Nuclear Norm-Based 2-DPCA for Extracting Features From Images

Fanlong Zhang, Jian Yang, *Member, IEEE*, Jianjun Qian, and Yong Xu, *Member, IEEE*

***Abstract*— The 2-D principal component analysis (2-DPCA) is a widely used method for image feature extraction. However, it can be equivalently implemented via image-row-based principal component analysis. This paper presents a structured 2-D method called nuclear norm-based 2-DPCA (N-2-DPCA), which uses a nuclear norm-based reconstruction error criterion. The nuclear norm is a matrix norm, which can provide a structured 2-D characterization for the reconstruction error image. The reconstruction error criterion is minimized by converting the nuclear norm-based optimization problem into a series of F-norm-based optimization problems. In addition, N-2-DPCA is extended to a bilateral projection-based N-2-DPCA (N-B2-DPCA). The virtue of N-B2-DPCA over N-2-DPCA is that an image can be represented with fewer coefficients. N-2-DPCA and N-B2-DPCA are applied to face recognition and reconstruction and evaluated using the Extended Yale B, CMU PIE, FRGC, and AR databases. Experimental results demonstrate the effectiveness of the proposed methods.**

***Index Terms*— Feature extraction, nuclear norm, principal component analysis (PCA), subspace analysis.**

I. Introduction

PRINCIPAL component analysis (PCA) [1] is a classical feature extraction and data representation technique, which has been widely used in the areas of pattern recognition and computer vision. The PCA was applied to face recognition in [43]. Since then, PCA has been widely investigated and a number of its extended versions are presented, such as weighted PCA (WPCA) [45] and independent component analysis (ICA) [47], [48]. The WPCA uses a weighted distance to alleviate the effect of the outliers onto the projection directions. The ICA, as a generalization of PCA, concerns not only the second-order dependences between variables but also the high-order dependences between them. The PCA makes the data uncorrelated, while ICA makes the data as independent as possible. The PCA and ICA are both unsupervised methods, while Fisher linear discriminate analysis (LDA) is supervised method and turns out to be very effective for face recognition [50], [51]. To avoid over fitting, LDA is generally implemented in the PCA-transformed space, i.e., PCA + LDA. Discrete cosine transform (DCT) has also been employed for face recognition [49]. The advantage of the DCT is that its basis is data independent and it can be implemented very fast.

However, PCA ignores the structural information of the image since it often converts an image to a vector. To exploit the structural information, Yang *et al.* [2] proposed the 2-D PCA (2-DPCA) which is based on 2-D image matrix rather than 1-D vector. 2-DPCA has been widely applied in pattern recognition and face recognition [3], [4]. The relationship between PCA and 2-DPCA has been discussed in [5]–[10]. As shown in [5] and [6], 2-DPCA operates on image rows, and ignores the information behind the image columns. To combine both kinds of image information (in rows and columns), bilateral projection-based 2-DPCA (B2-DPCA) are developed in [5]–[7], respectively. They seek two projection matrices to extract row information and column information simultaneously. Further research can be seen in [11]–[13].

A family of kernel-based methods and manifold learning methods also aroused wide research interests. Scholkopf *et al.* [44] presented kernel PCA (KPCA), which performs PCA in a kernel-induced feature space. Liwicki *et al.* [36] presented Euler PCA (EPCA), which is a special KPCA with a complex kernel in an explicitly defined Hilbert space. Yang *et al.* [46] proposed a two-phase kernel discriminate analysis, i.e., KPCA + LDA. Zafeiriou *et al.* [31] put forward a regularized kernel discriminate analysis with a robust kernel for face recognition and verification. He *et al.* [32] proposed the locality preserving projections (LPPs), which is derived from Laplacian eigen-map. In contrast to most manifold learning algorithms, LPP possesses the remarkable advantage that it can generate an explicit map.

Recently, the idea of sparse representation is used to design some feature extraction methods. Clemmensen *et al.* [33] provided a sparse LDA. Lai *et al.* [34] suggested a sparse version of the 2-D local discriminate projections. Yang *et al.* [35] proposed a sparse representation classifier (SRC) steered discriminative projection method, which maximizes the ratio of the between-class reconstruction residual to the within-class reconstruction residual in the projected space and thus enables an SRC to achieve better performance.

The PCA, 2-DPCA, and many other variants [5], [7], [11], which are all based on L_2-norm metric, are not robust in

Manuscript received September 5, 2013; accepted November 24, 2014. Date of publication January 8, 2015; date of current version September 16, 2015. This work was supported in part by the National Science Fund for Distinguished Young Scholars under Grant 61125305, Grant 61472187, Grant 61233011, and Grant 61373063, in part by the Key Project through the Ministry of Education, China, under Grant 313030, in part by the 973 Program under Grant 2014CB349303, in part by the Fundamental Research Funds for the Central Universities under Grant 30920140121005, and in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT13072.

F. Zhang, J. Yang, and J. Qian are with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csfzhang@126.com; csjyang@njust.edu.cn; qjjtx@126.com).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen 518055, China (e-mail: laterfall2@yahoo.com.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2376530

the sense that outlying samples can arbitrarily skew the solution from the desired solution. Considering L_1-norm is more robust to outliers [15], Li *et al.* [16] developed L_1-norm-based 2-DPCA (L_1-2-DPCA), which is an extension of L_1-norm-based PCA (PCA-L_1) [14].

Although 2-DPCA, B2-DPCA, and L_1-2-DPCA are all based on 2-D matrices, they can be equivalently implemented via vector-based methods, such as PCA and PCA-L_1 [5], [9], [11]. The inherent reason is that these methods use the Frobenius norm (F-norm) or L_1-norm, which are essentially 1-D vector norm.

In addition, the F-norm-based methods essentially employ Euclidean metric to measure the similarity between different images. Nonetheless, F-norm-based metric is not very robust since the variations between the images of the same person due to the illumination and viewing direction are almost always larger than the image variations due to the change of identities [17], [18]. To alleviate this problem, Gu *et al.* [19] used nuclear norm metric instead of F-norm to measure the similarity between two images. They demonstrated that nuclear norm is less sensitive to illumination changes. They also presented a nuclear norm-based PCA method, i.e., Schatten 1-PCA [19].

For 2-D-based subspace methods, except Euclidian distance, many different metrics are also used to measure the similarity between two feature matrices [37], [40], [41]. Zuo *et al.* [37] proposed an assembled matrix distance metric (AMD). Xu *et al.* [40] proposed to learn similarity measure by boosting. Bajwa *et al.* [41] provided a comprehensive comparative analysis of some recent 1-D and 2-D subspace methods with four distance metrics, including Euclidean (L_2) and cosine for the image space and their counter parts in Mahalanobis space, Mahalanobis (L_2), and Mahalanobis cosine.

Recently, nuclear norm-based minimization problem has aroused broad interests in the fields of pattern recognition and compressed sensing. Nuclear norm is essentially the convex envelope of the matrix rank [20], [21]. The nuclear norm-based optimization has been used in low rank matrix recovery [22], [23], removing self-shadowing in face images [24], and so on. Fornasier *et al.* [25] presented an efficient algorithm to solve the nuclear norm minimization problem by converting nuclear norm problem into F-norm problem in conjunction with the iteratively reweighted tactics.

In this paper, inspired by [2], [5], [7], and [19], we propose the nuclear norm-based 2-DPCA (N-2-DPCA). Differing from 2-DPCA, our model used nuclear norm to measure the reconstruction error rather than F-norm. We provide justifications for using the nuclear norm to characterize the reconstruction error. Note that Gu *et al.* [19] also used nuclear norm to characterize the transformed data in their model. However, in their algorithm for maximizing the criterion, they imposed an additional constraint on the desired projection matrix $\mathbf{P}$, i.e., $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$, which requires that the projection matrix is an orthogonal matrix. In general, the matrix $\mathbf{P}$ is a column-rank-deficient matrix for real world dimensionality reduction tasks. So, Gu's Schatten 1-PCA is an approximate algorithm. In this paper, we develop an exact algorithm for N-2-DPCA. Inspired by [5] and [7], the N-2-DPCA is further extended to nuclear norm-based bilateral 2-DPCA (N-B2-DPCA).

The rest of this paper is organized as follows. Section II reviews the related work of 2-DPCA, B2-DPCA, and L_1-2-DPCA. Section III presents our model, N-2-DPCA. Section IV extends N-B2-DPCA. Section V reports experimental results. Finally, the conclusions are drawn in Section VI.

II. Related Works

Given s image matrices $\mathbf{A}_1, \ldots, \mathbf{A}_s$ in $R^{m\times n}$, without loss of generality and for simplicity of discussion, the samples are assumed to have zero mean, i.e., $(1/s)\sum_{i=1}^{s}\mathbf{A}_i = 0$.

A. 2-DPCA and B2-DPCA

The 2-DPCA aims to find an $n \times r$ projection matrix $\mathbf{P}$ minimizing the following reconstruction error criterion:

$$\min_{\mathbf{P}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\right|\right|_F^2 \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r. \tag{1}$$

The image covariance (scatter) matrix of 2-DPCA is defined as $\mathbf{G} = 1/s\sum_{i=1}^{s}\mathbf{A}_i^T\mathbf{A}_i$. The optimal $\mathbf{P}$ can be obtained by finding orthogonal eigenvectors of $\mathbf{G}$ corresponding to the first r largest eigen-values. The 2-DPCA can be equivalently implemented via image-row-based PCA [5], [9]. To see this, let $\mathbf{a}_j^i$ be the jth row of image $\mathbf{A}_i$, and relabel $(\mathbf{a}_j^i)^T$, $j = 1\ldots m, i = 1\ldots s$ as $\mathbf{x}_1\ldots\mathbf{x}_{sm}$. Let the mean vector be $\mathbf{x}_0 \triangleq 1/sm\sum_{i=1}^{sm}\mathbf{x}_i$. Then, covariance matrix $\mathbf{G}$ can be rewritten as

$$\begin{aligned}\mathbf{G} &= \tfrac{1}{s}\sum_{i=1}^{s}\mathbf{A}_i^T\mathbf{A}_i = \tfrac{1}{s}\sum_{i=1}^{s}\sum_{j=1}^{m}\left(\mathbf{a}_j^i\right)^T\left(\mathbf{a}_j^i\right)\\ &= \tfrac{1}{s}\sum_{t=1}^{sm}(\mathbf{x}_t - \mathbf{x}_0)(\mathbf{x}_t - \mathbf{x}_0)^T\\ &= m\left[\tfrac{1}{sm}\sum_{t=1}^{sm}(\mathbf{x}_t - \mathbf{x}_0)(\mathbf{x}_t - \mathbf{x}_0)^T\right]\end{aligned} \tag{2}$$

where the matrix in square brackets is the covariance matrix of vector samples set $\{\mathbf{x}_t\}_{t=1}^{sm}$. Hence, 2-DPCA performed on the matrices is essentially the PCA performed on the rows of all the images.

The 2-DPCA adopts a unilateral projection scheme, and produces more coefficients than PCA when representing an image. To remedy this drawback, B2-DPCA is proposed in [5]. The B2-DPCA seeks two projection matrices $\mathbf{Q} \in R^{m\times t}$ and $\mathbf{P} \in R^{n\times r}$ simultaneously by the following reconstruction error criterion:

$$\min_{\mathbf{P},\mathbf{Q}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T A_i\mathbf{P}\mathbf{P}^T\right|\right|_F^2 \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_t. \tag{3}$$

Equation (3) is a generalization of (1).

The problem in (3) is solved by alternately iterative algorithm [5], which contains two steps as follows.

1) Given $\mathbf{Q}$, the optimal $\mathbf{P}$ is formed by the first r eigenvectors corresponding to the first r largest eigen-values of $\mathbf{G}_1$, where $\mathbf{G}_1 = (1/s)\sum_{i=1}^{s}(\mathbf{A}_i^T\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i)$.

(a) (b) (c)

Fig. 1. (a)–(c) Three face images from the Extended Yale B database.

2) Given $\mathbf{P}$, the optimal $\mathbf{Q}$ is formed by the first t eigenvectors corresponding to the first t largest eigen-values of $\mathbf{G}_2$, where $\mathbf{G}_2 = (1/s)\sum_{i=1}^{s}(\mathbf{A}_i\mathbf{P}\mathbf{P}^T\mathbf{A}_i^T)$.

B. L_1-Norm-Based 2-DPCA

Image-to-matrix methodology [2] motivates a series of research efforts in pattern recognition. One representative work is L_1-norm-based 2-DPCA (L_1-2-DPCA) [16], which maximizes the L_1-norm variance in low-dimensional feature space

$$\max_{\mathbf{u}} \sum_{i=1}^{s} ||\mathbf{A}_i\mathbf{u}||_1 \quad \text{s.t. } \mathbf{u}^T\mathbf{u} = 1. \tag{4}$$

L_1-2-DPCA is robust to outliers [16]. However, compared with 2-DPCA, L_1-2-DPCA consumes more time than 2-DPCA due to that each principal vector is obtained via iteration operation.

Actually, L_1-2-DPCA is also equivalent to the L_1-norm-based PCA (PCA-L_1) [14], [12]. The objective function of (4) can be rewritten as

$$\sum_{i=1}^{s} ||\mathbf{A}_i\mathbf{u}||_1 = \sum_{i=1}^{s}\sum_{j=1}^{m} \left|\mathbf{a}_j^i\mathbf{u}\right| = \sum_{t=1}^{sm} \left|\mathbf{u}^T\mathbf{x}_t\right|. \tag{5}$$

The right term is exactly the objective function of PCA-L_1 [14].

III. Nuclear Norm-Based 2-DPCA

In this section, we present N-2-DPCA model and use the iteratively reweighted method to solve it.

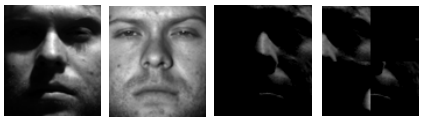
A. Motivation and Model

One justification of using nuclear norm is that, as a distance metric, it is more reliable than L_1 or L_2 norm (F-norm). For example, Fig. 1(a) and (b) are from one person and Fig. 1(c) is from another person. The distance between the images of the same person and different person are computed using the nuclear norm, L_2-norm and L_1-norm, as shown in Table I. We can see that both L_2-norm and L_1-norm fail to classify image in Fig. 1(a) correctly since the distance between Fig. 1(a) and (b) is larger than that between Fig. 1(a) and (c). However, the nuclear norm gives the correct result, since the nuclear norm-based distance between Fig. 1(a) and (b) is smaller than that between Fig. 1(a) and (c). This example motivates us to use the nuclear norm-based criterion.

Another justification is that nuclear norm is more suitable for characterizing the reconstruction error than L_1 or L_2 norm in the case of illumination change. Fig. 2(a) and (b) shows images of one person under different illuminations,

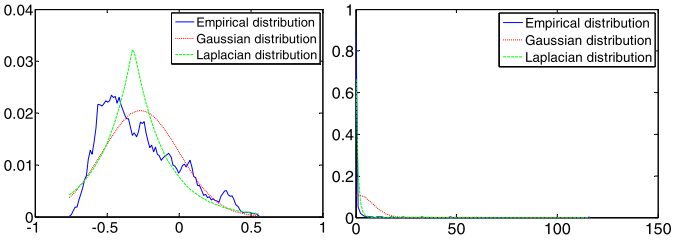
TABLE I
COMPARISON OF DISTANCE BETWEEN IMAGES USING DIFFERENT NORMS

	(a)-(b)	(a)-(c)
$\|\bullet\|_*$	88.1	101.9
$\|\bullet\|_2$	24.4	22.6
$\|\bullet\|_1$	3742	3105



(a) (b) (c) (d)

Fig. 2. Example images. (a) Image with bad lighting condition. (b) Image with good lighting condition. (c) Error image. (d) Rearranged image.



(a) (b)

Fig. 3. (a) Empirical distribution and the fitted distributions of the noise image $\mathbf{E}$. (b) Empirical distribution and the fitted distributions of the singular value vector of error image $\mathbf{E}$.

where Fig. 2(a) is taken under a bad lighting condition, and Fig. 2(b) under a good lighting condition. We view Fig. 2(b) as a ground truth so we expect Fig. 2(b) to be a reconstructed image of Fig. 2(a), and thus the reconstruction error image is $\mathbf{E}$ in Fig. 2(c), i.e., the difference between Fig. 2(a) and (b). Fig. 3(a) shows the error term $\mathbf{E}$ fitted by different distributions. One can see that Gaussian and Laplace distributions are far away from the empirical distribution. Instead, Fig. 3(b) shows that singular values of error matrix $\mathbf{E}$ fit Laplace distribution well.

From the probability distribution point of view, we know that L_1-norm provides an optimal characterization for errors with the Laplace distribution [42], while L_2-norm is optimal for Gaussian distribution [1]. However, from Fig. 3(a), one can see that the error $\mathbf{E}$ does not follow Laplace or Gaussian distributions. So, L_2-norm (or L_1-norm)-based methods cannot describe this kind of reconstruction error effectively.

Nuclear norm of a matrix is the sum of all singular values of the matrix, which is actually L_1-norm of the singular value vector. From Fig. 3(b), we can see that the singular values of the error image $\mathbf{E}$ follow Laplace distribution well. It means nuclear norm of the error image is more suitable for characterizing the structural noise caused by illumination changes than L_1-norm or L_2-norm. This motivates us to use the nuclear norm to characterize the reconstruction error matrix. Thus, the objective function of N-2-DPCA is defined as

$$\min_{\mathbf{P}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\right|\right|_* \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r. \tag{6}$$

In (6), $||\cdot||_*$ denotes nuclear norm. It is believed that nuclear norm can describe structural information more effectively than L_1-norm or L_2-norm. To see this, we arrange the pixels of $\mathbf{E}$ and obtain the image $\mathbf{F}$. The L_2-norm values of matrices $\mathbf{E}$ and $\mathbf{F}$ are equal (the value is 33.96), but their nuclear norm values are different (the values are 91.99 for $\mathbf{E}$ and 98.52 for $\mathbf{F}$). For previous 2-D methods based on L_2 or L_1 norm, the measure of the error image is still based on pixel values, so the structural information of the error image cannot be revealed.

B. Algorithm

We discuss how to solve (6) in this section. Motivated by [25], we convert a nuclear norm optimization problem to the F-norm (L_2-norm) optimization problem. To this end, let us give the following lemma.

Lemma 1 [25]: For matrix $\mathbf{X} \in R^{p\times q}$, one has

$$||\mathbf{X}||_* = ||(\mathbf{X}\mathbf{X}^T)^{-1/4}\mathbf{X}||_F^2. \tag{7}$$

Lemma 1 represents the nuclear norm in the form of F-norm, and provides a base for solving our model.

In Lemma 1, the αth power of a matrix $\mathbf{X}$ of rank r is defined by

$$\mathbf{X}^\alpha = \mathbf{U}\Sigma^\alpha\mathbf{V}^T, \quad \Sigma^\alpha = \mathrm{diag}\big(\sigma_1^\alpha, \ldots, \sigma_r^\alpha\big) \tag{8}$$

where $\mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition of $\mathbf{X}$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. From Lemma 1, the objective function in model (6) can be rewritten as

$$J(\mathbf{P}) = \sum_{i=1}^{s} \big|\big|\mathbf{W}_i\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)\big|\big|_F^2 \tag{9}$$

where $\mathbf{W}_i$ is the weight matrix and defined by

$$\mathbf{W}_i = \Big(\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)^T\Big)^{\frac{-1}{4}}. \tag{10}$$

Now, we use iteratively reweighted method to solve our model. The procedure consists of the following iterations.

1) Given $\mathbf{W}_i = \mathbf{W}_i^k$, updating $\mathbf{P}$ by

$$\mathbf{P}^{k+1} = \arg\min_{\mathbf{P}} \sum_{i=1}^{s} \big|\big|\mathbf{W}_i\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)\big|\big|_F^2 \quad \text{s.t. } \mathbf{P}\mathbf{P}^T = \mathbf{I}_r. \tag{11}$$

2) Given $\mathbf{P} = \mathbf{P}^{k+1}$, updating $\mathbf{W}_i$ by

$$\mathbf{W}_i^{k+1} = \Big(\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)^T\Big)^{\frac{-1}{4}}. \tag{12}$$

The key step is to solve the optimization problems (11). Its objective function can be rewritten as

$$\begin{aligned} J(\mathbf{P}) &= \sum_{i=1}^{s} \big|\big|\mathbf{W}_i\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)\big|\big|_F^2 \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\big(\mathbf{I} - \mathbf{P}\mathbf{P}^T\big)\mathbf{A}_i^T\mathbf{W}_i^T\big) \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\mathbf{A}_i^T\mathbf{W}_i^T\big) \\ &\quad -\mathrm{Tr}\left(\mathbf{P}\mathbf{P}^T\sum_{i=1}^{s}\big(\mathbf{A}_i^T\mathbf{W}_i^T\mathbf{W}_i\mathbf{A}_i\big)\right) \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{A}_i^T\mathbf{W}_i^T\mathbf{W}_i\mathbf{A}_i\big) \\ &\quad -\mathrm{Tr}\left(\mathbf{P}^T\sum_{i=1}^{s}\big(\mathbf{A}_i^T\mathbf{W}_i^T\mathbf{W}_i\mathbf{A}_i\big)\mathbf{P}\right) \end{aligned} \tag{13}$$

where the third equation is derived from the fact that the matrix $\mathbf{I} - \mathbf{P}\mathbf{P}^T$ is idempotent. Denote $\mathbf{D} = \sum_{i=1}^{s}(\mathbf{A}_i^T\mathbf{W}_i^T\mathbf{W}_i\mathbf{A}_i)$, problem (11) can be rewritten as

$$\mathbf{P}^{k+1} = \arg\max_{\mathbf{P}} \mathrm{Tr}(\mathbf{P}^T\mathbf{D}\mathbf{P}) \quad \text{s.t. } \mathbf{P}\mathbf{P}^T = \mathbf{I}_r. \tag{14}$$

So, $\mathbf{P}^{k+1}$ is the matrix formed by r orthonormal eigenvectors of $\mathbf{D}$ corresponding to the first r largest eigenvalues.

Now, we consider how to update (12) efficiently. Let $\mathbf{X}_i = \mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T$, $\mathbf{W}_i$ can be rewritten as $\mathbf{W}_i = (\mathbf{X}_i\mathbf{X}_i^T)^{-1/4}$. When some of the singular values of $\mathbf{X}_i\mathbf{X}_i^T$ become small, the computation of $\mathbf{W}_i$ becomes ill conditioned. To improve the stability of the algorithm, let us replace $\mathbf{X}_i$ by its ε-stabilization $(\mathbf{X}_i)_\varepsilon$. The ε-stabilization of one matrix $\mathbf{X}$ is defined by

$$\mathbf{X}_\varepsilon = \mathbf{U}\sum_{\varepsilon}\mathbf{V}^T, \quad \Sigma_\varepsilon = \mathrm{diag}(\max\{\sigma_i, \varepsilon\}_{i=1:r}). \tag{15}$$

However, for a fixed ε, we would no longer expect the algorithm to converge to the nuclear norm solution of (6). We select $\varepsilon_i^k = \min\{\varepsilon_i^{k-1}, \sigma_K(\mathbf{X}_i^k)\}$ at step k, and then one may hope for the stability and convergence toward the solution of (6). Above all, we update $\mathbf{W}_i$ by

$$\mathbf{W}_i^{k+1} = \Big[\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)_{\varepsilon_i^k}\big(\mathbf{A}_i - \mathbf{A}_i\mathbf{P}\mathbf{P}^T\big)_{\varepsilon_i^k}^T\Big]^{\frac{-1}{4}}. \tag{16}$$

The algorithm is summarized in Algorithm 1.

The convergence of the iteratively reweighted algorithm can be guaranteed when the constraint is linear [25]. Fig. 4 shows that the objective function value of N-2-DPCA converges well. Generally speaking, the variation of objective function value is $<10^{-6}$ when the number of iteration time is over 10.

After obtaining the projection matrix $\mathbf{P}$ by Algorithm 1, for a given image sample $\mathbf{A}$, the feature matrix $\mathbf{B}$ of the image sample $\mathbf{A}$ is obtained by $\mathbf{B} = \mathbf{A}\mathbf{P}$. The feature matrix $\mathbf{B}$ is used to represent image $\mathbf{A}$ for classification.

Algorithm 1 Iteratively Reweighted Method for N-2-DPCA

Input: Training data $\mathbf{A}_1, \ldots, \mathbf{A}_s$, projection dimensional r.

1: Initialize: $\mathbf{W}_{i=1:s}^0 = \mathbf{I}$, $\varepsilon_{i=1:s}^0 = 1$, $k = 0$, $K = 4$;
2: Compute $\mathbf{D}$: $\mathbf{D}^{k+1} = \sum_{i=1}^{s} \left(\mathbf{A}_i^T \left(\mathbf{W}_i^k\right)^T \mathbf{W}_i^k \mathbf{A}_i\right)$;
3: Update $\mathbf{P}$: $\mathbf{P}^{k+1} = [\mathbf{d}_1, \ldots, \mathbf{d}_r]$, where $\mathbf{d}_i$ is orthonormal eigenvectors of $\mathbf{D}^{k+1}$ corresponding to the i-th largest eigenvalues;
4: Compute $\mathbf{X}_i$ and ε_i:
$\mathbf{X}_i^{k+1} = \mathbf{A}_i - \mathbf{A}_i$, $\mathbf{P}^{k+1}(\mathbf{P}^{k+1})^T$,
$\varepsilon_i^{k+1} = \min\left\{\varepsilon_i^k, \sigma_K\left(\mathbf{X}_i^{k+1}\right)\right\}$;
5: Update $\mathbf{W}_i$: $\mathbf{W}_i^{k+1} = \left(\left(\mathbf{X}_i^{k+1}\right)_{\varepsilon_i^{k+1}} \left(\mathbf{X}_i^{k+1}\right)_{\varepsilon_i^{k+1}}^T\right)^{-1/4}$;
6: If $\varepsilon_i = 0$, go to 7; otherwise go to 2.
7: **Output:** Optimal projection matrix $\mathbf{P}^{k+1}$.

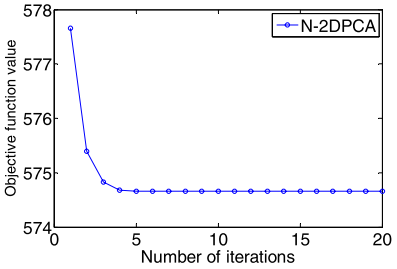


Fig. 4. Objective function value versus iteration times. Here, the training data are formed by 64 samples of subject 1 in the Extended Yale B database, and each sample is resized to 48×42. The number of projection axes is $r = 8$.

C. Connections to Existing 2-D Methods

Compared with 2-DPCA, the projection axes of N-2-DPCA are eigenvectors of the matrix $\mathbf{D} = \sum_{i=1}^{s} (\mathbf{A}_i^T \mathbf{W}_i^T \mathbf{W}_i \mathbf{A}_i)$, which can be viewed as weighted image covariance matrix $\mathbf{G}$ in 2-DPCA. In particular, each image sample $\mathbf{A}_i$ is weighted by the corresponding matrix $\mathbf{W}_i$. Meanwhile, the weighting matrix $\mathbf{W}_i$ is updated after completing the iteration each time. In the N-2-DPCA algorithm, we set the initial $\mathbf{W}_i$ as the identity matrix. That is, the 2-DPCA solution provides an initial solution for N-2-DPCA.

The model of the Schatten1-norm PCA [19] is

$$\max_{\mathbf{P}} \sum_{i=1}^{s} ||\mathbf{A}_i \mathbf{P}||_* \quad \text{s.t. } \mathbf{P}^T \mathbf{P} = \mathbf{I}_r. \tag{17}$$

Its criterion function is different from our method. Moreover, the solution of (17) is obtained on the condition that $\mathbf{P}^T \mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$, which is a very strong constraint. In general, a projection matrix $\mathbf{P}$ has only a small number of columns. So, the algorithm of Schatten1-norm PCA is not an exact algorithm for calculating the column-rank-deficient matrix $\mathbf{P}$. In contrast, our N-2-DPCA algorithm does not need any additional condition. It is an exact algorithm.

Table II shows the nuclear norm value of the optimal projection matrices of 2-DPCA, L_1-2-DPCA, Schatten 1-norm PCA, and N-2-DPCA (note that we use same experiment setting as that for Fig. 4). We can see that our model obtain minimal nuclear norm value among all methods.

TABLE II
OBJECTIVE FUNCTION VALUES AT DIFFERENT PROJECTION MATRICES

2DPCA	Schatten 1-PCA	L_1-2DPCA	N-2DPCA
577.7	582.3	580.5	574.7

IV. NUCLEAR NORM-BASED BILATERAL 2-DPCA

N-2-DPCA adopts a unilateral projection (right multiplication) scheme, which needs more coefficients for representing an image than PCA. As an extension of N-2-DPCA, N-B2-DPCA is developed, where left and right projection directions are calculated simultaneously. N-B2-DPCA can represent an image with much less coefficients than N-2-DPCA. The bilateral N-2-DPCA is formulated as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right|\right|_* \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_t. \tag{18}$$

Equation (18) is a generalization of (6). In (18), $\mathbf{P} \in R^{n \times r}$ and $\mathbf{Q} \in R^{t \times m}$ are the left and right multiplying projection matrices, respectively.

We update variables $\mathbf{P}$ and $\mathbf{Q}$ alternatively since there is no close-form solution for the problem (18).

Given $\mathbf{Q} = \mathbf{Q}^k$, update $\mathbf{P}$ by

$$\min_{\mathbf{P}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right|\right|_* \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r. \tag{19}$$

Given $\mathbf{P} = \mathbf{P}^{k+1}$, update $\mathbf{Q}$ by

$$\min_{\mathbf{Q}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right|\right|_* \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_t. \tag{20}$$

Similar to the way of solving N-2-DPCA, we use the iteratively reweighted method to solve (19) and (20).

A. Algorithms for Solving (19) and (20)

For (19), the procedure consists of the following iterations.

1) Given $\mathbf{W}_i = \mathbf{W}_i^k$, update $\mathbf{P}$ by

$$\mathbf{P}^{k+1} = \arg\min_{\mathbf{P}} \sum_{i=1}^{s} \left|\left|\mathbf{W}_i\left(\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right)\right|\right|_F^2 \quad \text{s.t. } \mathbf{P}\mathbf{P}^T = \mathbf{I}_r. \tag{21}$$

2) Given $\mathbf{P} = \mathbf{P}^{k+1}$, update $\mathbf{W}_i$ by

$$\mathbf{W}_i^{k+1} = \left(\left(\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right) \times \left(\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T \mathbf{A}_i \mathbf{P}\mathbf{P}^T\right)^T\right)^{\frac{-1}{4}}. \tag{22}$$

The key step is to solve the optimization problem (21). The objective function of (21) can be rewritten as

$$\begin{aligned} J(\mathbf{P}) &= \sum_{i=1}^{s} \left|\left|\mathbf{W}_i\left(\mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\mathbf{P}\mathbf{P}^T\right)\right|\right|_F^2 \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\big(\mathbf{A}_i\mathbf{A}_i^T - 2\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\mathbf{P}\mathbf{P}^T\mathbf{A}_i^T \\ &\qquad + \mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\mathbf{P}\mathbf{P}^T\mathbf{P}\mathbf{P}^T\mathbf{A}_i^T\mathbf{Q}\mathbf{Q}^T\big)\mathbf{W}_i^T\big) \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\mathbf{A}_i^T\mathbf{W}_i^T\big) \\ &\quad - \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\mathbf{P}\mathbf{P}^T\mathbf{A}_i^T(2\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{W}_i^T\mathbf{W}_i\big) \\ &= \sum\nolimits_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\mathbf{A}_i^T\mathbf{W}_i^T\big) \\ &\quad - \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{P}\mathbf{P}^T\mathbf{A}_i^T(2\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{W}_i^T\mathbf{W}_i\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\big) \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\mathbf{A}_i^T\mathbf{W}_i^T\big) \\ &\quad + \mathrm{Tr}\left(\mathbf{P}^T\sum_{i=1}^{s}\big(\mathbf{A}_i^T(\mathbf{Q}\mathbf{Q}^T - 2\mathbf{I})\mathbf{W}_i^T\mathbf{W}_i\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i\big)\mathbf{P}\right) \\ &= \sum_{i=1}^{s} \mathrm{Tr}\big(\mathbf{W}_i\mathbf{A}_i\mathbf{A}_i^T\mathbf{W}_i^T\big) + \mathrm{Tr}(\mathbf{P}^T\mathbf{D}\mathbf{P}) \end{aligned} \tag{23}$$

where $\mathbf{D} \triangleq \sum_{i=1}^{s}(\mathbf{A}_i^T(\mathbf{Q}\mathbf{Q}^T - 2\mathbf{I})\mathbf{W}_i^T\mathbf{W}_i\mathbf{Q}\mathbf{Q}^T\mathbf{A}_i)$. Then, the problem (21) can be recast as

$$\mathbf{P}^{k+1} = \arg\min_{\mathbf{P}} F(\mathbf{P}) \triangleq \mathrm{Tr}(\mathbf{P}^T\mathbf{D}\mathbf{P}) \quad \text{s.t. } \mathbf{P}\mathbf{P}^T = \mathbf{I}_r. \tag{24}$$

In (24), the matrix $\mathbf{D}$ is not symmetric. One cannot obtain the solution of (24) by seeking eigenvectors as in (14), because the eigenvalue and eigenvector may not be real for nonsymmetric matrix.

Fortunately, Wen and Yin [26] develop a curvilinear search algorithm for optimization with orthogonality constraints. For (24), given a feasible point $\mathbf{P}_k$ and the gradient matrix $\mathbf{G} \triangleq \nabla F(\mathbf{P}_k)$, the search direction $\mathbf{Y}_k(\tau)$ and step size τ_k of the next iteration point are given by (25) and (26), respectively

$$\mathbf{Y}(\tau) = \left(\mathbf{I} + \tfrac{\tau}{2}\mathbf{A}\right)^{-1}\left(\mathbf{I} - \tfrac{\tau}{2}\mathbf{A}\right)\mathbf{P}_k \tag{25}$$

$$\tau_k = \frac{tr\left(\mathbf{S}_{k-1}^T\mathbf{S}_{k-1}\right)}{\left|tr\left(\mathbf{S}_{k-1}^T\mathbf{Y}_{k-1}\right)\right|} \tag{26}$$

where $\mathbf{A} = \mathbf{G}\mathbf{P}_k^T - \mathbf{P}_k\mathbf{G}^T, \mathbf{S}_{k-1} = \mathbf{P}_k - \mathbf{P}_{k-1}$, and $\mathbf{Y}_{k-1} = \nabla F(\mathbf{P}_k) - \nabla F(\mathbf{P}_{k-1})$.

The curvilinear search algorithm for (24) is shown in Algorithm 2. It should be mentioned that the Algorithm 2 runs very fast and returns solution no worse than those from other state-of-the-art algorithms [26].

Based on the Algorithm 2, the algorithm for solving (19) can be summarized in Algorithm 3.

Problem (20) is equivalent to

$$\min_{\mathbf{Q}} \sum_{i=1}^{s} \left|\left|\mathbf{A}_i^T - \mathbf{P}\mathbf{P}^T\mathbf{A}_i^T\mathbf{Q}\mathbf{Q}^T\right|\right|_* \quad \text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_t. \tag{27}$$

Algorithm 2 Curvilinear Search Method for (24)

1: **Initialize**: $\mathbf{P}_0$, r, stop error ε, $k = 0$;
2: Generate search direction $\mathbf{Y}(\tau)$ by (25);
3: Chose a proper step size τ_k by (26);
4: Update $\mathbf{P}_{k+1} = \mathbf{Y}(\tau_k)$;
5: Stopping check. If $||\nabla F(\mathbf{P}_{k+1})|| \le \varepsilon$, stop; otherwise, $k \leftarrow k+1$ and go to step 2.
6: **Output:** Optimal matrix $\mathbf{P}_{k+1}$.

Algorithm 3 Iteratively Reweighted Method for (19)

Input: Training data $\mathbf{A}_1, \ldots, \mathbf{A}_s$, projection number r, left projection matrix $\mathbf{Q}$.

1: Initialize: $\mathbf{W}_{i=1:s}^0 = \mathbf{I}$, $\varepsilon_{i=1:s}^0 = 1$, $k = 0$, $K = 4$, $\mathbf{P}^0$;
2: Update $\mathbf{P}$ by Algorithm 2;
3: Update $\mathbf{W}_i$: $\mathbf{W}_i^{k+1} = \left(\left(\mathbf{X}_i^{k+1}\right)_{\varepsilon_i^{k+1}}\left(\mathbf{X}_i^{k+1}\right)_{\varepsilon_i^{k+1}}^T\right)^{-1/4}$; where $\mathbf{X}_i^{k+1} = \mathbf{A}_i - \mathbf{Q}\mathbf{Q}^T A_i\mathbf{P}^{k+1}(\mathbf{P}^{k+1})^T, \varepsilon_i^{k+1} = \min\left\{\varepsilon_i^k, \sigma_K\left(\mathbf{X}_i^{k+1}\right)\right\}$;
4: If $\varepsilon_i = 0$, go to 5; otherwise go to 2.
5: **Output:** Optimal projection matrix $\mathbf{P}^{k+1}$.

Algorithm 4 Alternatively Iterative Method for (18)

Input: Training data $\mathbf{A}_1, \ldots, \mathbf{A}_s$, projection numbers r, t

1: Initialize: $\mathbf{Q}^0 = \mathbf{I}$, $k = 0$;
2: Update $\mathbf{P}^{k+1}$ using the Algorithm 3;
3: Update $\mathbf{Q}^{k+1}$ using the Algorithm 3;
4: If the criterion (29) is satisfied, go to 5; otherwise go to 2
5: **Output:** Optimal projection matrices $\mathbf{P}^{k+1}, \mathbf{Q}^{k+1}$.

Obviously, Algorithm 3 can be also used to solve (27). One only needs to replace the input $\{\mathbf{A}_1, \ldots, \mathbf{A}_s, r, \mathbf{Q}\}$ of Algorithm 3 with the new input $\{\mathbf{A}_1^T, \ldots, \mathbf{A}_s^T, t, \mathbf{P}\}$.

B. Algorithm for N-B2-DPCA

The algorithm for N-B2-DPCA can be summarized in Algorithm 4. In Algorithm 4, we use the relative reduction of the mean reconstruction error value to check the convergence of N-B2-DPCA. The mean reconstruction error in step k is defined as

$$\mathrm{mre}(k) = \tfrac{1}{s}\sum_{i=1}^{s}\left|\left|\mathbf{A}_i - \mathbf{Q}^k(\mathbf{Q}^k)^T\mathbf{A}_i\mathbf{P}^k(\mathbf{P}^k)^T\right|\right|_*. \tag{28}$$

The convergence of Algorithm 4 can be judged by the relative difference-based convergence criterion

$$\left|\frac{\mathrm{mre}(k-1) - \mathrm{mre}(k)}{\mathrm{mre}(k-1)}\right| \le \mu \tag{29}$$

where μ is a small positive number.

After obtaining the projection matrix $\mathbf{P}$ and $\mathbf{Q}$ via Algorithm 4, for a given image sample $\mathbf{A}$, the feature matrix $\mathbf{C}$ of the image sample $\mathbf{A}$ is obtained by $\mathbf{C} = \mathbf{Q}^T\mathbf{A}\mathbf{P}$. The feature matrix $\mathbf{C}$ is used to represent image $\mathbf{A}$ for classification.

Fig. 5. Sample images from five subsets of the Extended Yale B database. (a) Subset 1. (b) Subset 2. (c) Subset 3. (d) Subset 4. (e) Subset 5.

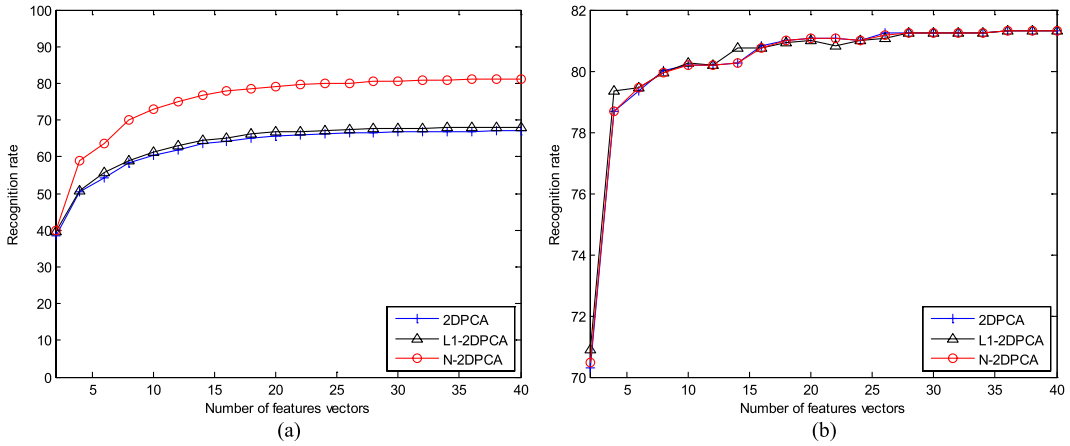


Fig. 6. Recognition rates of 2-D-based methods with varying feature number on the Extended Yale B database under (a) and (b) NN classifier and SVM.

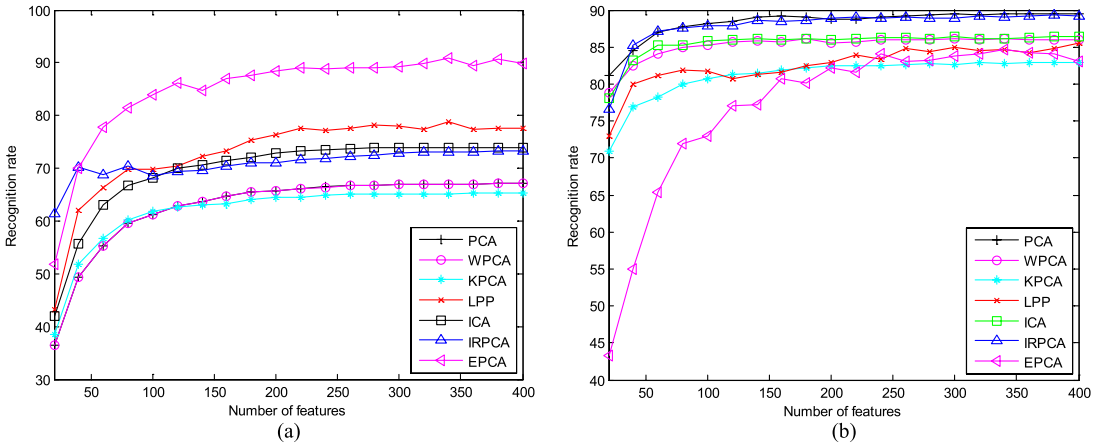


Fig. 7. Recognition rates of 1-D-based methods with varying feature number on the Extended Yale B database under (a) and (b) NN classifier and SVM.

V. Experiments

The proposed N-2-DPCA and N-B2-DPCA are evaluated and compared with the other feature extraction or classification methods, including 1-D-based methods, 2-D-based methods, and bilateral 2-D-based methods, on four well-known face image databases: 1) the CMU PIE [27]; 2) the Extended Yale B [28]; 3) the FRGC [29]; and 4) the AR database [53]. The 1-D-based methods include PCA, WPCA [45], KPCA [44], LPP [32], ICA [47], inductive robust principal component analysis (IRPCA) [38], two-stage sparse representation classifier (TSR) [39], EPCA [36], and PCA + LDA [50]. The 2-D-based methods include 2-DPCA and L_1-2-DPCA. The bilateral 2-D-based methods include B2-DPCA and DCT [49]. Finally, the nearest neighbor (NN) classifier and support vector machine (SVM) are employed for classification, respectively. The MATLAB code of our methods is available at: http://pcalab.chinandy.com/pcalab/code/code_Nuclear_2-DPCA.zip

A. Experiments on the Extended Yale B Database

In the Extended Yale B database [28], there are 38 subjects. Every image is resized to 96 × 84. The database is divided into five subsets according to different lighting conditions. The two sample images of each subset are shown in Fig. 5. For each subject, half of the images are randomly selected for training (i.e., 32 images per subject) and the rest images for testing. We let the number of features vary from 20 to 400 with an interval 20 for 1-D-based methods, while the varying number of features from 2 to 40 with an interval 2 for 2-D-based methods.

Figs. 6 and 7 show recognition rates of all methods with varying features number of 2-D-based methods and 1-D-based methods under two classifiers: 1) the NN classifier and 2) SVM. Tables III and IV list the recognition rates at varying features number of bilateral 2-D-based methods. Table V lists the top recognition rate of each method and the corresponding dimension and running time.

TABLE III
RECOGNITION RATES (%) OF BILATERAL 2-D-BASED METHODS WITH THE NN CLASSIFIER

Right / Left	8			16			24			32		
	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA
8	41.28	43.67	52.14	46.38	48.11	58.63	47.62	49.18	59.95	48.11	49.42	60.61
16	54.03	54.85	66.04	58.39	60.12	74.10	59.95	61.02	76.56	60.36	61.35	76.89
24	57.07	57.24	68.59	62.34	62.66	77.06	64.31	64.23	78.70	64.56	64.72	79.36
32	58.06	57.32	69.08	63.65	63.73	77.06	65.46	65.46	79.44	66.04	66.12	80.43

TABLE IV
RECOGNITION RATES (%) OF BILATERAL 2-D-BASED METHODS WITH SVM

Right / Left	8			16			24			32		
	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA
8	84.62	85.53	85.77	86.35	86.35	85.94	86.51	86.43	85.94	86.43	86.27	86.27
16	87.75	88.16	88.32	87.99	89.06	88.90	88.32	88.98	88.98	88.32	88.90	88.90
24	88.73	88.73	88.32	89.47	89.47	89.47	89.23	89.23	89.14	89.64	89.47	89.47
32	89.06	88.73	88.32	89.31	89.14	89.23	89.72	89.47	89.56	89.88	90.05	89.88

TABLE V
TOP RECOGNITION RATE, CORRESPONDING DIMENSION, AND RUNNING TIME OF EACH METHOD ON THE EXTENDED YALE B DATABASE UNDER THE NN CLASSIFIER AND SVM

	NN			SVM		
Method	Rate	Dim	Times(s)	Rate	Dim	Times(s)
PCA	67.11	380	37.5	89.47	300	41.6
WPCA	67.11	380	70.1	86.10	180	76.3
KPCA	65.38	380	14.1	82.98	360	14.2
LPP	78.87	340	77.5	85.53	400	86.6
ICA	73.85	280	199.9	86.43	380	346.6
IRPCA	73.27	380	9179.3	89.31	380	9179.3
EPCA	90.95	340	84.9	84.70	340	84.9
TSR	67.19	20x20	548.3	\	\	\
2DPCA	67.02	96x36	1.0	81.33	96x34	1.0
L_1-2DPCA	68.02	96x36	1310.4	81.33	96x34	1254.9
DCT	66.69	40x40	1.67	90.21	34x38	1.67
B2DPCA	66.78	40x36	6.32	90.05	34x30	6.46
N-2DPCA	81.09	96x34	227.9	81.33	96x34	227.9
N-B2DPCA	80.92	40x40	167.18	90.30	38x34	181.97
PCA+LDA	92.11	31	15.0	91.86	34	14.8
N-2DPCA+ LDA	93.09	37	243	92.02	28	289.6

For 2-D-based and bilateral 2-D-based methods, N-2-DPCA and N-B2-DPCA outperform others with the NN classifier, as shown in Fig. 6(a) and Table III. However, from Fig. 6(b) and Table IV, we can see that when SVM is used, there is no significant performance difference between these methods.

For 1-D-based methods in Fig. 7(a), we can see that EPCA achieves better performance than the other methods with the NN classifier. The possible reason is EPCA utilizes a robust dissimilarity measure based on the Euler representation of complex numbers, and EPCA retains PCAs desirable properties while suppressing outliers [36]. However, when SVM is used for classification, EPCA performs worse than N-2-DPCA and N-B2-DPCA, as shown in Fig. 7(b).

It should be mentioned that in this paper, we focus on the unsupervised methods. It is unfair to compare an unsupervised method, such as N-2-DPCA, with supervised methods, such as LDA directly. Since LDA is always done in the PCA-transformed space, i.e., PCA + LDA, we here implement N-2-DPCA + LDA for fair comparison. Table V shows that N-2-DPCA + LDA achieves a slightly better result than LDA.

In the second experiment, we evaluate the robustness of each method to noise caused by synthetic occlusions. Images in Subsets 1 and 2 are used for training and the images in Subset 3 for testing. Here, all testing samples and half training samples are imposed the black block occlusion with varying block size. The block size determines the occlusion rate of an image. Fig. 8(a) shows images with occlusion rates from 10% to 60%.

Fig. 8(b) shows recognition rates of all methods under the NN classifier with different occlusion rates. It can be seen that N-2-DPCA + LDA outperforms other methods in all cases. Except two supervised models, N-2-DPCA + LDA and PCA + LDA, N-2-DPCA and N-B2-DPCA achieves the best recognition results when the occlusion rate is no more than 40%. It should be noted that EPCA outperforms N-2-DPCA and N-B2-DPCA when the occlusion rate is more than 40%. However, when SVM is used for classification, we can see from Fig. 8(c) there is no significant performance difference between various methods.

B. Experiments on the CMU PIE Databases

For CMU PIE database [27], there are 68 subjects in total, and images of each person were taken across 13 different

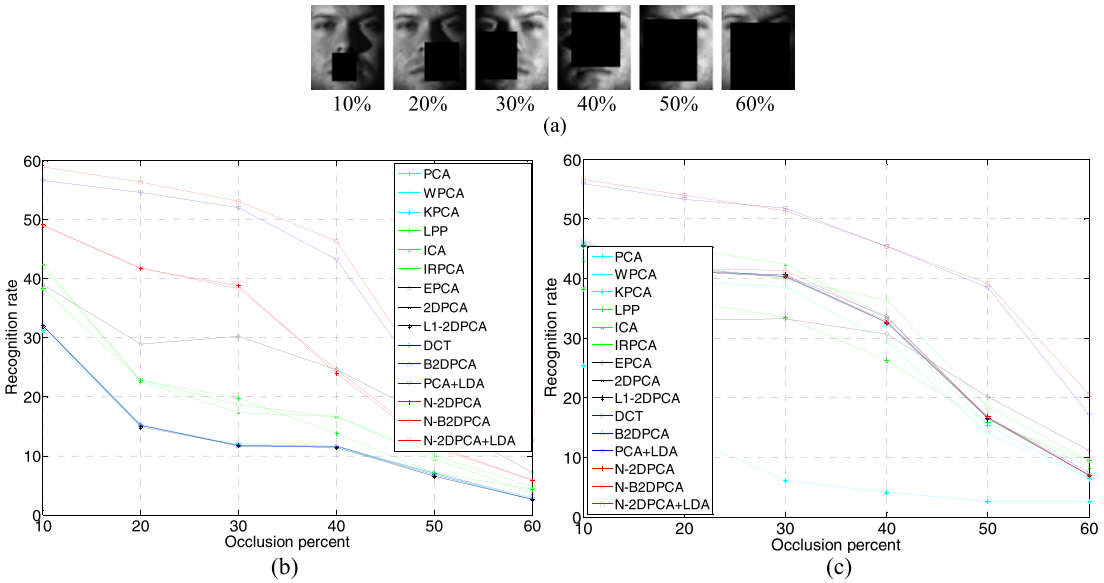


Fig. 8. Recognition rates (%) of all methods with different occlusion rates under the NN classifier and SVM. (a) Occluded images with different occlusion rates. (b) NN. (c) SVM.

poses, under 43 different lighting conditions, and with four different expressions. We choose one subset (Pose C9) for experiment. Each image is resized to 64×64 pixels. For each subject, half of the images are randomly selected for training (i.e., 12 images per subject), and the rest 12 images for testing.

Here, we use the same methodology as adopted in the first experiment on the Extended Yale B database. The top recognition rate of each method, corresponding dimension, and running time are listed in Table VI. It is obvious that N-2-DPCA + LDA achieves best performance among all methods. When the NN classifier is used for classification, N-B2-DPCA outperforms most competing methods except the two supervised ones.

The convergences of Algorithms 1 and 4 are shown in Fig. 9. We can see that they both converge after three iterations.

Tables VIII and IX show the computation time of each method with varying number of features (or components) N-2-DPCA and N-B2-DPCA consume less CPU time than L_1-2-DPCA and IRPCA. Due to the computation of singular values in iteration steps, N-2-DPCA and N-B2-DPCA are computationally more expensive than 2-DPCA and B2-DPCA.

C. Experiments on Large-Scale Face Database: FRGC

The FRGC version2.0 is a large scale face image database, including controlled and uncontrolled images [29], [30]. This database contains 12 776 training images (6360 controlled images and 6416 uncontrolled ones) from 222 individuals, 16 028 controlled target images, and 8014 uncontrolled query images from 466 persons for the FRGCv2.0 Experiment 4. The controlled images have good image quality, while the uncontrolled images display poor image quality, such as large illumination variations, low resolution of the face region, and possible blurring. We use a subset (222 subjects having 36 samples in the training set) of the Experiment 4. The face region of each image is first cropped from the original high-resolution still images and resized to a spatial resolution of 32×32. No further preprocessing is applied.

For each subject, half of the images are randomly selected for training (i.e., 18 images per subject) and the rest 18 images for testing. We use the same methodology as adopted in the

TABLE VI
TOP RECOGNITION RATE, CORRESPONDING DIMENSION, AND RUNNING TIME OF EACH METHOD ON THE CMU PIE DATABASE UNDER THE NN CLASSIFIER AND SVM

	NN			SVM		
Method	Rate	Dim	Time(s)	Rate	Dim	Times(s)
PCA	75.98	160	4.4	92.03	320	4.3
WPCA	76.10	340	16.4	92.03	380	16.4
KPCA	75.98	160	3.2	92.03	320	2.9
LPP	87.01	240	11.7	90.93	80	7.1
ICA	84.93	120	5.3	92.52	120	5.3
IRPCA	85.17	340	718.5	92.52	320	696.4
EPCA	85.54	320	28.5	76.10	280	20.5
TSR	61.40	20x20	572.6	\	\	\
2DPCA	77.57	64x26	0.6	92.28	64x12	0.6
L_1-2DPCA	78.57	64x26	454.4	92.28	64x12	186.6
DCT	77.70	36x32	0.7	92.52	10x32	0.7
B2DPCA	77.70	38x32	2.0	92.40	14x12	1.9
N-2DPCA	89.95	64x30	70.9	92.28	64x12	85
N-B2DPCA	90.07	38x34	205.6	92.40	10x24	55.3
PCA+LDA	92.52	34	3.3	92.40	40	4.5
N-2DPCA+ LDA	94.63	67	88.7	93.63	16	87.6

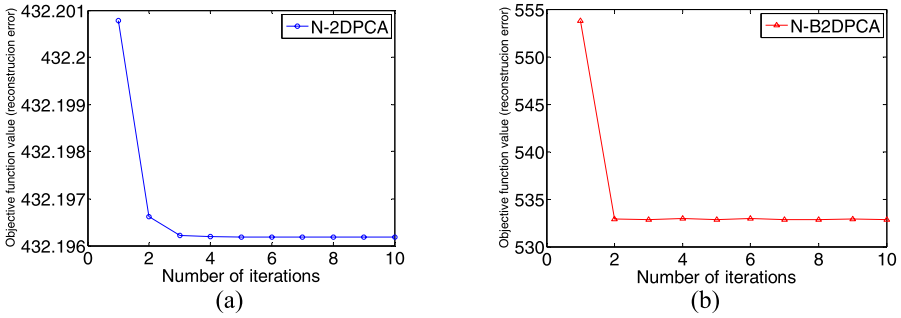


Fig. 9. Objective function values (reconstruction errors) versus iteration times on the CMU PIE database. (a) N-2-DPCA, projection number $r = 20$. (b) N-B2-DPCA, $r = 20$, $t = 20$.

TABLE VII
TOP RECOGNITION RATE, CORRESPONDING DIMENSION, AND RUNNING TIME OF EACH METHOD ON THE FRGC DATABASE UNDER THE NN CLASSIFIER AND SVM

	NN			SVM		
Method	Rate	Dim	Time(s)	Rate	Dim	Time(s)
PCA	73.32	280	6.4	92.09	400	6.0
WPCA	73.32	280	339.3	92.09	400	298.7
KPCA	73.35	400	252.7	77.63	400	252.7
LPP	83.66	240	217.3	92.77	140	148.5
ICA	81.66	380	173.5	92.64	320	138.3
IRPCA	81.36	360	1671.3	92.49	380	1655.3
EPCA	86.19	320	182.5	87.64	400	181.0
TSR	76.82	20x20	1682.2	\	\	\
2DPCA	73.30	32x20	0.6	92.07	32x20	0.6
L_1-2DPCA	74.30	32x20	618.1	92.07	32x20	618.1
DCT	73.30	20x20	2.2	92.34	14x20	2.3
B2DPCA	73.22	20x20	2.3	91.99	16x20	2.1
N-2DPCA	82.88	32x18	9.2	92.07	32x20	8.7
N-B2DPCA	83.03	18x20	63.4	92.92	16x20	69.4
PCA+LDA	96.17	101	486.2	97.17	81	532.8
N-2DPCA+ LDA	97.30	51	491.0	97.70	121	490.6

TABLE VIII
RUNNING TIME(S) OF 1-D-BASED METHODS ON THE CMU PIE DATABASE

Feature Number	PCA	WPCA	KPCA	LPP	ICA	IRPCA	EPCA
20	1.1	10.8	2.6	2.3	1.6	715.8	1.2
200	3.9	17.3	2.4	10.4	13.6	706.8	11.8
400	7.4	26.4	2.4	27.5	76.7	718.5	53.9

first experiment on the Extended Yale B database. The top recognition rate of each method, the corresponding dimension, and running time are listed in Table VII. It is obvious that N-2-DPCA + PCA achieves better performance than other methods in Table VII. When the NN classifier is used for classification, except the two supervised models, there

TABLE IX
RUNNING TIME(S) OF 2-D-BASED METHODS ON THE CMU PIE DATABASE FOR B2-DPCA AND N-B2-DPCA, THE LEFT PROJECTION COMPONENT NUMBER IS FIXED 20

Component Number	2DPCA	L_1-2DPCA	B2DPCA	N-2DPCA	N-B2DPCA
10	0.1	76.6	0.6	40.9	25.6
20	0.1	165.0	0.6	42.4	26.5
30	0.1	256.3	0.7	34.0	26.4

(a) (b) (c) (d) (e) (f) (g)

Fig. 10. (a)–(g) Images of one person in the AR face database.

Fig. 11. Kinds of objects used for occlusion of samples.

Fig. 12. Some outlying images of the AR database.

is only one method, i.e., EPCA, outperforming our method N-B2-DPCA. In general, N-2-DPCA needs more coefficients for image representation than PCA. However, its bilaterally extended version, N-B2-DPCA can overcome this drawback, since it needs as less coefficients as PCA for image representation.

D. Experiments on the AR Database

For the classification task, it is generally believed that supervised algorithms (e.g., LDA) are superior to unsupervised algorithms (e.g., PCA). The above experimental results reinforce this judgment again. However, it is not always the case, in [52], the authors concluded that when the training data set is small, PCA can outperform LDA. In this section, it is verified that proposed N-2-DPCA and N-B2-DPCA can

TABLE X
AVERAGE RECOGNITION RATE ON THE AR DATABASE UNDER THE NN CLASSIFIER AND SVM

method	PCA	WPCA	KPCA	LPP	ICA	IRPCA	EPCA	2DPCA	L_1-2DPCA	DCT	B2DPCA	N-2DPCA	N-B2DPCA	PCA +LDA	N-2DPCA +LDA
NN	72.17	72.19	73.48	72.72	72.23	72.17	68.63	73.09	73.05	74.10	73.49	85.73	**87.26**	82.25	83.28
SVM	80.00	80.00	80.67	81.18	80.00	80.00	71.98	80.13	80.11	80.53	80.30	83.15	**84.36**	82.34	83.24

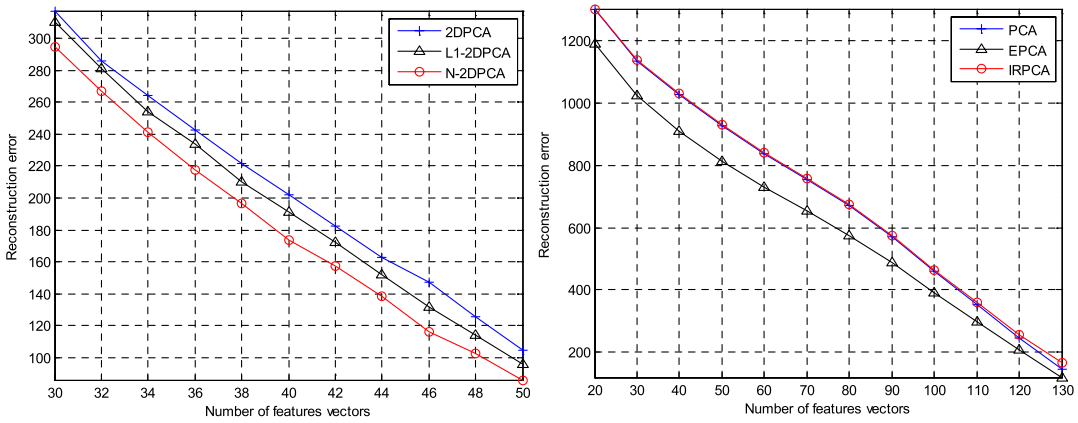


Fig. 13. Average reconstruction error versus feature number. (a) 2-D-based methods. (b) 1-D-based methods.

also outperform supervised algorithms when the size of the training database is small.

We here experiment on the AR database [53], which contains over 4000 color face images of 126 people with different facial expressions, lighting conditions, and occlusions (by sunglass or scarf). The pictures of most persons were taken in two sessions (separated by two weeks). Each section contains 13 images.

The experiments setting are similar with that in [52]. For each individual, only the nonoccluded images of two sessions are used. Each image is cropped and resized to 85×60 pixels. The images of one person are shown in Fig. 10(a)–(g) and the details of the images are neutral expression, smile, anger, scream, left light on, right light on, and all sides light on.

To simulate the effects of a small training data set, 50 different individuals were randomly selected, and two images per person are used for training and five for testing. There are a total of 21 ways of selecting two for training and five for testing. We will use all these 21 different ways of separating the data into the training and the testing parts.

To each of the 21 different training and testing data sets created in the manner described above, the top recognition rate is computed. Table X lists the average recognition rates of 21 ways. These results show that un-supervised methods (N-2-DPCA and N-B2-DPCA) outperform supervised methods (e.g., PCA + LDA and N-2-DPCA + LDA).

In the second experiment, as done in [16], we compare the performances of different methods in terms of reconstruction error for inliers when outliers are present in the training data. Let $\{\mathbf{A}_1, \ldots, \mathbf{A}_t, \mathbf{A}_{t+1}, \ldots, \mathbf{A}_s\}$ be the training data, where the first t samples are nonoutliers and the last s-t are outliers. The averaged recovered error for nonoutliers is defined by

$$e \triangleq \frac{1}{t} \sum_{i=1}^{t} \left|\left|\mathbf{A}_i - \mathbf{A}_i^{\text{rec}}\right|\right|_F \tag{30}$$

where $\mathbf{A}_i^{\text{rec}}$ are the reconstructed image.

For the training set, 140 images of 10 subjects are used. Each subject contains 14 images nonoccluded images from two sessions. Sample images in one session are shown in Fig. 10. Our purpose is to investigate the reconstruction capacity of the projection matrix learned from the training set when the training set is spoiled by outliers and 40% of images in the training set were used to generate the outliers. Each outlying image was formed by adding block noise. The occlusion size is 30% of the image size. Ten kinds of blocks can be used, as shown in Fig. 11. Fig. 12 shows some samples of the outliers (outlying images).

The average reconstruction error of different methods for nonoutlying images is shown in Fig. 13 and Table XI. One can see that: 1) for 2-D-based methods, N-2-DPCA outperforms L_1-2-DPCA and 2-DPCA; 2) for bilateral 2-D-based methods, the proposed N-B2-DPCA outperforms B2-DPCA and DCT; and 3) for 1-D-based methods, EPCA outperform PCA and IRPCA.

Fig. 14 shows the reconstructed images of different methods. The first column gives two original images. The other columns are the corresponding images reconstructed by different algorithms. One can see the following.

1) For 1-D-based method, the reconstructed images are seriously affected by outliers and some important detailed information of images is lost.

TABLE XI
AVERAGE RECONSTRUCTION ERROR OF BILATERAL 2-D-BASED METHODS

Right / Left	25			35			45		
	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA	DCT	B2DPCA	N-B2DPCA
25	735.4	613.7	**603.9**	663.4	551.6	**544.2**	634.8	523.5	**520.1**
35	582.4	503.9	**489.6**	476.3	411.0	**401.2**	428.4	366.0	**365.2**
45	514.1	451.8	**440.5**	381.6	336.2	**325.9**	314.5	275.9	**270.1**

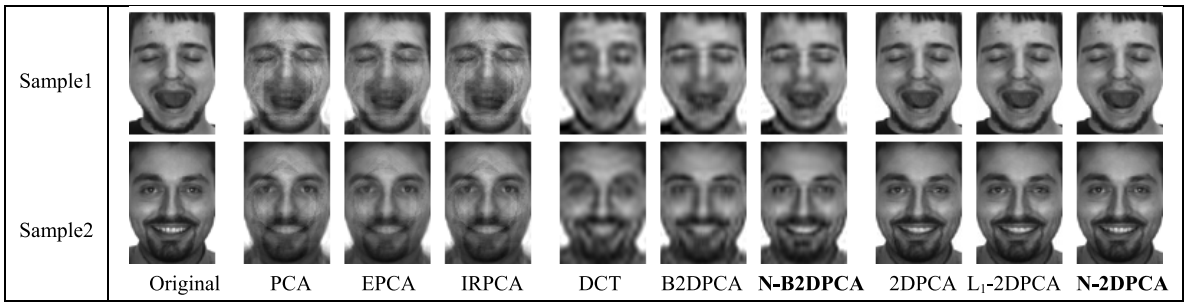


Fig. 14. Original images (first column) and reconstructed images (other columns) by different methods.

2) For bilateral 2-D-based methods, the reconstructed images by N-B2-DPCA are clearer than that by DCT and B2-DPCA.
3) The reconstructed images by 2-D-based methods are far superior to that by 1-D-based methods and bilateral 2-D-based methods.

VI. CONCLUSION

This paper presents a 2-D-based subspace learning model, namely, the nuclear norm-based 2-DPCA for extracting features from images. The key idea of the model is to use the nuclear norm instead of the F-norm to measure the reconstruction error. The model is solved via the iteratively reweighted algorithm. In addition, N-2-DPCA is further extended to N-B2-DPCA, which achieves a higher compression rate than N-2-DPCA. Experimental results on face image databases show the proposed methods performs better than or comparably with state-of-the-art feature extraction methods.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions.

REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
[2] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.
[3] Z. Lai, W. Wong, Z. Jin, J. Yang, and Y. Xu, "Sparse approximation to the eigensubspace for discrimination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1948–1960, Dec. 2012.
[4] M. Visani, C. Garcia, and C. Laurent, "Comparing robustness of two-dimensional PCA and eigenfaces for face recognition," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 3212. Berlin, Germany: Springer-Verlag, 2004, pp. 717–724.
[5] H. Kong, L. Wang, E. K. Teoh, X. Li, J.-G. Wang, and R. Venkateswarlu, "Generalized 2D principal component analysis for face image representation and recognition," *Neural Netw.*, vol. 18, nos. 5–6, pp. 585–594, Jul./Aug. 2005.
[6] D. Q. Zhang and Z.-H. Zhou, "$(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, Dec. 2005.
[7] J. Yang and C. J. Liu, "Horizontal and vertical 2DPCA-based discriminant analysis for face verification on a large-scale database," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 4, pp. 781–792, Dec. 2007.
[8] L. Wang, X. Wang, X. Zhang, and J. Feng, "The equivalence of two-dimensional PCA to line-based PCA," *Pattern Recognit. Lett.*, vol. 26, no. 1, pp. 57–60, Jan. 2005.
[9] Q.-X. Gao, "Is two-dimensional PCA equivalent to a special case of modular PCA?" *Pattern Recognit. Lett.*, vol. 28, no. 10, pp. 1250–1251, Jul. 2007.
[10] L. Wang, X. Wang, and J. Feng, "On image matrix based feature extraction algorithms," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 1, pp. 194–197, Feb. 2006.
[11] A. Mashhoori and M. Z. Jahromi, "Block-wise two-directional 2DPCA with ensemble learning for face recognition," *Neurocomputing*, vol. 108, no. 2, pp. 111–117, May 2013.
[12] H. Wang, "Block principal component analysis with L_1-norm for image analysis," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 537–542, Apr. 2012.
[13] D. Wang and H. Lu, "Object tracking via 2DPCA and ℓ_1-regularization," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 711–714, Nov. 2012.
[14] N. Kwak, "Principal component analysis based on L_1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.
[15] L. Xu and A. L. Yuille, "Robust principal component analysis by self-organizing rules based on statistical physics approach," *IEEE Trans. Neural Netw.*, vol. 6, no. 1, pp. 131–143, Jan. 1995.
[16] X. Li, Y. Pang, and Y. Yuan, "L_1-norm-based 2DPCA," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1170–1175, Aug. 2009.
[17] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721–732, Jul. 1997.
[18] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[19] Z. Gu, M. Shao, L. Li, and Y. Fu, "Discriminative metric: Schatten norm vs. vector norm," in *Proc. 21st IEEE Conf. Pattern Recognit.*, Nov. 2012, pp. 1213–1216.

[20] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2002.
[21] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proc. Amer. Control Conf.*, 2001, pp. 4734–4739.
[22] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[23] R. He, Z. Sun, T. Tan, and W.-S. Zheng, "Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2889–2896.
[24] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, May 2011, Art. ID 11.
[25] M. Fornasier, H. Rauhut, and R. Ward, "Low-rank matrix recovery via iteratively reweighted least squares minimization," *SIAM J. Optim.*, vol. 21, no. 4, pp. 1614–1640, Oct. 2011.
[26] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, nos. 1–2, pp. 397–434, 2012. [Online]. Available: http://optman.blogs.rice.edu/
[27] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[28] K.-C. Lee, J. Ho, and D. Driegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[29] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 947–954.
[30] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, and W. Worek, "Preliminary face recognition grand challenge results," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, Southampton, U.K., 2006, pp. 15–24.
[31] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[32] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[33] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.
[34] Z. Lai, M. Wan, Z. Jin, and J. Yang, "Sparse two-dimensional local discriminant projections for feature extraction," *Neurocomputing*, vol. 74, no. 4, pp. 629–637, 2011.
[35] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1023–1035, Jul. 2013.
[36] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Euler principal component analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 498–518, 2013.
[37] W. Zuo, D. Zhang, and K. Wang, "An assembled matrix distance metric for 2DPCA-based image recognition," *Pattern Recognit. Lett.*, vol. 27, no. 3, pp. 210–216, 2006.
[38] B.-K. Bao, G. Liu, C. Xu, and S. Yan, "Inductive robust principal component analysis," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3794–3800, Aug. 2012.
[39] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Two-stage nonnegative sparse representation for large-scale face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 35–46, Jan. 2013.
[40] Z. Xu, J. Zhang, and X. Dai, "Boosting for learning a similarity measure in 2DPCA based face recognition," in *Proc. WRI World Congr. Comput. Sci. Inf. Eng.*, 2009, pp. 130–134.
[41] U. I. Bajwa, I. A. Taj, M. W. Anwar, and X. Wang, "A multifaceted independent performance analysis of facial subspace recognition algorithms," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. ID e56510.
[42] J. Gao, "Robust L_1 principal component analysis and its Bayesian variational inference," *Neural Comput.*, vol. 20, no. 2, pp. 555–572, 2008.
[43] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 586–591.
[44] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.
[45] Y. Koren and L. Carmel, "Robust linear dimensionality reduction," *IEEE Trans. Vis. Comput. Graphics*, vol. 10, no. 4, pp. 459–470, Jul./Aug. 2004.
[46] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and J. Zhong, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.
[47] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.
[48] J. Yang, D. Zhang, and J.-Y. Yang, "Constructing PCA baseline algorithms to reevaluate ICA-based face-recognition performance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 4, pp. 1015–1021, Aug. 2007.
[49] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *Int. J. Comput. Vis.*, vol. 43, no. 3, pp. 167–188, 2001.
[50] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[51] J. Yang and J.-Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognit.*, vol. 36, no. 2, pp. 563–566, 2003.
[52] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[53] A. M. Martinez and R. Benavente, "The AR face database," CVC, New Delhi, India, Tech. Rep. 24, Jun. 1998.

Fanlong Zhang received his B.S. and M.S. degrees in applied mathematics from the Liaocheng University in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

His current research interests include pattern recognition and optimization.

Jian Yang (M'08) received the B.S. degree in mathematics from Xuzhou Normal University, Xuzhou, China, in 1995, the M.S. degree in applied mathematics from Changsha Railway University, Changsha, China, in 1998, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

He was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain, in 2003. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, Hong Kong Polytechnic University, Hong Kong, and the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA, from 2006 to 2007. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored over 80 scientific papers in pattern recognition and computer vision. His journal papers have been cited over 3000 times in the ISI Web of Science, and 7000 times on Google Scholar. His current research interests include pattern recognition, computer vision, and machine learning.

Prof. Yang is also an Associate Editor of *Pattern Recognition Letters* and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, respectively.

Jianjun Qian received the BS degree in computer science from the Zhengzhou University in 2007. He received the MS degree in computer software and theory from Northwest University for Nationalities in 2010.

He is currently an Assistant Professor with the School of Computer Science and Engineering, NUST. His current research interests include pattern recognition, computer vision, and face recognition in particular.

Yong Xu (M'06) received his B.S. and M.S. degrees from the Air Force Institute of Meteorology in1994 and 1997, respectively. He received the B.S. and M.S. degrees, in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 2005.

He is currently with the Bio-Computing Research Center, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, biometrics, machine learning, image processing, and video analysis.