2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Sparse learning for salient facial feature description*

Yue Zhao[1] and Jianbo Su[1]

***Abstract*— High dimension of the features employed for face recognition is the main reason to slow down the recognition speed. Additionally, selecting salient facial features has significant impact on the efficiency of face recognition. In order to get the sparse and salient facial features, this paper propose a new sparse learning approach for salient facial feature description. This approach is to learn the feature evaluation vector with the training samples composed of within- and between-class distance vector sets. Then, the feature evaluation vector is employed to construct a new model for salient facial feature description. Experimental results show that the proposed method achieves much better face recognition performance with lower feature dimensionality.**

I. INTRODUCTION

As a hot research area of artificial intelligence, face recognition has generated wide applications including person identification [1], video surveillance [2], intelligent camera [3]–[5], and so on. To achieve good face recognition results, the key issue is to afford an efficient facial feature descriptor. In the past couple of years, many descriptors have been presented for face recognition. A hybrid approach, Eigenface [6], is proposed by M.Turk and A.Pentland. It produces features by performing Principal Component Analysis (PCA) to local face regions independently [7]. Based on the theory of Fisher Linear Discriminant Analysis (FLDA), Fisherface [8] finds the optimal projector that maps the original data into a low-dimensional feature space. This process follows the restriction that the ratio of the trace of the between-class scatter to the trace of the within-class scatter is maximized [9]. Elastic Bunch Graph Matching (EBGM) [10] uses the the Gabor filter responding in certain facial landmarks and a graph describing the spatial relations of these landmarks [11] to describe faces. Local Binary Pattern (LBP) is firstly adopted to describe facial features by Ahonen *et al.* [12]. To describe the face feature more efficiently, Ahonen *et al.* assign different weights to different regions based on the psychology research. This leads to good results of face recognition. A three-level operator, Local Ternary Patterns (LTP), is proposed by Tan *et al.* [13]. LTP features are more discriminant and less sensitive to noise. Guo *et al.* present a hierarchical mutliscale LBP (HM-LBP) model [14] to dig out useful information from those “non-uniform” patterns.

*This work was partially financially supported by the Key Project of National Natural Science Foundation of China (NSFC) under grant 60935001

[1]Yue Zhao and Jianbo Su are both with Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240 zhaoyue0609@sjtu.edu.cn; jbsu@sjtu.edu.cn

For face recognition, LTP, HM-LBP and the other region-based features have achieved good results [13], [14]. However, they normally give rise to feature sets of high dimensionality, which essentially slows down the speed of face recognition. In addition, they ignore the different contributions of the features (a feature is an element of the feature vector) in the same region, which leads to features with different salience have the same importance for face recognition in one region. To get the sparse and salient facial features, this paper propose a new sparse learning approach for salient facial feature description based on our previous work [15]. Firstly, a facial feature transformation is presented to generate the training samples composed of within- and between-class distance vector sets. Then, a sparse learning approach is presented to learn the feature evaluation vector with the training samples. Lastly, the feature evaluation vector is employed to construct a new model for salient facial feature description. As the sparse learning approach encourages the sparsity at both the group and individual levels, the new facial feature description model can distinguish the contributions of different dimensions of the features in the same region and thus achieve dimensionality reduction.

The rest of the paper is organized as follows. Regional facial feature descriptors are presented in Section II. Section III describes the general facial feature description models. In Section IV, a sparse learning approach is proposed to evaluate the salience of the facial features. Then, the salient facial feature description model is derived in Section V. Extensive experiments results are employed to illustrate the performance of the proposed method in Section VI, followed by the conclusions in Section VII.

II. REGIONAL FACIAL FEATURE DESCRIPTORS

Regional feature descriptors are the dominant descriptors in face recognition, such as LTP [13] and HM-LBP [14].

A. LTP

The LBP operator takes a local neighborhood around each pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary-valued image patch as a local image descriptor. LBPs threshold exactly at the value of the central pixel, they tend to be sensitive to noise, specially in near-uniform image regions. So LTP descriptor employs 3-valued codes to generate ternary code. Suppose g_c is the gray value of the center pixel of the circular template, and g_e is the gray value of the e^{th}

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

neighborhood, where $e = 0,1,2,\dots,E-1$. The original gray level distribution of the these points T is:

$$T = \{g_c, g_0, g_1, \dots, g_{E-1}\}. \tag{1}$$

The original gray level distribution can be described by the joint difference distribution approximatively, which is:

$$T \approx \{g_0 - g_c, g_1 - g_c, \dots, g_{E-1} - g_c\}. \tag{2}$$

In order to achieve the variance with respect to the scaling of the gray scale, LTP descriptor only considers the signs of the differences instead of their exact values. Thus the original gray level distribution T can be expressed by the following equation:

$$T \approx \{s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{E-1} - g_c)\}. \tag{3}$$

The gray value of the neighborhoods in a zone of width $\pm t$ around g_c are quantized to "0". Ones above this are quantized to "1" and ones below it to "-1". This procedure can be described by the following formula:

$$s(g_e - g_c) = \begin{cases} 1 & if\ g_e - g_c \geq t \\ 0 & if\ |g_e - g_c| < t \\ -1 & if\ g_e - g_c \leq -t \end{cases} \qquad e = 0,1,2,\dots,E-1. \tag{4}$$

where t is a threshold specified manually. In this way, LTP codes are more resistant to noise. Fig. 1 presents the LTP encoding process. LTP descriptor also use the "uniform" pattern to describe the texture pattern in a local area.

B. HM-LBP

LBP descriptor adopts "uniform" patterns to reduce the number of binary patterns, which has been validated to play an important role in face recognition [12]–[14]. Incorporating uniform idea, many patterns, which are not uniform patterns, are clustered into a set of non-uniform patterns. By this way, many discriminant but non-uniform patterns fail to provide useful features, and the percentage of non-uniform patterns increases as the radius increases, which leads to much information lost. Different to LBP descriptor, HM-LBP descriptor uses many circular templates with different sizes to generate uniform pattern. After the LBP map of the biggest radius for each pixel is gotten, the whole pixels are divided into two groups: one group with the uniform pattern, and the other with the non-uniform pattern. For the latter, another circular template with a smaller radius will be employed to extract the LBP patterns. The process will be stopped, if the smallest radius has arisen. [14].

III. General Facial Feature Description Models

For face recognition, descriptors LTP and HM-LBP all statistics the number of each uniform pattern in a local face image area to obtain a set of histogram vectors, and then concatenated them into a whole histogram vector as a facial feature vector. The detailed procedures are as follows.

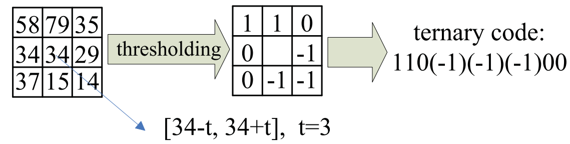

Fig. 1. The theories of LTP coding.

LTP and HM-LBP both divide a face image into r different regions: $D_1, D_2, \dots, D_r$, then an m-dimensional histogram vector is calculated for each region. Lastly, r histogram vectors $H[1], H[2], \dots, H[r]$ compose a $1 \times M$ feature vector to describe the face image I:

$$\begin{aligned} H &= [h_1, \dots, h_j, \dots, h_M] \\ &= [\underbrace{h_1, \dots, h_m}_{H[1]}, \underbrace{h_{m+1}, \dots, h_{2m}}_{H[2]}, \dots, \underbrace{h_{(r-1)m+1}, \dots, h_{rm}}_{H[r]}], \end{aligned} \tag{5}$$

where $M = r \times m$, and h_j is an element of the feature vector H.

Eq. (5) is an unweighted facial feature description model, and we simply call it as Unweighted Model. In this model, each feature plays the same role.

As the psychophysical studies show that some facial regions contain more useful information than others for distinguishing different persons [2], Ahonen *et al.* [12] assign different weights to different regions, but all the features in the same region still have the same weight. We visualize the region weights in Fig. 2. Based on this idea, we also assign different weights to different regions in model (5), and then get a region weighted facial feature description model shortly named as Region Weighted Model, which can be described as following:

$$\begin{aligned} K &= [k_1, \dots, k_j, \dots, k_M] \\ &= [\underbrace{\rho_1 k_1, \dots, \rho_1 k_m}_{K[1]}, \underbrace{\rho_2 k_{m+1}, \dots, \rho_2 k_{2m}}_{K[2]}, \dots, \\ &\quad \underbrace{\rho_m k_{(r-1)m+1}, \dots, \rho_m k_{rm}}_{K[r]}], \end{aligned} \tag{6}$$

where K is the local feature vector, divided into $K[1], K[2], \dots, K[r]$, corresponding to the r regions of the face image. k_j is an element of the feature vector K. ρ_m is the weight for all the features in the m^{th} region.

Model (6) has considered different contributions of different facial regions, but it ignores the variety of feature contributions in the same region.

IV. Salience Evaluation of Facial Features

A. Sample Sets Generation

For a face image dataset, LTP or HM-LBP descriptor is adopted to extract feature vectors, which compose a feature vector set. Because each feature vector has not been assigned

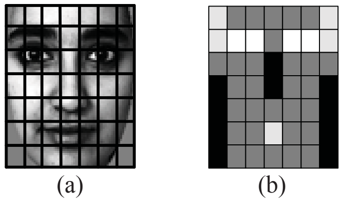

Fig. 2. Weights for each region: (a) a facial image divided into 7×7 windows; (a) a visualization of the weights for each region. Black squares indicate weight 0, dark gray 1, light gray 2, and white 4.

a label, this feature vector set can not be used to learn the feature evaluation vector. As a solution for this problem, a feature transformation is proposed. Each transformation is applied to two different images. If these two images are from the same person, then a within-class distance vector is obtained, and we assign label "1" to this vector. If these two images are from two different persons, then a between-class distance vector is gotten, and label "0" is assigned to this vector.

Suppose two feature vectors from the feature vector set are:

$$\begin{cases} H &= [h_1, \ldots, h_M] \\ \tilde{H} &= [\tilde{h}_1, \ldots, \tilde{h}_M]. \end{cases} \tag{7}$$

The facial feature transformation could be described as:

$$\begin{aligned} X_i &= f(H, \tilde{H}) \\ &= \left[\frac{(h_1 - \tilde{h}_1)^2}{h_1 + \tilde{h}_1}, \ldots, \frac{(h_j - \tilde{h}_j)^2}{h_j + \tilde{h}_j}, \ldots, \frac{(h_M - \tilde{h}_M)^2}{h_M + \tilde{h}_M} \right] \\ &= [x_1, \ldots, x_j, \ldots, x_M] \\ &= [\underbrace{x_1, \ldots, x_m}_{X_i[1]}, \underbrace{x_{m+1}, \ldots, x_{2m}}_{X_i[2]}, \ldots, \underbrace{x_{(r-1)m+1}, \ldots, x_{rm}}_{X_i[r]}], \end{aligned} \tag{8}$$

where x_j is the χ^2 distance of corresponding features h_j and $\tilde{h}_j$, and X_i is the χ^2 distance vector divided into $X_i[1], X_i[2], \ldots, X_i[r]$, corresponding to the features in r regions. If H and $\tilde{H}$ are both from the images of the same person, X_i is a within-class distance vector, denoted as X_i^+; Otherwise, X_i is a between-class distance vector, denoted as X_i^-. Finally, the within- and between-class distance vector sets are obtained respectively as:

$$\begin{cases} X^+ &= \{X_i^+, i = 1, \ldots, p\} \\ X^- &= \{X_i^-, i = 1, \ldots, q\}, \end{cases} \tag{9}$$

where p and q are the total numbers of elements in X^+ and X^-. As the within- and between-class distance vectors are respectively from the face images of the same person and different persons, it is easy to know that $p \ll q$.

B. Sparse learning for Facial Feature Evaluation

In section IV-A, the within-class distance vector set X^+ and between-class distance vector set X^- are obtained, respectively. However, there is still a problem for learning the feature evaluation vector, because the number of the between-class distance vector is greatly larger than that of the within-class distance vector. For balancing the numbers of within- and between-class distance vectors, Bootstrap strategy [16], [17] is employed to sample on set X^- to get p samples as the negative sample set $\tilde{X}^-$, and label the negative samples as "0". Simultaneously, we take the whole set X^+ as the positive sample set $\tilde{X}^+$, and label the positive samples as "1". Consequently, $\tilde{X}^+$ and $\tilde{X}^-$ compose the training set S, which contains both p within-class distance vectors and p between-class distance vectors, i.e. $W = 2p$. Thus, S is also a distance matrix.

Suppose the label vector of the training samples are L. The core idea is to model the relationship between the training samples and their labels. Namely, we can use the feature evaluation vector β to do regressing between S and L. Thus, we have

$$L = S\beta. \tag{10}$$

L, S and β satisfy the following constrains:

$$\begin{cases} L = [l_1, \ldots, l_f, \ldots, l_W]^{\mathrm{T}} \\ S = \begin{bmatrix} \tilde{X}^+ \\ \tilde{X}^- \end{bmatrix} \\ \beta = [\beta_1, \ldots, \beta_j, \ldots, \beta_M]^{\mathrm{T}}, \end{cases} \tag{11}$$

where W is the total number of the samples in the training set S; M is the number of the elements in the each distance feature vector; l_f is the label of the f^{th} sample; β_j is the evaluation coefficient of the j^{th} element in the distance feature vector. Fig. 3 presents a visual explanation for the facial feature transformation (8) and the facial feature evaluation model (10).

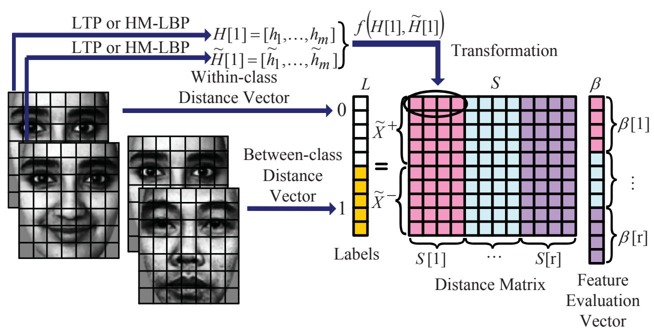

Fig. 3. A visual expression of the facial feature transformation and the facial feature evaluation model.

Because the dimensionality of the distance vector is much larger than the total number of samples, i.e. $M > W$, β in (10) has an ununique solution. Thus, a regularization term, Group lasso scheme [18]–[21], is introduced to equation (10) for a unique solution of β. As the predictors belong to predefined groups in this scheme, it is desirable to shrink and select the members of a group together. Dividing the distance matrix S and feature evaluation vector β into r sub-matrices respectively, we have:

$$\begin{cases} S &= [\underbrace{s_1,\ldots,s_m}_{S[1]},\underbrace{s_{m+1},\ldots,s_{2m}}_{S[2]},\ldots,\underbrace{s_{(r-1)m+1},\ldots,s_{rm}}_{S[r]}] \\ \beta &= [\underbrace{\beta_1,\ldots,\beta_m}_{\beta^{\mathrm{T}}[1]},\underbrace{\beta_{m+1},\ldots,\beta_{2m}}_{\beta^{\mathrm{T}}[2]},\ldots,\underbrace{\beta_{(r-1)m+1},\ldots,\beta_{rm}}_{\beta^{\mathrm{T}}[r]}]^{\mathrm{T}}, \end{cases} \tag{12}$$

where s_m is the m^{th} column of S.

By introducing the group lasso scheme to the problem (10), the following sparse minimization problem could be formulated to get the solution of β:

$$\underset{\beta\in\mathbb{R}^M}{\arg\min}\,\|L-\sum_{\eta=1}^{r}S[\eta]\beta[\eta]\|_2^2+\lambda\sum_{\eta=1}^{r}\|\beta[\eta]\|_2, \tag{13}$$

where $S[\eta]$ denotes the η^{th} group of distance vectors, $\beta[\eta]$ is the corresponding feature evaluation vector of the η^{th} group of features, and $\|\cdot\|_2$ is the Euclidean norm. Here $\lambda > 0$ is a complexity parameter which controls the amount of shrinkage: the larger the value of λ is, the greater the amount of shrinkage could be. The solution of (13) can be obtained with the method presented in [22]. As each element of the feature evaluation vector must be nonnegative, the following formula is used to update β:

$$\overline{\beta}_j = \begin{cases} \beta_j & if\ \beta_j > \tau \\ 0 & if\ \beta_j \le \tau, \end{cases} \tag{14}$$

where $\overline{\beta}_j$ is the updated value of β_j, and $\tau > 0$ is a threshold which controls the dimensionality of the facial feature vector. The larger the value of τ is, the smaller the dimensionality of the facial feature vector will be. Suppose the final evaluating indicator vector by updating β is:

$$\overline{\beta} = [\overline{\beta}_1,\ldots,\overline{\beta}_j,\ldots,\overline{\beta}_M]^{\mathrm{T}}. \tag{15}$$

As many elements of $\overline{\beta}$ are zeros, the corresponding features of these zero elements can be omitted directly, which greatly reduces the facial feature dimensionality.

V. Salient Facial Feature Description Model

A new facial feature description model is obtained by combining (15) and (5):

$$\begin{aligned} A &= [a_1,\ldots,a_j,\ldots,a_M] \\ &= [\overline{\beta}_1\times h_1,\ldots,\overline{\beta}_j\times h_j,\ldots,\overline{\beta}_M\times h_M] \\ &= [\underbrace{\overline{\beta}_1 h_1,\ldots,\overline{\beta}_m h_m}_{A[1]},\underbrace{\overline{\beta}_{m+1}h_{m+1},\ldots,\overline{\beta}_{2m}h_{2m}}_{A[2]},\ldots, \\ &\quad \underbrace{\overline{\beta}_{(r-1)m+1}h_{(r-1)m+1},\ldots,\overline{\beta}_{rm}h_{rm}}_{A[r]}], \end{aligned} \tag{16}$$

where A is a feature vector, and divided into $A[1]$, $A[2]$, ..., $A[r]$, corresponding to the r regions of the face image. Each feature is afforded a weight to evaluate its contribution for face recognition in model (16). Thus, this model is a feature weighted facial feature description model, and we call it Feature Weighted Model simply. As β assigns larger weights to more salient features, model (16) is also a salient facial feature description model.

VI. Experiments

In this section, three facial feature description models: Unweighted Model, Region Weighted Model and Feature Weighted Model are employed to perform the experiments. The region weights for Region Weighted Model are all the same to those in [12], which can be seen in Fig. 2. In different databases, the region weights are fixed. The feature weights for Feature Weighted Model are generated automatically with the sparse learning approach described in Section IV, and various in different databases. Two feature descriptors: LTP [13] and HM-LBP [14] are adopted for extracting the original image features.

A. Data Description

Two face databases FERET [23] and CAS-PEAL-R1 [24] are used in our experiments. FERET database is one of the well-known face database, which contains five subsets varying in lighting, ages and expressions. Experiments are performed with the same gallery and probe sets specified by the FERET evaluation protocol. 551 positive samples and 658,445 negative samples are employed to learn the feature evaluation vector β. These positive and negative samples are generated with all the training samples from the standard FERET training set [23] and "subfc training set" [25].

Another famous face database is also applied to the experiments. In this database, there are 30,900 images of 1,040 individuals containing different poses, expressions, and so on. The frontal images from subsets of accessory, background, distance, expression, and aging are taken as probe sets and normal gallery set in our experiments. 603 positive samples and 200,196 negative samples generated by the standard training set of CAS-PEAL-R1 database are used to learn the feature evaluation vector β.

B. Results

The results of the three models: Unweighted Model, Region Weighted Model and Feature Weighted Model are drawn in Tables I, II and Figs. 4, 5. From Tables I and II, we can see that Feature Weighted Model achieves the highest recognition rates on most of the subsets containing variations in lighting, expressions, accessory, distance, background, and aging.

From the last column of Table I, we can see that for each feature descriptor of LTP and HM-LBP, the average recognition rate of Feature Weighted Model is obviously better than Unweighted Model and Region Weighted Model. Namely, Feature Weighted Model has the highest average recognition rate.

TABLE I

COMPARISON RESULTS OF DIFFERENT MODELS ON FERET DATABASE

Model	Feature	Dimensionality	Recognition rate (%) on the subset				Average recognition rate (%)
			fb	fc	dupI	dupII	
Unweighted Model [13]	LTP	5782	92.8	36.1	54.8	44.4	74.6
Region Weighted Model	LTP	4602	96.0	61.9	59.1	**50.4**	80.2
Feature Weighted Model	LTP	**800**	**97.3**	**80.4**	**61.6**	47.4	**83.7**
Unweighted Model [14]	HM-LBP	8575	94.2	48.5	65.2	56.8	80.1
Region Weighted Model	HM-LBP	6825	97.4	75.3	68.8	64.1	85.6
Feature Weighted Model	HM-LBP	**2900**	**98.3**	**91.2**	**74.5**	**65.0**	**89.5**

TABLE II

COMPARISON RESULTS OF DIFFERENT MODELS ON CAS-PEAL-R1 DATABASE

Model	Feature	Dimensionality	Recognition rate (%) on the subset					Average recognition rate (%)
			Accessory	Distance	Background	Expression	Aging	
Unweighted Model [13]	LTP	5782	61.4	96.9	83.9	88.4	81.8	72.3
Region Weighted Model	LTP	4602	53.4	96.3	78.5	88.2	81.8	67.2
Feature Weighted Model	LTP	**2000**	**70.5**	**97.5**	**86.9**	**93.3**	**86.4**	**79.5**
Unweighted Model [14]	HM-LBP	8575	**75.9**	97.5	99.4	89.1	87.9	82.6
Region Weighted Model	HM-LBP	6825	66.6	**97.8**	**99.5**	88.7	89.4	77.5
Feature Weighted Model	HM-LBP	**2900**	75.1	**97.8**	**99.5**	**92.4**	**90.9**	**83.3**

All the results show that Feature Weighted Model can get the best performance on face recognition comparing to the Unweighted Model and Region Weighted Model. This is due to the fact that Feature Weighted Model not only considers the contributions of different face regions, but also can distinguish the contributions of different dimensions of the features in the same region, which can represent the face more precisely.

From Tables I and II, we can see that the feature vector generated by Feature Weighted Model has the lowest dimensionality comparing with the feature vectors produced by Unweighted Model and Region Weighted Model. This indicates that Feature Weighted Model has the best sparsity among the three models.

Figs. 4 and 5 present the relationship between feature dimensionality and recognition rate generated by Feature Weighted Model with LTP and HM-LBP features. From these two figures, Feature Weighted Model achieves satisfactory performance with much lower feature dimensionality, which shows that Feature Weighted Model can remarkably reduce the feature dimensionality.

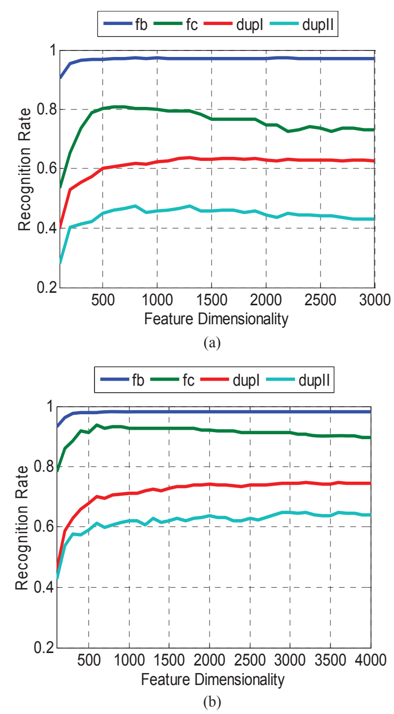

Fig. 4. Relationship between feature dimensionality and recognition rate produced by Feature Weighted Model on FERET database: (a) with LTP features; (b) with HM-LBP features.

VII. CONCLUSIONS

Face recognition attracts more and more attention for both wide applications and scientific challenges. This paper focuses on constructing an efficient facial feature description model for face recognition, which can complete two goals: (1) facial feature dimensionality reduction; (2) salient facial feature selection. Thus, this paper proposes a new sparse learning approach for salient facial feature description. The proposed method firstly presents a facial feature transformation to generate the training samples composed of within- and between-class distance vectors sets. Then, it proposes a sparse learning approach to learn the feature evaluation vector based on the training samples. Finally, the feature evaluation

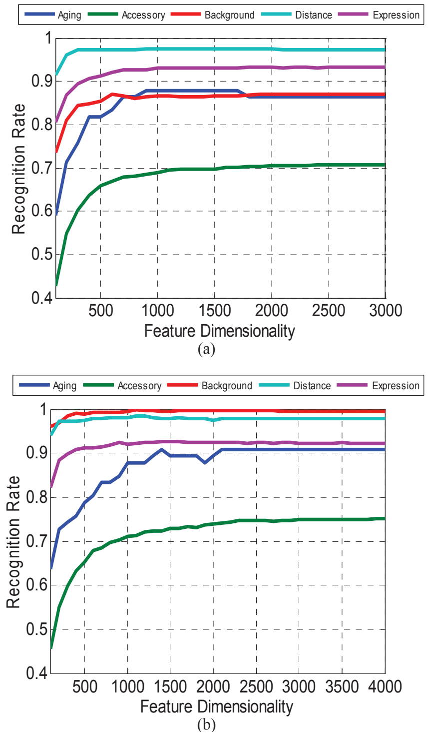

Fig. 5. Relationship between feature dimensionality and recognition rate produced by Feature Weighted Model on CAS-PEAL-R1 database: (a) with LTP features; (b) with HM-LBP features.

vector is employed to construct a new model for salient facial feature description.

Because the sparse learning approach encourages the sparsity at both group and individual levels, the proposed method can greatly reduce the dimensionality of the facial feature vector. As the feature evaluation vector can distinguish the contributions of different features in the same region, the proposed method can recognize the salience of different features. Experimental results show that the proposed method achieves much better face recognition performance with lower feature dimensionality. We really believe that the proposed method can also be successfully used in object recognition, face recognition, texture classification, an so on.

REFERENCES

[1] A. Rice, P. J. Phillips, and A. O'Toole, "Variable use of the face and body in person identification," *Journal of Vision*, vol. 13, no. 9, pp. 977–977, 2013.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[3] C. Ye, T. Wu, Y. Chen, P. He, P. Xie, Y. Zhang, S. Teng, Y. Chen, and P. Hsiung, "Smart video camera design–real-time automatic person identification," in *Advances in Intelligent Systems and Applications*. Springer, 2013, vol. 2, pp. 299–309.

[4] F. Huang and J. Su, "Face contour detection using geometric active contours," in *Proceedings of World Congress on Intelligent Control and Automation*, vol. 3. IEEE, 2002, pp. 2090–2093.

[5] J. Dai, D. Liu, and J. Su, "Rapid eye localization based on projection peak," *Pattern Recognition and Artificial Intelligence*, vol. 4, 2009.

[6] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[7] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1991, pp. 586–591.

[8] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[9] X. Jing, H. Wong, and D. Zhang, "Face recognition based on 2d fisherface approach," *Pattern Recognition*, vol. 39, no. 4, pp. 707–710, 2006.

[10] L. Wiskott, J. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[11] D. S. Bolme, "Elastic bunch graph matching," Ph.D. dissertation, Colorado State University, 2003.

[12] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[13] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[14] Z. Guo, L. Zhang, D. Zhang, and X. Mou, "Hierarchical multiscale LBP for face and palmprint recognition," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2010, pp. 26–29.

[15] Y. Zhao and J. Su, "LBP-based hierarchical sparse patch learning for face recognition," in *IEEE International Conference on Information and Automation*. IEEE, 2013, pp. 868–872.

[16] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics, 2001.

[17] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[18] Y. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[19] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2009.

[20] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *Arxiv preprint arXiv:1001.0736*, 2010.

[22] J. Liu and J. Ye, "Fast overlapping group lasso," *Arxiv preprint arXiv:1009.0306*, 2010.

[23] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[24] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2008.

[25] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European Conferenceon Computer Vision (ECCV)*. Springer, 2004, pp. 469–481.