(1,3) there is only one cluster that does not coincide with the other clusters.

In addition to the above advantages, ERF also can make the training samples to have different densities in low dimensional space. For example, the training samples marked by pink in Fig.1 (e) (1,2) and the testing samples marked by pink in Fig.1 (f) (1,2) are more widely distributed, so that the topological structure of neighbors can be more preserved. This is much beneficial to CGM as a classifier, because it makes full use of density information that to some extent can characterize the topological structure of neighbors. The case is similar to the other kinds of training samples and testing samples, such as marked by blue, cyan and black in Fig.1 (e)(1,2).

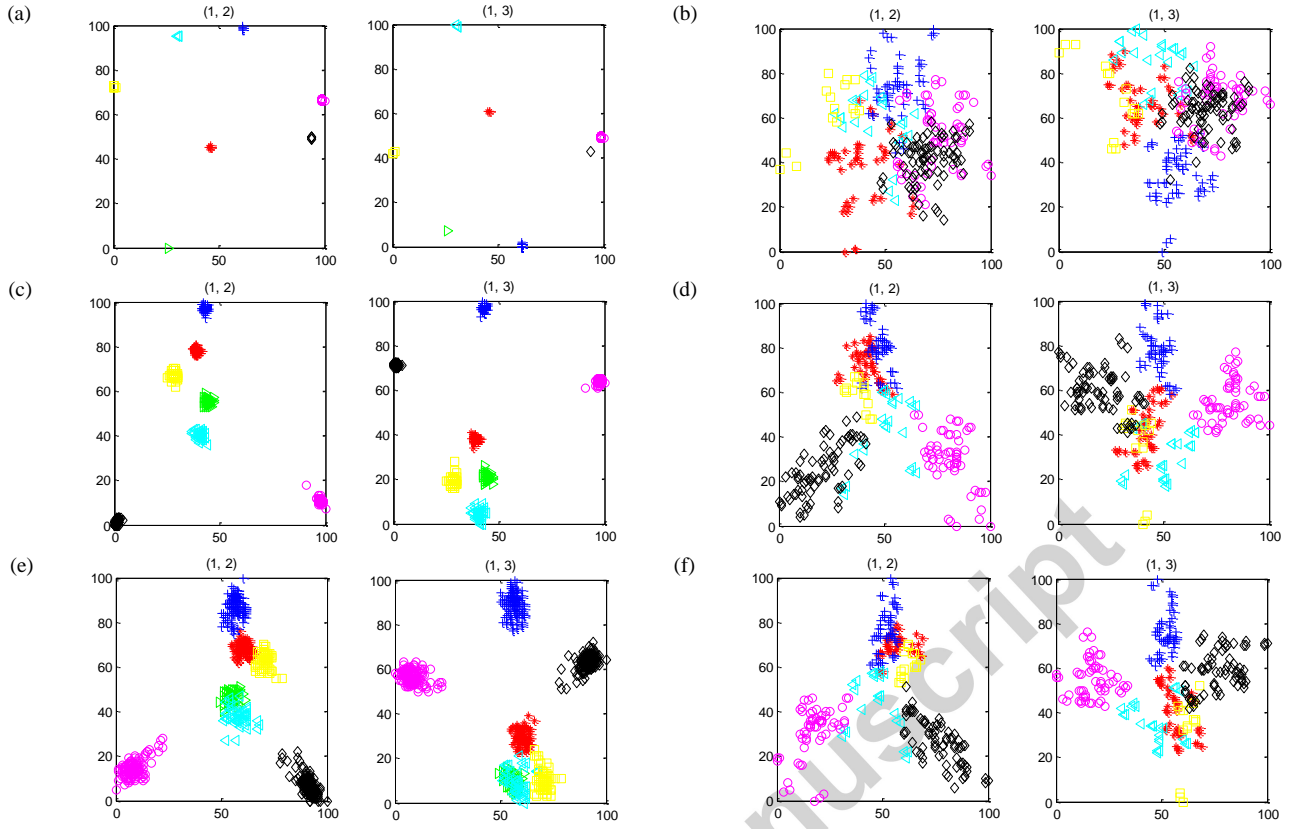


Fig. 1 The visual results of in 2-dimensional spaces, where the 2-dimensional spaces are respectively obtained by LDA, RF and ERF from top to bottom

3.3 Enhanced Cognitive Gravitation Model (ECGM).

The classifier with CGM can utilize the density information to increase the recognition accuracy[22]. However, it has a defect derived from its mass definition based on the self-information. This can be illustrated by an example shown in Fig.2, where the coordinates of samples in this two-dimensional space are the features of samples. In this example, the density of Class 1 is smaller than that of Class 2 and $\|x - x_A\|_2 > \|x - x_B\|_2$, where $\|\cdot\|_2$ is L_2 normalization. We define $\|x - x_A\|_2 > \|x - x_B\|_2$, because in such case, x is classified to class 2 by tradition classifiers, such as k-nearest neighbor (KNN), nearest neighbor (NN), support vector machine (SVM), but x should be classified to class 1 according to the theory of CGM, which is a good example to explain when the CGM is better than traditional classifiers. To reach the goal of CGM, the condition that $F(x, x_A) > F(x, x_B)$ should be met. Specifically, the samples in bigger circle should be classified to class 1 and the samples in smaller circle should be classified to class 2. According to Eq.(5):

$$\begin{aligned}
 & F(x, x_A) > F(x, x_B) \\
 \Rightarrow & G \frac{m(x)m(x_A)}{\|x - x_A\|_2^2} > G \frac{m(x)m(x_B)}{\|x - x_B\|_2^2} \\
 \Rightarrow & \frac{m(x_A)}{m(x_B)} > \frac{\|x - x_A\|_2^2}{\|x - x_B\|_2^2}
 \end{aligned} \tag{14}$$

It can be seen from Eq.(14) that $m(x_A) / m(x_B)$ must meet a condition. Specially, if x is at the junction of two circles in Fig.2, the condition $F(x, x_A) = F(x, x_B)$ must be met, and then $m(x_A) / m(x_B) = \|x - x_A\|_2^2 / \|x - x_B\|_2^2$. However, in CGM, $m(x)$ is defined by self-information $I(x)$ that varies with both the density of x and the number of training samples in the non-linear way, which are proved by Proposition 1 and Proposition 2. As a result, to better define cognitive gravity model, a more reasonable method should be used to estimate the mass $m(x)$ of a sample x .

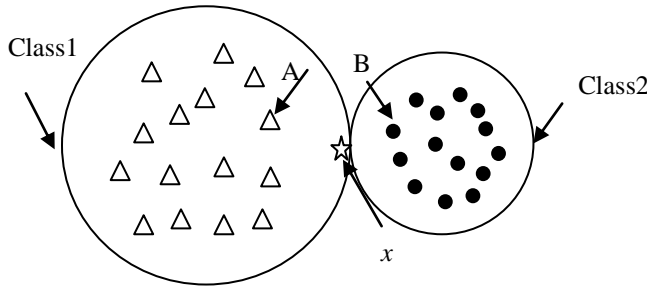


Fig.2 An example used to show the defect of CGM

Proposition 1: $I(x)$ decreases with the increasing of density.

Proof: Given n training samples, $\sum_{i=1}^n f_d(x_i)$ can be seen as a constant. Let $\sum_{i=1}^n f_d(x_i) = \eta$, $dI(x) / d\|x - x_i\|_2^2$ can be computed as follow:

$$\begin{aligned} \frac{dI(x)}{d\|x - x_i\|_2^2} &= -\log\left(\frac{f_d(x)}{\eta}\right)' = -\frac{1}{\eta f_d(x)} f_d'(x) \\ &= \frac{\exp(-\|x - x_i\|_2^2 / 2\gamma^2)}{2\eta\gamma^2 \sum_{j=1}^n \exp(-\|x_i - x_j\|_2^2 / 2\gamma^2)} \end{aligned} \quad (15)$$

It can be seen from Eq.(15) that $dI(x) / d\|x - x_i\|_2^2 > 0$, illustrating that $I(x)$ increases positively with the distance $\|x - x_i\|_2$. On the other hand, generally $\|x - x_i\|_2$ decreases with the increasing of density of x when $x_i \in N(x)$, where $N(x)$ is the neighborhood of x with the given neighborhood size. Therefore $I(x)$ decreases with the increasing of density.

Proposition 2: $I(x)$ nonlinearly increases with the increasing of number of training samples.

Proof: Let $\sum_{i=1}^n f_d(x_i) = z$ where z is a variable as $\sum_{i=1}^n f_d(x_i)$ increases with the increasing of number of training samples, rewrite $I(x)$ as:

$$I(x) = \log\left(\frac{1}{p(x)}\right) = \log\left(\frac{z}{f_d(x)}\right) \quad (16)$$

$dI(x) / dz$ can be computed as:

$$dI(x) / dz = \log(z / f_d(x))' = 1 / (f_d(x) * z) \quad (17)$$

Obviously, $dI(x) / dz > 0$ and $dI(x) / dz$ is not a constant, so that $I(x)$ nonlinearly increases with the increasing of number of training samples.

It can be seen from Eq.(15) that $dI(x) / d\|x - x_i\|_2^2$ increases with the contribution of x_i in computing $f_d(x)$, which is consistent with the human perception about the requirement of mass in cognitive gravity model. If the defect of $I(x)$ that $dI(x) / d\|x - x_i\|_2^2$ varies with η can be reasonably solved, $m(x_i) / m(x_j)$ and $m(x)$ could be correctly estimated. Fortunately, the defect of $I(x)$ can be solved by scaling and shifting $I(x_i), i = 1, 2, \dots, n$ to $\delta(m(x))$ and $\bar{m}(x)$, where $\delta(m(x))$ and $\bar{m}(x)$ are the correct standard deviation and mean of $m(x)$. Obviously, there is a Chicken-and-egg problem, which can be solved by another method that can be used to estimate $m(x)$, which is $\|x - x^K\|_2$.

$\|x - x^K\|_2$ can be used to estimate $m(x)$ when samples are evenly distributed around x , where x^K is the K^{th} neighbor of x , but $\|x - x^K\|_2$ cannot be used to correctly estimate $m(x)$ when samples are arbitrarily distributed around x . However, $\|x - x^K\|_2$ can be used to estimate the correct $\bar{m}(x)$ and $\delta(m(x))$.

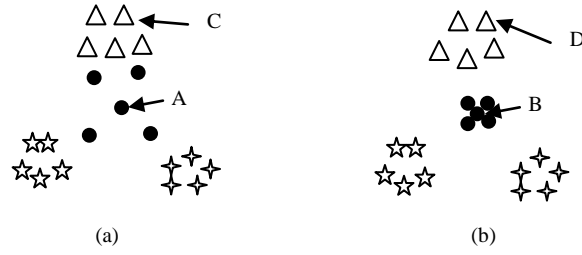


Fig.3 Two cases that $m(x)$ cannot be correctly estimated by $\|x - x^K\|_2$

Suppose that the samples are arbitrarily distributed around x_A, x_B respectively, shown as Fig.3, $m(x_A)$ defined by $\|x_A - x_C\|_2$ would be underestimated and $m(x_B)$ defined by $\|x_B - x_D\|_2$ would be overestimated. Statistically, the probabilities of above two cases are the same, so that $\bar{m}(x)$ can be estimated as follows:

$$\bar{m}(x) = \frac{1}{n} \sum_{i=1}^n \|x_i - x^K\|_2 \quad (18)$$

By the similar reasons, $\delta(m(x))$ can be estimated as follows:

$$\delta(m) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\|x_i - x^K\|_2 - \bar{m}(x))^2} \quad (19)$$

From the above discussion, $m(x)$ can be correctly estimated by Eq.(20), where $\bar{I}(x)$ and $\delta(I(x))$ are the mean and variance of $I(x)$, where a parameter λ is applied to correct the estimated error.

$$m(x) = (I(x) - \bar{I}(x)) * \lambda * \delta(m(x)) / \delta(I(x)) + \bar{m}(x) \quad (20)$$

$$\bar{I}(x) = \sum_{i=1}^n I(x_i) / n, \delta(I(x)) = \sum_{i=1}^n (I(x_i) - \bar{I}(x))^2 / n$$

To reduce the gravities from far away samples, exponential function is applied to redefine the cognitive gravity model as

$$F(x_i, x_j) = \exp(-w_{ij}), w_{ij} = \frac{\|x_i - x_j\|_2^2}{m(x_i)m(x_j)} \quad (21)$$

The algorithm of computing mass for CGM can be summarized as algorithm 2. It can be seen that computing the mass for a sample is only related to itself and the training samples, so that the masses of training samples can be computed in advance in real applications.

Algorithm 2 CM

Input: training samples $X = [x_1, x_2, \dots, x_n]$, mass parameter K , a testing sample x .

Output: the mass $m(x)$ of x

1: Compute the density of x by

$$f_d(x) = \sum_{i=1}^n \exp(-\|x - x_i\|_2^2 / \gamma^2), \gamma^2 = \sum_{i=1}^n \|x - x_i^K\|_2^2 / n$$

2: Compute $I(x)$ through density:

$$I(x) = \log\left(\frac{1}{p(x)}\right) = -\log(p(x))$$

$$P(x) = \frac{f_d(x)}{\sum_{i=1}^n f_d(x_i)}$$

3: Compute $m(x)$ by normalizing $I(x)$

$$m(x) = (I(x) - \bar{I}(x)) * \delta(m) * \lambda / \delta(I) + \bar{m}(x)$$

$$\bar{m}(x) = \frac{1}{n} \sum_{i=1}^n \|x_i - x_i^K\|_2, \delta(m) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\|x_i - x_i^K\|_2 - \bar{m}(x))^2}$$

$$\bar{I}(x) = \sum_{i=1}^n I(x_i) / n, \delta(I) = \sum_{i=1}^n (I(x_i) - \bar{I}(x))^2 / N$$

In Algorithm 2, the most time consuming operation is computing $f_d(x)$ in step 1, whose time complexity is $O(n)$.

3.4 The proposed CFER

CFER is summarized as the algorithm 3, which integrates ERF and ECGM. Before using CFER, the training steps are required. In training steps, the features of all training samples are extracted by PHOG. Then the linear transform matrix is trained by ERF and all training samples are projected to the low-dimensional space. The training samples in the low dimensional space are denoted as $[y_1, y_2, \dots, y_n]$. Finally, the masses of $[y_1, y_2, \dots, y_n]$ are computed by Algorithm 2. the most time consuming operation of the training step is ERF, which is $O(Q^3)$, where Q is the dimension of training samples.

Algorithm 3 CFER

Input: training samples in the low dimensional space denoted as $[y_1, y_2, \dots, y_n]$, their labels denoted as $\{l_1, l_2, \dots, l_n\}$, their masses denoted as $M = [m(y_1), m(y_2), \dots, m(y_n)] \in R^n$, linear transform matrix W , the classifier parameter k , and a testing sample x .

Output: label l of the testing sample x

1: Extract features for x by PHOG.

2: Project x to the low-dimensional space by $y = W^T x$.

3: Compute the mass of y by the algorithm 2.

4: Compute the gravities $[F_1, F_2, \dots, F_n]$ between y and $[y_1, y_2, \dots, y_n]$ by Eq.(21)

For $c=1$ to C

5: Find k biggest gravities $[F_{c1}, F_{c2}, \dots, F_{c\alpha}]$ from c^{th} class, where $F_{c\alpha}$ is the α^{th} biggest gravity between y and the training samples of c^{th} class.

6: Compute the gravity between y to c^{th} class by

$$F_c = F_{c1} + F_{c2} + \dots + F_{c\alpha}$$

End

7: Obtain the label of x by $l = \arg \max_{c=1,2,\dots,C} F_c$

After obtaining the linear transform matrix W , the computational complexity of CFER is composed of the computational complexity of PHOG, algorithm 2, and finding k biggest gravities from each class, whose computational complexity are $O(RH)$, $O(n)$, and $O(k * \log n)$ respectively, where R is the width of face image and H is the height of face image. As a result, the computational complexity of CFER is $O(RH+n+k * \log n)=O(n)$.

4. Experiments

4.1 Facial expression Databases

There two benchmark databases are used in the experiments. The first database is Japanese female facial expression (JAFPE) database [37]. JAFPE database contains facial images of 10 Japanese females, where each person has two to four samples for each expression. In total, there are 213 grayscale facial expression images in this database, with pixel resolution 256×256 .

The second is extended Cohn-Kanade (CK+) database [38]. The CK+ database is a larger database commonly used for facial expression recognition. The experiments involve 123 subjects and 327 emotion-labeled image sequences, where emotion labels are happiness, sadness, anger, fear, disgust, contempt, and surprise. Three final frames of each sequence with peak expression are used as the expressive images in our experiments.

4.2 The compared methods and Performance Evaluation Criteria

In CFER, ERF is used to perform the dimensionality reduction and the classifier based on ECGM is applied to perform the classification. Therefore, many dimensionality reduction methods are compared. They are linear discriminant analysis (LDA) [41], semi-supervised discriminant analysis (SDA) [42], local Fisher discriminant analysis (LFDA) [5], PCA[41], neighborhood

preserving embedding (NPE) [43], locality preserving projections (LPP) [44], and RF [21]. In all methods, the graph constructing parameter and the neighbor size of ERF is selected from [3:3:18]. SDA has a trade-off parameter, which is selected from [0.2:0.2:1.0]. RF also has a trade-off parameter that is selected from $[10^{-2}:10^{-1}:10^2]$. ERF has two trade-off parameters that are selected from [0.2:0.2:1.0].

To make comparison with ECGM, the following classifiers are selected: support vector machine(SVM) [45], sparse representations classifier (SRC) [46], local mean classifier (LMC)[47], and cognitive gravitation model classifier (CGMC)[22]. The polynomial kernel is used for SVM, whose kernel parameter is selected from [2:1:4], where the polynomial kernel is chose, as the result of SVM with polynomial kernel is nearly the same as that of SVM with radial basis function(RBF) kernel. (Sparse Modeling Software) SPAMS package [48] is used to optimize the objective of SRC. The neighborhood sizes for both LMC and CFER are selected from [3:3:18]. The mass parameter of CFER is selected from [6:6:60]. The parameter λ in CFER takes 1.5, as the experiment results vary little when $1.3 \leq \lambda \leq 1.7$.

Accuracy is taken as the performance evaluation criteria [5,6,7]. All algorithms are evaluated by leave one subject out (LOSO) approach [24] or 10-fold cross-validation (10-FOLD) approach[52]. The parameters of dimensionality reduction methods are firstly selected, where SVM with polynomial kernel is used for classification, whose kernel parameter is 3. Subsequently the parameters of classifiers are selected, where dimensionality reduction methods use the selected parameters. 5-fold cross-validation experiments on the training samples are used for selecting parameters.

4.3 Experiments on JAFFE

In this subsection, we will use LOSO and 10-FOLD experiments on JAFFE to evaluate the CFER. Faces are detected by Adaboost [49], as the landmarks of faces are not offered in this database and cannot be precisely located by active appearance model (AAM) [50] on JAFFE. Some wrong results of landmarks of faces located by AAM are presented in Fig.4. Some examples of faces detected by Adaboost are presented in Fig.5.

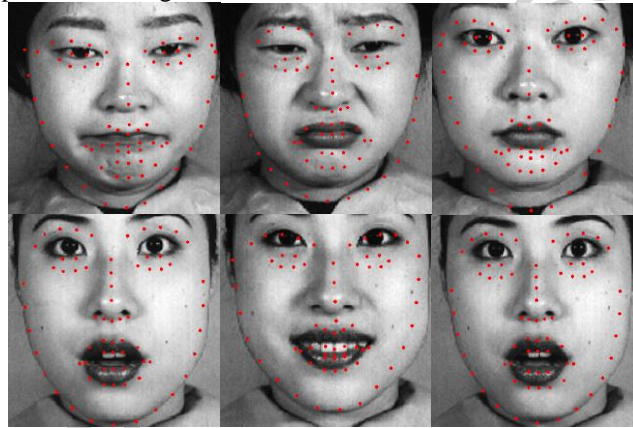


Fig.4 Examples that the landmarks cannot be exactly located by AAM



Fig.5 Example faces obtained by Adaboost on JAFFE

4.3.1 Experiments on JAFFE with LOSO

In most applications of facial expression recognition, the samples of testing subjects cannot be used for training, so that we firstly evaluate our algorithm by the experiments with LOSO.

Table 1 THE PERFORMANCES WITH LOSO ON JAFFE, WHERE THE FACES ARE OBTAINED BY ADABOOST

Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
SVM	68.08	SRC	61.97	LMC	62.91
LDA+SVM	69.01	LDA+SRC	61.50	LDA+LMC	69.01
SDA+SVM	69.48	SDA+SRC	65.73	SDA+LMC	67.61
LFDA+SVM	72.30	LFDA+SRC	67.14	LFDA+LMC	71.36
PCA+SVM	68.08	PCA+SRC	63.85	PCA+LMC	64.32
NPE+SVM	69.95	NPE+SRC	64.32	NPE+LMC	68.54

LPP+SVM	72.77	LPP+SRC	66.67	LPP+LMC	72.30
RF+SVM	69.48	RF+SRC	56.34	RF+LMC	65.73
ERF+SVM	73.24	ERF+SRC	69.95	ERF+LMC	74.18
CFER	76.46	ERF+CGMC	74.58		

The performances with LOSO of all compared classifiers are given in Table.1. It can be seen from Table 1 that most of the accuracies of SVM, SRC and LMC with ERF are higher than that of SVM, SRC and LMC with the other compared dimensionality reduction methods, illustrating that ERF is better than the compared dimensionality reduction methods including RF. Secondly, the accuracies of CFER and CGMC with ERF are much higher than that of SVM, SRC and LMC with ERF, which proves that the classifiers based on cognitive gravity model are much better than the compared classifiers for facial expression recognition. Thirdly, the accuracies of CFER are much higher than that of CGMC with ERF, illustrating that ECGM is better than CGM.

4.3.2 Experiments on JAFFE with 10-FOLD

Many advances[52,60, 65,69] are evaluated by 10-FOLD on JAFFE, so that we also evaluate our method by 10-FOLD on JAFFE.

The performances with 10-FOLD of all compared classifiers are given in Table.2. It can be seen from Table.2 that ERF is also better than the compared dimensionality reduction methods, and the accuracies of CFER and CGMC with ERF are also higher than that of SVM, SRC and LMC with ERF, which prove the effectiveness of ERF and the classifiers based on cognitive gravity model. Furthermore, it also can be seen from Table.2 that the accuracy of CFER is the highest, which proves the improvement of ECGM.

Table 2 THE PERFORMANCES WITH 10-FOLD ON JAFFE, WHERE THE FACES ARE OBTAINED BY ADABOOST

Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
SVM	94.30	SRC	89.29	LMC	83.52
LDA+SVM	94.69	LDA+SRC	91.74	LDA+LMC	95.31
SDA+SVM	94.46	SDA+SRC	93.52	SDA+LMC	94.18
LFDA+SVM	95.82	LFDA+SRC	94.23	LFDA+LMC	95.92
PCA+SVM	95.16	PCA+SRC	90.04	PCA+LMC	67.32
NPE+SVM	95.87	NPE+SRC	95.12	NPE+LMC	95.87
LPP+SVM	95.26	LPP+SRC	91.64	LPP+LMC	95.26
RF+SVM	95.63	RF+SRC	94.88	RF+LMC	95.02
ERF+SVM	96.05	ERF+SRC	95.40	ERF+LMC	95.87
CFER	96.24	ERF+CGMC	96.00		

4.3.3 Compared with recent advances

In Table 3, we compare the performance of CFER on JAFFE to the other algorithms in the literature. The performances of these algorithms are directly cited from its original references.

It can be seen from Table.3 that the best performance with LOSO is 76.46%, which is obtained by our method, and the second best performance with LOSO is 74.18%, which is obtained by Ref.[26]. Obviously, our method is better than the listed advances with LOSO.

It also can be seen from Table.3 that the best performance with 10-FOLD is 96.33%, which is obtained by Ref.[65], the second best performance with 10-FOLD is 96.24%, which is obtained by our method, and the third best performance with 10-FOLD is 96.20%, which is obtained by Ref.[69]. Compared with these advances, our method owns following advantages. Firstly, one training data or testing data of our method only need a frame, however, the method proposed by Ref.[65] is a sequence-based method, whose one testing data or training data should contains two or more frames, which means that samples of each expression of each person on JAFFE database are either used for training or used for testing in once time. Secondly, the landmarks are not used in our method, however, the landmarks must be used in the above two advances. Thirdly, the classifier in our method is a single classifier, however, the classifier in the method proposed by Ref.[69] is a ensemble classifier, which contains 5 base classifiers, and the best performance among these base classifiers is 91.6%, which is much lower than our performances. Fourthly, 7 expressions are recognized by our method, however, only 6 expressions are recognized by the method proposed by Ref.[65]. Obviously, our method is also better than the listed advances with 10-FOLD.

Table 3 THE PERFORMANCES OF RECENT ADVANCES ON JAFFE, WHERE 6-FOLD REPRESENTS 6-FOLD CROSS-VALIDATION, 10-FOLD REPRESENTS 10-FOLD CROSS-VALIDATION, LOO REPRESENTS LEAVING ONE SAMPLES OUT, AND LOSO REPRESENTS LEAVE ONE SUBJECT OUT, RS REPRESENTS RANDOM SELECTING MANY SAMPLES AS TRAINING SAMPLES

Ref	Classes	Dynamic	Measure	Using Landmarks	Accuracy
[2]	7	NO	6-FOLD	NO	85.90
[5]	7	NO	LOO	NO	94.70
[6]	6	NO	LOSO	Eye Center	53.01
[26]	7	NO	LOSO	NO	74.18
[51]	7	NO	LOSO	Eye Center	71.00
[69]	7	NO	LOSO	20 Face landmarks	70.00
[69]	7	NO	10-FOLD	20 Face landmarks	96.20
[35]	6	NO	RS	NO	89.09

[36]	7	NO	LOO	68 Face landmarks	92.90
[52]	7	NO	10-FOLD	NO	85.00
[60]	7	NO	10-FOLD	19 Face landmarks	91.80
[65]	6	NO	10-FOLD	Lips, eyes, forehands	96.33
[67]	7	NO	5-FOLD	NO	85.70
[68]	7	NO	LOO	NO	88.89
[30]	7	NO	RS	NO	87.75
CFER	7	NO	LOSO	NO	76.46
CFER	7	NO	10-FOLD	NO	96.24

4.4 Experiments on CK+

In this subsection, we perform experiments with LOSO strategy on CK+ to evaluate the CFER. Because the landmarks of faces are offered by CK+, the faces in CK+ are obtained by two ways: (1) the faces are aligned and cropped by landmarks; (2) the faces are detected by Adaboost. Using the first way to obtain faces can facilitate comparison with the existed facial expression recognition methods, as most existed methods used the landmarks for facial expression recognition. Using the second way to obtain faces can evaluate the effectiveness of CFER when CFER is used in a complex environment, as where the landmarks of faces cannot be exactly located.

4.4.1 Experiments on CK+ when faces are obtained and corrected by landmarks.

The input faces of the most of existing methods are aligned before extracting features [6, 15, 24-25, 28-29, 31-33], which can reduce the influence of angles of input faces. Here the input faces are aligned by the three steps shown in Fig.6. Firstly, the left eye center p_l and the right eye center p_r are respectively computed by the mean of 38th to 42th landmarks and the mean of 43th to 48th landmarks of the face. Secondly, the image is rotated according to the angle between two eyes centers. Thirdly, the face is cropped from the square region in the image, where the top-left vertex of square is p_f and the width of square is $2.5 * d_e$. p_f and d_e are computed as:

$$p_{fx} = p_{lx} - 0.5 * d_e, p_{fy} = p_{ly} - 0.3 * d_e$$

$$d_e = \sqrt{(p_{lx} - p_{rx})^2 + (p_{ly} - p_{ry})^2} \quad (22)$$

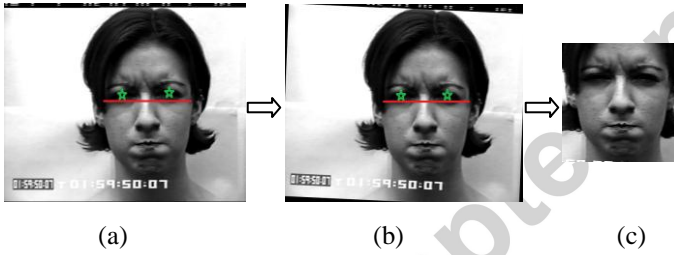


Fig.6 The process of obtaining face by landmarks



Fig.7 Example faces obtained by landmarks

Table 4 THE PERFORMANCES ON CK+, WHERE THE FACES ARE OBTAINED BY LANDMARKS

Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
SVM	94.50	SRC	89.81	LMC	86.95
LDA+SVM	91.85	LDA+SRC	83.28	LDA+LMC	91.85
SDA+SVM	92.46	SDA+SRC	87.77	SDA+LMC	92.15
LFDA+SVM	92.35	LFDA+SRC	87.77	LFDA+LMC	92.35
PCA+SVM	94.50	PCA+SRC	90.01	PCA+LMC	89.30
NPE+SVM	92.05	NPE+SRC	88.18	NPE+LMC	92.46
LPP+SVM	92.76	LPP+SRC	86.24	LPP+LMC	92.86
RF+SVM	95.21	RF+SRC	88.38	RF+LMC	93.07
ERF+SVM	95.31	ERF+SRC	96.53	ERF+LMC	95.92
CFER	97.66	ERF+CGMC	97.04		

Example faces obtained by landmarks of faces are shown in Fig.7. The performances of all compared classifiers are given in

Table.4. It can be seen that ERF is better than the compared dimensionality reduction methods. CFER and CGMC with ERF are also better than SVM, SRC and LMC with ERF. Finally, CFER is better than CGMC with ERF.

4.4.2 Experiments on CK+ when faces are obtained by Adaboost

Because in many real applications, the landmarks of faces in the training database such as JAFFE are not provided, the experiments that the input faces are obtained by Adaboost should be also performed on CK+. Example faces are presented in Fig.8.



Fig.8 Example faces obtained by Adaboost on CK+

Table 5 THE PERFORMANCES ON CK+, WHERE THE FACES ARE OBTAINED BY ADABOOST

Algorithm	Accuracy	Algorithm	Accuracy	Algorithm	Accuracy
SVM	92.24	SRC	87.36	LMC	82.47
LDA+SVM	89.09	LDA+SRC	82.67	LDA+LMC	89.09
SDA+SVM	90.83	SDA+SRC	86.24	SDA+LMC	91.03
LFDA+SVM	88.79	LFDA+SRC	87.05	LFDA+LMC	88.69
PCA+SVM	92.24	PCA+SRC	87.77	PCA+LMC	87.05
NPE+SVM	89.40	NPE+SRC	83.59	NPE+LMC	89.40
LPP+SVM	89.50	LPP+SRC	89.40	LPP+LMC	89.40
RF+SVM	93.37	RF+SRC	83.79	RF+LMC	90.42
ERF+SVM	92.25	ERF+SRC	93.17	ERF+LMC	92.76
CFER	94.87	ERF+CGMC	93.68		

The performances of all compared classifiers are given in Table.5. It can be observed that most of the accuracies of SVM, SRC and LMC with ERF are higher than that of SVM, SRC and LMC with the compared dimensionality reduction methods, illustrating that ERF is better than the compared dimensionality reduction methods for facial expression recognition. Especially, the accuracies of SRC and LMC with ERF are respectively higher than that of SRC and LMC with RF about 9.38%, and 2.34%, which proves the significant improvement of ERF from RF. Secondly, by using the same ERF, the accuracies of CFER and CGMC are much higher than that of SVM, SRC and LMC, illustrating that the classifiers based on cognitive gravity model are much better than the compared classifiers for facial expression recognition. Finally, CFER outperforms CGMC with ERF about 1.19%, showing that ECGM is better than CGM.

4.4.3 Compared with recent advances

In Table 6, we compare the performance of CFER on CK+ to the other algorithms in the literature, as CK+ is popular used by these algorithms in the literature[15, 24~29,31-36,53-61,64,65]. The performances of these algorithms are directly cited from its original references. It can be seen from Table.6 that the best performance is 99.60%, which is obtained by Ref.[61]. However, in this performance, the 10-FOLD(SD) is used, and then the samples of a testing subject may be also used for training, which may be difficultly used in real-world environments. Obviously, other performances with the 10-FOLD(SD) also have the same problems.

To overcome the above problem, we evaluate our method by LOSO. Compared with other performances with LOSO, Ref.[56], Ref.[32] Ref.[64], and Ref.[33] have similar results with our methods'. However, the method represented by Ref.[56] and Ref.[64] are two dynamic facial expression recognition method, which have a greater limit when it is used in the real environment, for example, the head angle cannot be changed over a period of time; and the methods represented in Ref.[32] and Ref.[33] are only evaluated by the subset of CK+ with only 6 classes. Furthermore, both LOSO and 10-FOLD(SI) are subject independent, so that the performance of our method also should be compared with the performances with 10-FOLD(SI). The best performance on CK+ with 10-FOLD(SI) is 96.40% when 8 classes are recognized and is 98.30% when 6 classes are recognized, which is obtained by Ref.[58]. This performance may be similar with our performance when 7 class are recognized. However, convolutional neural networks (CNN) is used for feature extraction in this performance, which need too many convolution operation, and then this method is much slower than our method.

Table 6 THE PERFORMANCES OF RECENT ADVANCES ON CK+, WHERE 10-FOLD(SD) REPRESENTS THAT ALL SAMPLES DIVIDED INTO TEN EQUAL SIZE, 10-FOLD(SI) REPRESENTS THAT ALL SAMPLES DIVIDED INTO 10 PERSON-INDEPENDENT SUBSETS BY SAMPLING IN ID ASCENDING ORDER WITH STEP SIZE EQUALS 10, SD REPRESENTS SUBJECT DEPENDENT, AND SI REPRESENTS SUBJECT INDEPENDENT.

Ref	Classes	Dynamic	Measure	Using Landmarks	Accuracy
[24]	6	NO	LOSO	68 Face landmarks	95.37
[25]	6	YES	LOSO	32 Face landmarks	76.00
[26]	6	NO	LOSO	NO	90.92
[27]	6	YES	LOSO	24 Face landmarks	95.10
[28]	7	NO	LOSO	51 Face landmarks	86.30

[29]	6	NO	LOSO	3 Face landmarks	91.59
[15]	7	NO	LOSO	68 Face landmarks	78.10
[31]	8	NO	LOSO	Eye Center	87.40
[32]	6	NO	LOSO	68 Face landmarks	96.53
[33]	6	NO	LOSO	3 Face landmarks	97.13
[34]	6	NO	LOSO	NO	89.21
[35]	7	NO	LOSO	NO	90.50
[36]	7	NO	LOSO	68 Face landmarks	89.50
[53]	7	YES	10-FOLD(SD)	68 Face landmarks	97.10
[54]	7	NO	10-FOLD(SI)	Eye center	93.62
[56]	7	YES	LOSO	NO	97.20
[57]	7	NO	10-FOLD(SI)	49 Face landmarks	97.25
[58]	8	NO	10-FOLD(SI)	NO	96.40
[58]	6	NO	10-FOLD(SI)	NO	98.30
[59]	7	NO	LOSO	44 Face landmarks	90.50
[60]	6	NO	10-FOLD(SD)	NO	94.09
[61]	7	NO	10-FOLD(SD)	NO	99.60
[64]	7	YES	10-FOLD(SI)	NO	95.66
[65]	7	YES	10-FOLD(SD)	NO	96.83
CFER	7	NO	LOSO(SI)	Eye Center	97.66
CFER	7	NO	LOSO(SI)	NO	94.87

4.5 Parametric Analysis for CFER

Generally, it is ideal that a classifier is not sensitive to its parameters. In this subsection, we analyze the sensitivity of parameters of CFER by the experiments on JAFFE and CK+ with LOSO, as in most applications of facial expression recognition, the samples of testing subjects cannot be used for training. The input faces of algorithms in experiments with JAFFE is detected by Adaboost and that with CK+ are obtained by two ways described in the subsection 5.4, where the input faces of “CK+landmarks” are obtained by landmarks, and the input faces of “CK+Adaboost” are obtained by Adaboost. Four parameters are needed for CFER, such as the trade-off parameters a_1 and a_2 , the mass parameter K , and the classifier parameter k .

We first test two trade-off parameters a_1, a_2 . The testing range of a_1, a_2 is from 0.2 to 1.0 with the step 0.2, and the values of K and k are fixed as $K=30$ and $k=9$ that are selected from the above experiments. The testing results are presented in Fig.9. It can be seen from the first column of Fig.9 that the max accuracy is 96.31% and the min accuracy is 82.82%, however, the min accuracy is 94.98% when a_1 is bigger than 0.2. Secondly, the second column shows that the max accuracy is 77.00% and the min accuracy is 75.59% when a_1, a_2 is set to $a_1 > 0.4, a_2 > 0.4$. Thirdly, the third column shows that the max accuracy is 97.55% and the min accuracy is 97.04% when a_1, a_2 is set to $a_1 \geq a_2$. Fourthly, the fourth column shows that the max accuracy is 95.01% and the min accuracy is 94.01% when a_1, a_2 is set to $a_1 \geq a_2$. Finally, it can be seen from Fig.9 that the highest accuracies are nearly obtained by setting the same value for a_1 and a_2 . These indicate that CFER is not much sensitive to a_1 and a_2 .

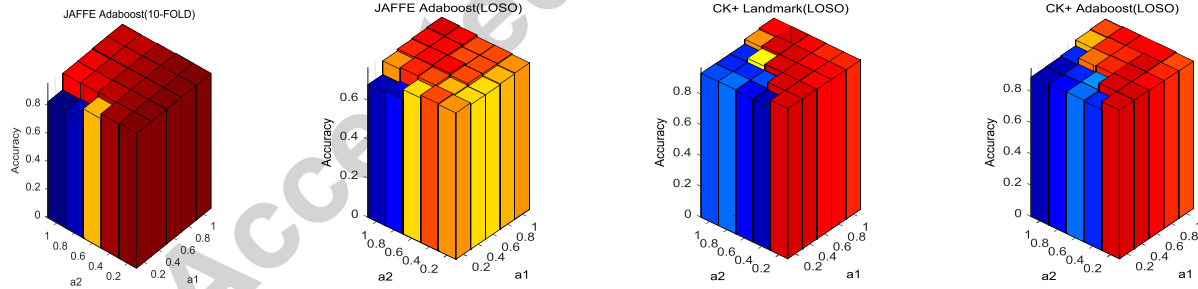


Fig.9 The performances of testing parameters a_1, a_2 for CFER

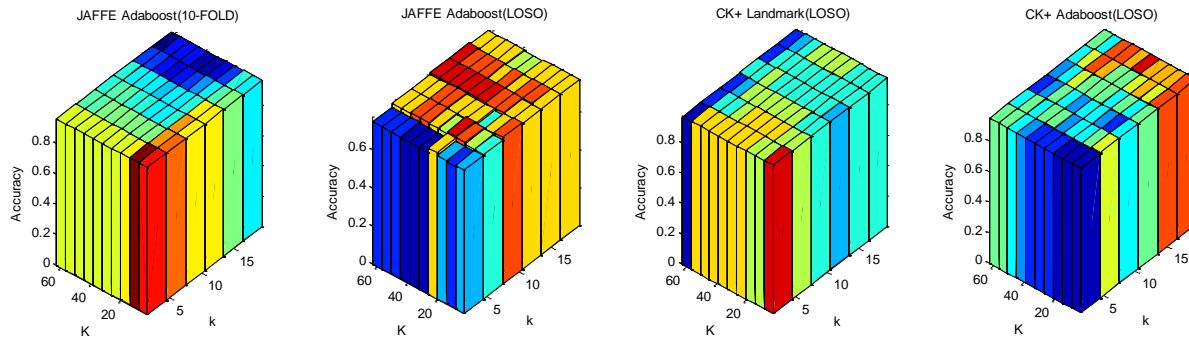


Fig.10 The performances of testing parameters K and k for CFER

Now we turn to test the mass parameter K and the classifier parameter k . The testing range of K is from 6 to 60 with step 6, the testing range of k is from 3 to 18 with step 3, $a_1 = 1, a_2 = 1$ on JAFFE, and $a_1 = 0.2, a_2 = 0.2$ on CK+, which is set according to Fig.9. It can be seen from Fig.10 that CFER is not much sensitive to K and k . For example, most accuracies of experiments on JAFFE with 10-FOLD distribute from 95.60% to 96.63%, and most accuracies of experiments on JAFFE with LOSO distribute from 76.50% to 77.46% when $k \geq 6$. The max, min and mean accuracies of experiments on “CK+landmark” are 97.86%, 97.15% and 97.50% respectively. The max, min, and mean accuracies of experiments on “CK+Adaboost” are 95.41%, 94.60% and 94.98% respectively. The difference between them is much small. These indicate that CFER is also not much sensitive to K and k .

4.6 Time cost Analysis

Although the computational complexities of proposed algorithms have been given in the section 3, the time costs of the algorithms are also simply described in this subsection, as the time cost of algorithms can show a more clear result for the execution speed of the algorithm. Algorithms were implemented in a Matlab environment on a PC platform with a Intel Core i5-5300U Duo CPU with a clock speed of 2.30GHz, 4GB RAM, and Windows 7.

Table 7 AVERAGE TIME COST OF THE TRAINING PROCESSING OF EACH COMPARED DIMENSIONAL REDUCTION ALGORITHM, WHICH IS MEASURED IN SECONDS.

	JAFFE (10-fold)	JAFFE (LOSO)	CK+ (Adaboost)	CK+ (Landmark)
LDA	0.048	0.049	0.976	0.610
SDA	0.093	0.094	1.932	0.172
LFDA	0.154	0.160	3.865	3.480
PCA	0.049	0.051	0.618	0.628
NPE	0.548	0.531	6.125	6.133
LPP	0.085	0.088	1.373	1.368
RF	0.178	0.168	12.631	12.591
ERF	0.213	0.221	13.685	13.700

As we all know, the compared dimensional reduction algorithms contain the training processing and the testing processing, where the testing processing of these algorithms are the same, so that only the average time cost of the training processing is compared in this subsection. The average time cost of the training processing of each dimensional reduction algorithm in each fold is given in Table 7. It can be seen from Table 7 that the average time costs of NPE are the biggest on JAFFE, and the average time costs of ERF are the biggest on CK+, so that the time cost of ERF are increased more quickly with the number of training samples than that of other dimensional reduction methods. However, in real application, the training processing only needs to be perform offline one time, and then we are able to accept that the training processing performs slower.

Table 8 AVERAGE TIME COST OF EACH COMPARED CLASSIFIER, WHICH IS MEASURED IN SECONDS.

	JAFFE (10-fold)	JAFFE (LOSO)	CK+ (Adaboost)	CK+ (Landmark)
SVM(Training)[45]	0.495	0.498	3.556	3.556
SVM(Testing)[45]	0.078	0.082	0.178	0.171
SRC[46]	1.346	1.253	3.483	3.482
LMC[47]	0.033	0.036	0.074	0.076
CGMC[22]	0.047	0.048	0.135	0.137
CFER	0.053	0.052	0.141	0.142

The average time cost of each compared classifier is given in Table 8. It can be seen from Table 8 that the time cost CFER is only bigger than that of LMC and CGMC, and much smaller than that of SRC, they prove that CFER can quickly recognize the facial expression.

5. Conclusion

We have developed a novel hybrid approach named CFER for the facial expression recognition. The results demonstrate the increased performance of the proposed approach over existing methods. The primary advantages of CFER are its flexibility and adaptability derived from its integration of two new techniques. The first technique is the newly enhanced dimensionality reduction method which can extract some density information from the training samples. Training the classifier on data dimensionally reduced by this approach yields a good classifier for the testing data. The second technique is the newly designed classifier that can further make full use of the density information provided by the first technique to more reasonably define masses for cognitive gravity model.

We benchmark our method against several other popular algorithms. Our approach significantly outperformed those popular machine learning algorithms in facial expression recognition. It is found that the accuracy of CFER whose input faces are obtained by landmarks are much higher than that of CFER whose input faces are obtained by Adaboost on database CK+, where the landmarks are only used for aligning faces. However, the landmarks of faces cannot be exactly located in many real applications for facial expression recognition. In the future, we will design a new method to more exactly locate the landmarks and align the faces so as to further improve the performance of CEFR for facial expression recognition.

Acknowledgment

This work was supported by China National Science Foundation under Grants 60973083, 61273363, State Key Laboratory of Brain and Cognitive Science under grants 08B12, Jiaying National Science Foundation under Grants 2016AY13013.

REFERENCES

- [1] Tawari A, Trivedi, M.M, Face Expression Recognition by Cross Modal Data Association, *IEEE Transactions on Multimedia*. vol.15, no.7, pp.1543–1552, Nov.2013.
- [2] Kaimin Yu,Zhiyong Wang, Markus Hagenbuchner, David Dagan Feng, Spectral embedding based facial expression recognition with multiple features, *Neurocomputing*, vol.129, no.10, pp.136–145, 2014.
- [3] Niki Aifanti,Anastasios Delopoulos, Linear subspaces for facial expression recognition, *Signal Processing: Image Communication*, 29(1)(2014)177–188.
- [4] LUO Yuan, WU Cai-ming, ZHANG Yi, Facial expression feature extraction using hybrid PCA and LBP, *The Journal of China Universities of Posts and Telecommunications*, 20(2)(2013)120–124.
- [5] Yogachandran Rahulamathavan, Raphael C.-W. Phan, Jonathon A. Chambers, David J. Parish, Facial Expression Recognition in the Encrypted Domain Based on Local Fisher Discriminant Analysis, 4(1)(2013)83–92.
- [6] Yi-Han Tu, Chiou-Ting Hsu, Dual Subspace Nonnegative Matrix Factorization for Person-Invariant Facial Expression Recognition, in *proc. International Conference on Pattern Recognition (ICPR)*, (2012)2391–2394.
- [7] Guodong Guo, Rui Guo, Xin Li, Facial Expression Recognition Influenced by Human Aging, *IEEE Trans. Affective Computing*, 4(3)(2013)291–298.
- [8] Woo-han Yun, DoHyung Kim, Chankyu Park, Jaehong Kim, Multi-view Facial Expression Recognition Using Parametric Kernel Eigenspace Method Based on Class Features, in *proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (2013)2689–2693.
- [9] Zizhao Zhang, Yan Yan, Hanzi Wang, Discriminative filter based regression learning for facial expression recognition, in *proc. International Conference on Image Processing (ICIP)*, (2013)1192–1196.
- [10] Rui Xiao,Qijun Zhao, David Zhang, Pengfei Shi, Facial expression recognition on multiple manifolds, *Pattern Recognition*, vol.44, no.1, 2011, pp. 109–116.
- [11] Shuai Liu, Qiqui Ruan, Orthogonal Tensor Neighborhood Preserving Embedding for facial expression recognition, *Pattern Recognition*, vol.44, no.7, 2011, pp. 1497–1513.
- [12] Tan Dat Nguyen, Surendra Ranganath, Facial expressions in American sign language: Tracking and recognition, *Pattern Recognition*, vol.45, no.5 , pp. 1877–1891, 2012.
- [13] Anastasios Maronidis,Dimitris Bolis, Anastasios Tefas,Ioannis Pitas, Improving subspace learning for facial expression recognition using person dependent and geometrically enriched training sets, *Neural Networks*, 24(8)2011, pp. 814–828.
- [14] Yeongjae Cheon, Daijin Kim, Natural facial expression recognition using differential-AAM and manifold learning, *Pattern Recognition*, vol.42, no.7, 2009, pp. 1340–1350.
- [15] Shaohua Wan, J.K. Aggarwal, Spontaneous facial expression recognition: A robust metric learning approach, *Pattern Recognition*, vol.47, no.5, pp. 1859–1868, 2014.
- [16] Fei Long, Tingfan Wu, Javier R. Movellan, Marian S. Bartlett, Gwen Littlewort, Learning spatiotemporal features by using independent component analysis with application to facial expression recognition, *Neurocomputing*, vol.93, no.15, pp. 126–132, 2012.
- [17] Xuesong Wang , Yang Gao, Yuhu Cheng, “A non-negative sparse semi-supervised dimensionality reduction algorithm for hyperspectral data,” *Neurocomputing*, vol. 188, no. 5, pp. 275–283, May. 2016.
- [18] Khadijeh Sadatnejad, Saeed Shiry Ghidary, “Kernel learning over the manifold of symmetric positive definite matrices for dimensionality reduction in a BCI application,” *Neurocomputing*, vol. 179, no. 29, pp. 152–160, Feb. 2016.
- [19] Fanhua Shang, L.C. Jiao, Yuanyuan Liu, “Orthogonal component analysis: A fast dimensionality reduction algorithm,” *Neurocomputing*, vol. 177, no. 12, pp. 136–146, Feb. 2016
- [20] Wei Jia, Wen Guihua, Wang Wenfeng, Wang Jiabing, “Local reconstruction and global preserving based semi-supervised dimensionality reduction algorithm,” *computer science*, vol. 38, no. 8, pp. 201–204, Aug. 2011
- [21] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, Pan Yunhe, “A Multimedia Retrieval Framework Based on Semi-Supervised Ranking and Relevance Feedback,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, Apr. 2012
- [22] Guihua Wen, JiaWei, JiabingWang, TiangangZhou, L.Chen, "Cognitive gravitation model for classification on small noisy data," *Neurocomputing*, vol.118, no.22, pp.245–252, Nov.2013.
- [23] Lizhi Peng, Bo Yang, Yuehui Chen, Ajith Abraham, "Data gravitation based classification," *Information Sciences*, vol.179, no.6, pp.809–819, Mar.2009.
- [24] Sezer Ulukaya, Cigdem Eroglu Erdem, Gaussian mixture model based estimation of the neutral face shape for emotion recognition, *Digital Signal Processing*, vol.32, pp.11–23, Sept.2014.
- [25] M. Taner Eskiil, Kristin S. Benli, Facial expression recognition based on anatomy, *Computer Vision and Image Understanding*, vol.119, pp.1–14, Feb.2014.
- [26] Zheng Zhang, Chao Xu, Jiaxin Wang, Xiangning Chen. Facial Expression Recognition Based on MB-LGBP Feature and Multi-level Classification. *Advances in Multimedia, Software Engineering and Computing*, vol.2, pp 37-42, 2011.
- [27] Yongqiang Li, Shangfei Wang, Yongping Zhao, Qiang Ji, Simultaneous Facial Feature Tracking and Facial Expression Recognition, *IEEE Transactions on Image Processing*, vol.22, no.7, pp.2259–2273, Jul.2013.
- [28] Ziheng Wang, Shangfei Wang, Qiang Ji, Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition, In *Proc.International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3422–3429, June.2013.
- [29] Mohammadi M.R, Fatemizadeh. E, Mahoor M.H, Simultaneous Recognition of Facial Expression and Identity via Sparse Representation, In *Proc.Winter Conference on Applications of Computer Vision (WACV)*, pp.1066–1073, 2014.
- [30] Xiao-Hu Wang, An Liu, Shi-Qing Zhang. New facial expression recognition based on FSVM and KNN. *Optik - International Journal for Light and Electron Optics*, vol.126, no.21, pp.3132–3134, Jul.2015.
- [31] Mohammadi M.R, Fatemizadeh.E, Fuzzy local binary patterns: A comparison between Min-Max and Dot-Sum operators in the application of facial expression recognition, In *Proc. Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp.315–319, 2013.
- [32] Sadeghi H, Raie.A.-A, Mohammadi M.R, Facial Expression Recognition Using Geometric Normalization and Appearance Representation, In *Proc. Iranian Conference on Machine Vision and Image Processing (MVIP)*, pp.159–163, 2013.
- [33] Mohammadi M.R, Fatemizadeh. E, Mahoor M.H, PCA-based dictionary building for accurate facial expression recognition via sparse representation, *Journal of Visual Communication and Image Representation*, vol.25, no.5, pp.1082–1092, Jul.2014.
- [34] Sima Taheri, Vishal M. Patel, Rama Chellappa, Component-Based Recognition of Faces and Facial Expressions, *IEEE Transactions on Affective Computing*, vol.4, no.4, pp.360–371, Oct-Dec.2013.
- [35] Seung Ho Lee, Konstantinos N. Plataniotis, Intra -class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition, *IEEE Transactions on Affective Computing*, vol.4, no.4, pp.340–351, 2014.
- [36] Fu-Song Hsu, Wei-Yang Lin, Tzu-Wei Tsai, Facial expression recognition using bag of distances, *Multimedia Tools and Applications*, vol.73, no.1, pp.309–326, Nov.2014.

- [37] Lyons Michael J, Budynek J, Akamatsu S, Automatic classification of single facial images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.21, no.12, pp.1357–1362, Dec.1999.
- [38] Lucey P, Cohn J.F, Kanade T, Saragih J, Ambadar Z, Matthews I, The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.94–101, June.2010.
- [39] Dalal N, Triggs B, Histograms of oriented gradients for human detection, In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.886–893, June.2005.
- [40] Yan Ouyang, Nong Sang, Rui Huang, Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers, *Neurocomputing*, vol.149, no.3, pp.71–78, Feb.2015.
- [41] Martinez AM, Kak AC, “PCA Versus LDA,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb.2001
- [42] Deng Cai, Xiaohei He, Jiawei Han, Semi-supervised Discriminant Analysis, In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp.1–7, Oct.2007.
- [43] Xiaohei He, Deng Cai, Shuicheng Yan, Hong-Jiang Zhang, Neighborhood preserving embedding, In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp.1208–1213, Oct.2005.
- [44] Xiaohei He, Partha Niyogi, “Locality Preserving Projections,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, pp. 585–591, 2003.
- [45] Chih-Chung Chang, Chih-Jen Lin, LIBSVM -- A Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, 2(3)(2011)1–27.
- [46] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, Yi Ma, “Robust Face Recognition via Sparse Representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
- [47] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, Yi Ma, “A local mean-based nonparametric classifier,” *Pattern Recognition Letters*, vol. 27, no. 10, pp. 1151–1159, July. 2006.
- [48] Julien Mairal, “Stochastic Majorization Minimization Algorithms for Large Scale Optimization,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2283–2291, 2013.
- [49] Paul Viola, Michael J. Jones, “Robust Real-Time Face Detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May. 2004.
- [50] Tzimiropoulos G, Pantic M, Optimization Problems for Fast AAM Fitting in-the-Wild, In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp.593–600, Dec.2013.
- [51] Zheng Zhang, Chi Fang, Xiaoqing Ding. Facial expression analysis across databases. *International Conference on Multimedia Technology*, pp.317–320
- [52] Zhan Wang, QiuqiRuan, GaoyunAn. Facial expression recognition using sparse local Fisher discriminant analysis, *Neurocomputing*, vol.174, no.22, pp.137–154, Jan.2016
- [53] XiaorongPu, KeFan, XiongChen, LupingJi, ZhihuZhou. Facial expression recognition from image sequences using two fold random forest classifier, *Neurocomputing*, vol.168, no.30, pp.1173–1180, Nov.2015.
- [54] Mengyi Liu, Shaoxin Li, Shiguang Shan, Xilin Chen. AU-inspired Deep Networks for Facial Expressi on Feature Learning, *Neurocomputing*, vol.159, no.2, pp.126–136, Jul.2015.
- [55] Kaili Zhao, Honggang Zhang, Zhanyu Ma, Yi-Zhe Song, Jun Guo. Multi-label learning with prior knowledge for facial expression analysis, *Neurocomputing*, vol.157, no.1, pp.280–289, Jun.2015.
- [56] Yimo Guo, Guoying Zhao, Matti Pietikäinen. Dynamic Facial Expression Recognition With Atlas Construction and Sparse Representation. *IEEE Transactions on Image Processing*.
- [57] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, Junmo Kim. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. *IEEE International Conference on Computer Vision*, pp.2983–2991, 2015.
- [58] Pooya Khorrami Tom Le Paine Thomas S. Huang. Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition. *IEEE International Conference on Computer Vision Workshops*, pp.19–27, 2015
- [59] Seung Ho Lee, Yong Man Ro. Partial Matching of Facial Expression Sequence Using Over-complete Transition Dictionary for Emotion Recognition. *IEEE Transactions on Affective Computing*, inPress, 2016.
- [60] S L Happy, Aurobinda Routray. Automatic Facial Expression Recognition Using Features of Salient Facial Patches. *IEEE Transactions on Affective Computing*, vol.6, no.1, 2015.
- [61] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, DeXpression: Deep Convolutional Neural Network for Expression Recognition. *arXiv:1509.05371*, 2015
- [62] Jian Yang, A. F. Frangi, Jing-Yu Yang, David Zhang, Zhong Jin. KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.2, pp.230–244, Feb.2005.
- [63] Jinghua Wang, Jane You, Qin Li, Yong Xu. Orthogonal discriminant vector for face recognition across pose, *Pattern Recognition*, vol.45, no.12, pp.4069–4079, Dec.2012
- [64] S. Elaiwat, M.Bennamoun, F.Boussaid A. spatio-temporal RBM-based model for facial expression recognition. *Pattern Recognition*, vol.49, pp.152–161, Jan.2016
- [65] Muhammad Hameed Siddiqi, Rahman Ali, Adil Mehmood Khan, Young-Tack Park, Sungyoung Lee. Human Facial Expression Recognition Using Stepwise Linear Discriminant Analysis and Hidden Conditional Random Fields. *IEEE Transactions on Image Processing*, vol.24, no.4, APR.2015.
- [66] Birkan Tunc, VolkanDağlı, MuhittinÇokmen. Class dependent factor analysis and its application to face recognition. *Pattern Recognition*, vol.45, pp.4093–4101, Jan.2016
- [67] Kaimin Yu,Zhiyong Wang, Li Zhuo, Jiajun Wang, Zheru Chi, Dagan Feng. Learning realistic facial expressions from web images. *Pattern Recognition*, vol.46, no.8, pp.2144–2155, Aug.2013.
- [68] Ying Tong,Rui Chen,Yong Cheng. Facial expression recognition algorithm using LGC based on horizontal and diagonal prior principle. *Optik - International Journal for Light and Electron Optics*, vol.125, no.16, pp.4186–4189, Aug.2014.
- [69] Thiago H.H. Zavaschi, Alceu S. Britto Jr., Luiz E.S. Oliveira, Alessandro L. Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, vol.40, no.2, pp.646–655, Feb.2014.

Yaxin Sun is a Ph.D. candidate at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received his M.Sc. degree in Computer Application Technology from Guangzhou University of Technology in 2012. His current research interests include speech emotion recognition, pattern recognition and machine learning.

Guihua Wen, Ph.D., professor, doctor candidate supervisor. His main research interests are affective computing, machine learning and data mining, and cognitive computing. He has published papers in journal such as *Pattern Recognition*, *Neurocomputing*, *Journal of Software*, and international conferences such as *IJCAI*. He directed the projects from the China National Natural Science Foundation, State Key Laboratory of Brain and Cognitive Science etc al. He also focuses on developing affective computing open platform and the related applications. He has ever been a Council Member of Chinese Association for Artificial Intelligence and program committee member of many international conferences. He is also a reviewer for China National Natural Science Foundation