ENHANCED DICTIONARY PAIR LEARNING SPARSE REPRESENTATION MODEL FOR FACIAL EXPRESSION CLASSIFICATION

Jianquan Gu, Haifeng Hu, and Siyue Xie

School of Electronic and Information Engineering, Sun Yat-sen University, Guangzhou, China

ABSTRACT

Facial expression recognition (FER) is a challenging task in the community of affect analysis and pattern recognition. In this paper, we propose a novel framework, namely Enhanced Dictionary Pair Learning Sparse Representation (EDPLSR), for facial expression recognition. The key idea behind our model is that it jointly learns a synthesis dictionary as well as an analysis dictionary, which require that all coding vectors should be group sparse. Furthermore, inspired by the observation that the geometrical information of the data is discriminative, a manifold regularization term is introduced to obtain smoothly vary sparse representations along the geodesics of data manifold. This is distinctive from most of the existing approaches which fail to consider the geometrical structure of data space. The experimental results demonstrate the effectiveness of our method.

***Index Terms*—**dictionary pair learning, manifold regularization, sparse representation, facial expression recognition

1. INTRODUCTION

Facial expressions play a significant role in reflecting one's emotions. Many previous researches on expression analysis followed the work of Ekman *et al.* [1] and aimed to set up a model that could automatically classify different expressions. In these years, sparse representation classification (SRC) has been widely applied to solve the problem of facial expression recognition (FER) and performed well in experiments [2-8]. Most of existing dictionary learning methods focus on improving the structure of the dictionary. Some learn a shared dictionary for all class while others learn class-specific dictionaries [9-11]. Some researchers concentrate on enhancing the discrimination of the coding coefficients. Recently, learning a pair of dictionaries [12-13], which include a synthesis dictionary and a projective analysis dictionary, seems to work well in image classification. The method of dictionary pair learning (DPL), proposed by Gu *et al.* [12], can effectively reduce the time complexity as well as achieve competitive performance by using an extra projective analysis dictionary.

In this paper, we present a dictionary pair learning based framework, called Enhanced Dictionary Pair Leaning Sparse Representation (EDPLSR), for solving FER problem. Fig.1 shows the flowchart of our method. Training images are first aligned, and fifteen landmarks points will be located. As LBP can effectively describe the textures and shape changes in an image, we apply it to employing feature extraction. Two structured dictionaries, the so-called analysis dictionary and synthesis dictionary, are implemented to represent the query sample. We can generate discriminative codes and reconstruct the query face image in a specific class respectively. Inspired by the progress in manifold learning, a manifold regularization term is introduced to the pair dictionaries learning, which improves the ability on discrimination. Therefore, we can obtain a smoothly vary representation along the geodesics of the data manifold. This process can also optimize dictionary's structure and is more suitable for face expression classification. In classification stage, we utilize the dictionary pair as well as the coding coefficients to enhance the recognition accuracy. Experimental results demonstrate the effectiveness of our proposed method. Specifically, the contributions of our model can be summarized as follows:

1) By introducing a manifold regularization term, coding vectors with group sparsity can be obtained by EDPLSR. This means our model holds the properties of both individual sparsity as well as local similarity. Moreover, the group sparsity of coding vectors can effectively improve the recognition rate of FER.

2) Compared with existing sparse coding approaches, EDPLSR extends the discriminative synthesis dictionary learning framework. Moreover, it considers the geometric structure of the data space, which is important for discrimination.

The rest of this paper is organized as follows. Section 2 describes the EDPLSR method. Section 3 presents the experimental results on three facial expression databases. Finally, the conclusion of this paper is given in Section 4.

2. THE PROPOSED METHOD

2.1. Expressional Image Representation

To represent a facial expression image, we first detect the face region by the famous Viola-Jones face detector [14] and then locate fifteen landmarks (shown in Fig.2) using an

This work was supported in part by the NSFC, under Grant no. 61673402, the NSF of Guangdong (grant nos. 2014A030313173 and 2016B010109002), and by the Science and Technology Program of Guangzhou (grant no. 201704020180), and the fundamental Research Funds for the Central Universities of China.
Corresponding author: Haifeng Hu.

efficient face alignment algorithm [15]. High-dimensional LBP features [16] are extracted in the regions around these landmarks. Specifically, a fixed-size patch is cropped with 5 different scales. Each patch is divided into 4×4 cells, where a 59-dimensional uniform local binary pattern [17] is extracted from each cell. Features extracted from the same image are concatenated together as a descriptor. Thus the feature dimension for each image is 4425.

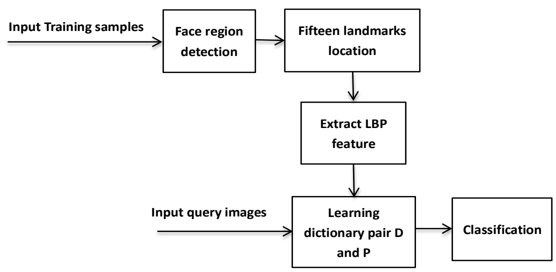

Fig.1 *The flowchart of the proposed method*

Fig.2 *The location of fifteen landmarks in expression face image*

2.2. Dictionary Pair Learning Model

The model of dictionary pair learning (DPL) is a dictionary learning algorithm for sparse representation. Denote $X = [X_1, \dots, X_k, \dots, X_K]$ as a p-dimensional training set from K expression classes. $X_k \in R^{p \times N}$ is the subset from class k and N denotes the number of samples in each class. D is the over-completed synthesis dictionary and A is the sparse coding coefficient matrix of X over D.

DPL additionally learns an analysis dictionary P to represent A by analytically code X, i.e., $A = PX$, where $P \in R^{NK \times p}$. The framework of the modified model can be formulated as:

$$\{P^*, D^*\} = arg\, min_{P,D,A} \sum_{k=1}^{K} ||X_k - D_k P_k X_k||_F^2 + \lambda ||P_k \bar{X}_k||_F^2, s.t. \|d_i\|_2^2 \leq 1. \quad (1)$$

$\bar{X}_k$ is the complementary set of X_k in the overall training set X, $D_k \in R^{p \times N}$ and $P_k \in R^{N \times p}$ are the k-th class sub-dictionary of the analysis dictionary $P = [P_1; \dots; P_k; \dots; P_K]$ and synthesis dictionary $D = [D_1, \dots, D_k, \dots, D_K]$ respectively. Using the sub-dictionary P_k in P, samples from class i ($i \neq k$) can be projected to a nearly null space, i.e.,

$$P_k X_i \approx 0, \quad \forall k \neq i \quad (2)$$

We can reconstruct X_k from its projective code matrix $P_k X_k$ according to sub-dictionary D_k, which minimizes the reconstruction error:

$$min_{P,D} \sum_{k=1}^{K} ||X_k - D_k P_k X_k||_F^2 \quad (3)$$

Thus the dictionary pair learning model can be written as:

$$\{P^*, D^*\} = arg\, min_{P,D,A} \sum_{k=1}^{K} \left|\left|X_k - D_k A_k\right|\right|_F^2 + \tau ||P_k X_k - A_k||_F^2 + \lambda ||P_k \bar{X}_k||_F^2, s.t. \|d_i\|_2^2 \leq 1. \quad (4)$$

where d_i is the i-th atom of the sub-dictionary D_k.

2.3. The proposed EDPLSR Model

In DPL model, the items of D are considered as basic vectors in projective space and columns of A are the new representation of the corresponding data point in the obtained space. Let x_i and x_j be two projection points of $\mathrm{a_i}$ and $\mathrm{a_j}$, i.e., two columns of A, in the new space. The manifold assumption [18-19] assumes x_i and x_j should be close to each other if they are ambient in the original space. Motivated by the assumption, we extend the DPL model with manifold properties, which is called Enhanced Dictionary Pair Learning Sparse Coding (EDPLSR) model.

As for a set of data point $x_1, \dots, x_M$ with p-dimension, where M is the number of training samples, we obtain a M vertices nearest neighbor graph G. Each vertex of the graph represents a data point. We denote a matrix W as the weight of G, which can be formulated as:

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ with the same label} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The degree of x_i is defined as $de_i = \sum_{j=1}^{M} W_{ij}$ and let $DE = diag(de_1, \dots, de_M)$.

To map the weighted graph G to the sparse coefficient matrix A, we minimize the following objective function:

$$\frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} (a_i - a_j)^2 W_{ij} = Tr(ALA^T) \quad (6)$$

where $L = DE - W$ is the Laplacian matrix and the Laplacian regularization $Tr(ALA^T)$ can be rewritten as:

$$\begin{aligned} Tr(ALA^T) &= Tr\left(\sum_{i,j=1}^{M} L_{ij} a_i a_j^T\right) \\ &= \sum_{i,j=1}^{M} L_{ij} a_j^T a_i \\ &= \sum_{i,j=1}^{M} L_{ij} a_i^T a_j \end{aligned} \quad (7)$$

By incorporating the manifold regularization term (6) into the dictionary pair learning model, the objective function of our EDPLSR model can be formulated as follow:

$$\{P^*, D^*, A^*\} = arg\, min_{P,A,D} \sum_{k=1}^{K} \left(\left|\left|X_k - D_k A_k\right|\right|_F^2 + \tau \left|\left|P_k X_k - A_k\right|\right|_F^2 + \lambda \left|\left|P_k \bar{X}_k\right|\right|_F^2\right) + \beta Tr(ALA^T) \quad (8)$$

where τ, λ, and β are scalar constants, X_k is the k-th class training samples, D_k and P_k are the sub-dictionary with respect to the dictionary D and P corresponding to the k-th class expressions. All terms are characterized by Frobenius norm, and (8) can be optimized as:

$$\begin{aligned} &\{P^*, D^*, A^*\} \\ &= arg \min_{P,A,D} \sum_{k=1}^{K} \sum_{t=1}^{N} \left(\left|\left|X_{kt} - D_k A_{kt}\right|\right|_F^2 + \tau \left|\left|P_k X_{kt} - A_{kt}\right|\right|_F^2\right) \\ &+ \lambda \sum_{k=1}^{K} \sum_{u=1, u \neq k}^{K} \sum_{t=1}^{N} \|P_k X_{ut}\|_F^2 \\ &+ \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} L_{i,j} A_{k_1 t_1}^T A_{k_2 t_2} \quad s.t. \|d_i\|_2^2 \leq 1. \end{aligned} \quad (9)$$

where $i = (k_1 - 1) \times N + t_1$ and $j = (k_2 - 1) \times N + t_2$.

Algorithm1 Enhanced Dictionary Pair Learning Sparse Representation (EDPLSR)

Input: Training samples for K classes $X = [X_1, \dots, X_k, \dots, X_K]$, parameter λ, τ, β, m,

1: Initialize $D^{(0)}$, $P^{(0)}$ and $A^{(0)}$ as random matrices with unit Frobenious norm, $h = 0$;

2: **While** *stopping criterion has not been met*, **do**

for $k = 1{:}K$ **do**

for $t = 1{:}N$ **do**

Fixing $P_k^{(h)}$ and $D_k^{(h)}$

$$A_{kt}^{(h+1)} \leftarrow \left(D_k^{T(h)} D_k{}^{(h)} + (\tau + \beta L_{ii}) I\right)^{-1} (D_k{}^{(h)} X_{kt} + \tau P_k^{(h)} X_{kt} - \beta \sum_{\substack{j=1 \\ j \neq i}}^{M} L_{ij} A_{kt}^{(h)})$$

end for

Fixing $A_k^{(h)}$

$$P_k^{(h+1)} \leftarrow \tau A_k^{(h)} X_k^T (\tau X_k X_k^T + \lambda \bar{X}_k \bar{X}_k^T + cI)^{-1}$$

$$D_k^{(h+1)} \leftarrow \begin{Bmatrix} min_{D_k^{(h)}} \sum_{k=1}^{K} \left\| X_k - D_k^{(h)} A_k^{(h)} \right\|_F^2, \\ s.t.\ D_k^{(h)} = B,\ \|B_i\|_2^2 \leq 1 \end{Bmatrix}$$

end for

$h \leftarrow h + 1$;

end while

$P \leftarrow P^{(h)}$

$D \leftarrow D^{(h)}$

$A \leftarrow A^{(h)}$

3: **Output**: Analysis dictionary P , synthesis dictionary D, coding coefficient matrix A.

X_{kt} denotes the t-th training sample of the k-th class expressions, A_{kt} is the corresponding code vector.

All terms in (9) are characterized by Frobenius norm. We initialize the analysis dictionary P , synthesis dictionary D and coding coefficient matrix A as random matrices with unit Frobenius norm. To obtain the optimal coefficient, A and $\{D, P\}$ can be optimized alternately by two steps, which can be described as follow:

(1) Fix $D$and P, update A

$$A_{kt}^{*} = (D_k^T D_k + (\tau + \beta L_{ii}) I)^{-1} (D_k X_{kt} + \tau P_k X_{kt} - \beta \sum_{\substack{j=1 \\ j \neq i}}^{M} L_{ij} A_{kt}) \tag{10}$$

where $i = (k-1) \times N + t, k = 1,2, \dots, K, t = 1,2, \dots, N$.

(2) Fix A, update D and P:

$$\begin{cases} P^{*} = arg\ min_P \sum_{k=1}^{K} \tau \|P_k X_k - A_k\|_F^2 + \lambda \|P_k \bar{X}_k\|_F^2 \\ D^{*} = arg\ min_D \sum_{k=1}^{K} \|X_k - D_k A_k\|_F^2,\quad s.t.\ \|d_i\|_2^2 \leq 1 \end{cases} \tag{11}$$

The first function in (11) aims to promote the discriminative ability of the dictionary P. The lower one focuses on minimizing the reconstruction error with the help of the coding coefficients. Balance between the ability of representation and discrimination of our model will be met as the minimization process converges. Thus we can express the close-form solution of P as:

$$P_k^{*} = \tau A_k X_k^T (\tau X_k X_k^T + \lambda \bar{X}_k \bar{X}_k^T + cI)^{-1} \tag{12}$$

where$c = 10e^{-4}$ is a small constant. The D problem can be optimized by introducing a variable B:

$$min_{\mathrm{D,B}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \|\mathrm{X_k} - \mathrm{D_k A_k}\|_{\mathrm{F}}^2,\quad \mathrm{s.t.\ D = B}, \|\mathrm{B_i}\|_2^2 \leq 1 \tag{13}$$

whereB_i is the$i$-th atom of B.

The optimal solution of (13) can be obtained utilizing the Alternating Direction Method of Multipliers (ADMM) algorithm [20]. In each step of optimization, we can obtain the solutions for A and P , and the ADMM-based optimization of D converges. We summarize the general algorithm of EDPLSR model in Algorithm 1, where m is the number of the sub-dictionary atoms of each class. The stopping criterion is set as 0.01.

2.4. Classification Scheme

In the training stage, a pair of sub-dictionaries D_k and $P_k$are learnt for each class according to (10). The synthesis sub-dictionary D_k is specially used to reconstruct the k-th class training set X_k , which means that the residual $\|X_k - D_k P_k X_k\|_F$ should smaller than $\|X_i - D_k P_k X_i\|_F$, where $i \neq k$.Therefore, a test sample from class k, denoted as y, can be well reconstructed by D_k and P_k with smallest residual. Thus we can categorize y to the correct class by the reconstruction residual of the following classification scheme:

$$class(y) = arg\ min_k \|y - D_k P_k y\|_2 \tag{14}$$

3. EXPERIMENTAL RESULTS

In this section, the performance of our EDPLSR method is evaluated using two public facial expression databases, including Cohn-Kanade (CK) [21] and Extended Cohn-Kanade dataset (CK+) [22]. We compare the proposed method with three classical methods including SRC [2], CRC [23] and DPL [12].

In all of the following experiments, for each image, the face region is detected automatically by Viola's face detector. We then locate fifteen landmarks [16] for extracting LBP features [24]. To guarantee robustness and simplicity, the regularization parameters λ, τ, β in Eq. (10) are fixed as $\lambda = 0.003$, $\tau = 0.05$, $\beta = 0.001$. For all dictionary pair learning, we fix the number of sub-dictionary atoms of each class as $m = 30$.

3.1 Cohn-Kanade Database

The Cohn-Kanade database contains image sequences of 97 subjects. Six among these images are the basic emotions. In our experiments, 373 sequences are selected. Each sequence is labeled as one of the six basic emotional expressions. In Table I, we list the number of selected sequences in each expression category. For each sequence, we choose the last three frames with the most intense expressions. That is, we actually collect 1119 (373×3) facial images for training samples and testing samples in total. Fig. 3 shows some samples with six expressions in the Cohn–Kanade database.

Table II shows the confusion matrix of six emotions using our method. The number of training sequences of each expression is 13, and the rest are used for testing. From the results, surprise expression achieves the best recognition rate which is characterized by open mouth and upward eyebrow movement. We can find that the most likely to be confused expressions pair is anger and sadness. The reason may be these two expressions share similar subtle changes and shape deformation in the face. Fig.4 plots the CRR curves compared with three algorithms under different number of training sequences. We can find that the proposed method performs the best, whose recognition performance improves as the number of training sequences increase.

Fig.3 *Example of six basic expressions from the Cohn–Kanade database(anger, disgust, fear, happiness, sadness, and surprise)*

TABLE I STATISTICS FOR THE SEQUENCE OF EACH EXPRESSION IN COHN-KANADE DATABASE USED IN THE EXPERIMENTS

anger	disgust	fear	happy	sadness	surprise
74	101	32	55	38	73

TABLE II CONFUSION MATRIX OF EDPLSR ON COHN-KANADE DATABASE (MEASURED BY RECOGNITION RATE: %)

	anger	disgust	fear	happy	sadness	surprise
anger	91.11	0	0	0	8.89	0
disgust	0	100	0	0	0	0
fear	0	2.63	85.97	7.02	0.88	3.51
happy	0.40	0.79	9.92	86.91	0	1.98
sadness	5.36	2.38	2.98	0	87.50	1.79
surprise	0	0	0	0	0	100

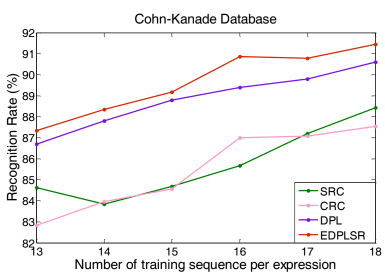

Fig.4 *Performance comparisons on Cohn–Kanade database with different number of training sequences per expression*

3.2. Extended Cohn-Kanade Database

The database of Extended Cohn-Kanade (CK+) consists of 593 sequences collected from 123 subjects. Among these sequences, only 327 sequences are labeled. We choose 309 sequences labeled with six basic expressions. Table III lists the number of sequences in each expression category. Only the last three frames in each sequence, which refer to the most discriminative expression images, are used in our experiment. Fig.6 shows some samples with all expressions in the Extended Cohn-Kanade database.

Table IV shows the confusion matrix of six expressions in the Extended Cohn-Kanade database using our method. The number of training sequences of each expression is 19, and the rest sequences are used for testing. As can be seen, the proposed method performs worst for anger expression and the sadness expression is more likely to be misclassified as anger. This is because anger and sadness are similar in face appearance. As the exaggerated degrees of anger and sadness are not intense as all other expressions, the recognition rate of these two kinds of expressions cannot achieve the best recognition accuracy. Surprise achieves the highest recognition rate for the exaggerated shape of opening the mouth. Fig. 6 shows the performance of the compared algorithms influenced by different number of training sequence per expression. Our method achieves the highest recognition rate, which may be due to the fact that manifold regularization enhances the group sparse property and thus improves the discriminating power of our model.

Fig.5 *Example of six basic expressions from the Extended Cohn–Kanade database*

TABLE III STATISTICS FOR THE SEQUENCE OF EACH EXPRESSION IN EXTENDED COHN-KANADE DATABASE USED IN THE EXPERIMENTS

anger	disgust	fear	happy	sadness	surprise
45	59	25	69	28	83

TABLE IV CONFUSION MATRIX OF EDPLSR ON EXTENDED COHN-KANADE DATABASE (MEASURED BY RECOGNITION RATE: %)

	anger	disgust	fear	happy	sadness	surprise
anger	65.48	2.38	3.57	0	28.57	0
disgust	7.14	85.71	3.17	0	2.38	1.59
fear	4.17	8.33	87.50	0	0	0
happy	0.64	0	5.77	92.95	0	0.64
sadness	9.09	0	9.09	0	72.73	9.09
surprise	0	0.51	3.03	0	3.03	93.43

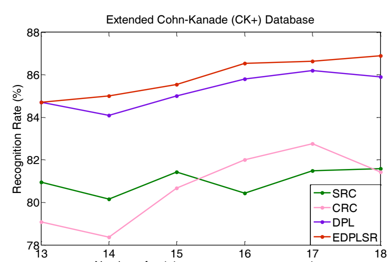

Fig.6 *Performance comparisons on Extended Cohn–Kanade database with different number of training sequences per expression*

4. CONCLUSION

In this paper, we propose a novel enhanced dictionary pair learning sparse representation (EDPLSR) model for facial expression classification tasks. EDPLSR can jointly learn a synthesis dictionary and an analysis dictionary, which works together to obtain better representation and discrimination simultaneously. A manifold regularization term is used to obtain sparse representation which may vary smoothly along the geodesics of the data manifold. Extensive experimental evaluations verify the effectiveness of the proposed model.

5. REFERENCES

[1] Friesen E, Ekman P. Facial action coding system: a technique for the measurement of facial movement, Palo Alto, 1978.

[2] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(2): 210-227.

[3] Zhang S, Zhao X, Lei B. Robust facial expression recognition via compressive sensing. Sensors, 2012, 12(3): 3747-3761.

[4] Wang Z., Huang M., Ying Z., The performance study of facial expression recognition via sparse representation, International Conference on Machine Learning and Cybernetics, 2010, 2: 824-827.

[5] Zafeiriou S, Petrou M. Sparse representations for facial expressions recognition via l_1 optimization, Computer Vision and Pattern Recognition Workshops (CVPRW). 2010: 32-39.

[6] Wang Z, Ying Z. Facial expression recognition based on local phase quantization and sparse representation[C], International Conference on Natural Computation (ICNC), 2012: 222-225.

[7] Bai S, Bai X, Liu W. Multiple stage residual model for image classification and vector compression. IEEE Transactions on Multimedia, 2016, 18(7): 1351-1362.

[8] Tang P, Wang X, Feng B, *et al.* Learning Multi-instance Deep Discriminative Patterns for Image Classification, IEEE Trans. on Image Processing, 2016

[9] Ramirez I, Sprechmann P, Sapiro G. Classification and clustering via dictionary learning with structured incoherence and shared features, CVPR, 2010: 3501-3508.

[10] Yang M, Zhang L, Feng X, et al. Fisher discrimination dictionary learning for sparse representation, ICCV, 2011: 543-550.

[11] Gao S, Tsang I W H, Ma Y. Learning category-specific dictionary and shared dictionary for fine-grained image categorization. IEEE Transactions on Image Processing, 2014, 23(2): 623-634.

[12] Gu S, Zhang L, Zuo W, et al. Projective dictionary pair learning for pattern classification, Advances in Neural Information Processing Systems. 2014: 793-801.

[13] Zeng X, Bian W, Liu W, *et al*. Dictionary Pair Learning on Grassmann Manifolds for Image Denoising, IEEE Transactions on Image Processing, 2015, 24(11): 4556-4569.

[14] Viola P, Jones M J, Robust real-time face detection, International journal of computer vision, 2004, 57(2): 137-154.

[15] Xiong X, De la Torre F, Supervised descent method and its applications to face alignment, CVPR, 2013: 532-539.

[16] Chen D, Cao X, Wen F, *et al*. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, CVPR, 2013: 3025-3032.

[17] Ojala T, Pietikainen M, Maenpaa T, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. on pattern analysis and machine intelligence, 2002, 24(7): 971-987.

[18] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering, NIPS, 2001, 14: 585-591.

[19] Cai D, He X, Wu X, *et al.* Non-negative matrix factorization on manifold, IEEE International Conference on Data Mining, 2008: 63-72.

[20] S. Boyd, N. Parikh, E. Chu, *et al*, Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Machine Learning, 2011, 3(1):1-122.

[21] Kanade T, Cohn J F, Tian Y, Comprehensive database for facial expression analysis, Automatic Face and Gesture Recognition, CVPR, 2000: 46-53.

[22] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, CVPR Workshops. 2010: 94-101.

[23] Zhang L, Yang M, Feng X. Sparse representation or collaborative representation: Which helps face recognition? ICCV, 2011: 471-478.

[24] Chen B C, Chen C S, Hsu W. Review and implementation of high-dimensional local binary patterns and its application to face recognition, Academia Sinica, Tech. Rep. TR-IIS-14-003, 2014.