Orthogonal Principal Coefficients Embedding for Unsupervised Subspace Learning

Xinxing Xu, Shijie Xiao, Zhang Yi, *Fellow IEEE,* Xi Peng, and Yong Liu

***Abstract*—As a recently proposed method for subspace learning, principal coefficients embedding (PCE) method can automatically determine the dimension of the feature space and robustly handle various corruptions in real-world applications. However, the projection matrix learned by PCE is not orthogonal, so the original data may be reconstructed improperly. To address this issue, we proposed a new method termed orthogonal principal coefficients embedding (OPCE). OPCE can not only automatically determine the dimension of the feature space, but also additionally considers the orthogonal property of the projection matrix for better discriminating ability. Moreover, OPCE can be solved in closed-form, thus making it computational efficient. Extensive experimental results from multiple benchmark data sets demonstrate the effectiveness and computational efficiency of the proposed method.**

***Index Terms*—Unsupervised feature extraction, bio-inspired data representation, dimension reduction, manifold learning.**

I. Introduction

Dimensional reduction methods have been widely applied to the machine intelligence and pattern recognition problems such as face recognition and gait recognition. The key idea of dimensional reduction is to learn a projection matrix that can map data from high-dimensional space into a low dimensional one. For the face recognition problem, Eigenface [1] was proposed to use the principal component analysis (PCA) to project the data into a low-dimensional feature space. The Fisherface [2] method applied the linear discriminant analysis (LDA) by utilizing the class label information for better discriminative ability. It is known that PCA, LDA, and their extensions [3] are based on the global structure of the data and may not consider local structure of the data.

In addition, there are also methods that learn the projection by considering local structure of the data. The locally linear embedding (LLE) [4] method utilizes the local structure of the data by reconstructing the data point using its neighborhood points. The neighborhood preserving embedding (NPE) [5] method learns a projection matrix for out-of-sample extension based on LLE. The locality preserving projections (LPP) [6] approach considers the locality by embedding the graph Laplacian matrix constructed from the data. Following LPP, the orthogonal locality preserving projections (OLPP) [7] method was proposed to enforce that the basis of the projection matrix are orthogonal for better discriminating ability.

Motivated by LLE, the sparsity preserving projections (SPP) [8] and L1Graph [9] were proposed to obtain the reconstruction combination coefficients based on the ℓ_1-norm regularized sparse coding algorithms. These methods can automatically select the basis for the reconstruction instead of specifying the neighborhood number as in the aforementioned algorithms. The nuclear norm regularization based algorithm robust principal component analysis (RPCA) [10] was proposed to recover the low-rank matrix from corrupted data by learning the robust subspace. L2graph [11]–[13] was proposed to eliminate the errors from projection space with ℓ_2-norm regularization, which have shown the state-of-the-art performance in clustering and feature extraction. Motivated by the great success of manifold learning and regularization techniques, many methods have been proposed and achieved impressive performance in manifold ranking [14], blind deconvolution [15], spectral embedding [16], [17], and so on.

Different from most existing subspace learning methods that usually require specifying the dimension of features for the projection or the number of neighbors for the reconstruction, principal coefficient embedding (PCE) [18] automatically determines the dimension of the subspace learning and handle the gross corruption properly. It firstly learns the reconstruction combination coefficients of the data by minimizing the Frobenius norm of both the reconstruction matrix and the error matrix; and then projects the data based on the learned reconstruction combination coefficients. However, the basis of the projection matrix in PCE is not orthogonal, and thus the learned projection may lack strong discriminating ability. Some works in biology and neuroscience have shown that the orthogonality plays an important role in human's information processing system [19], [20]. Motived by these biological evidences and OLPP, we improve PCE by enforcing the basis of the projection to be mutually orthogonal. To be specific, we propose a new algorithm called orthogonal principal coefficient embedding (OPCE) for dimension reduction. OPCE has the following advantages: (1) it can not only automatically select the dimension of the learned subspace for data with gross noise, but also keep the basis of projection matrix orthogonal for better discriminating ability. (2) it has an efficient closed-form solution, thus making the computation

This work was supported by the National Natural Science Foundation of China under Grant 61432012 and U1435213.

X. Xu and Y. Liu are with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632. E-mail: {xuxinx, liuyong}@ihpc.a-star.edu.sg.

S. Xiao is with the OmniVision Technologies Singapore Pte. Ltd. E-mail: shijie.xiao@ovt.com

Z. Yi is with College of Computer Science, Sichuan University, Chengdu, China 610065 (E-mail: zhangyi@scu.edu.cn).

X. Peng is with College of Computer Science, Sichuan University, Chengdu, China 610065 and with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632. E-mail: pangsaai@gmail.com.

X. Xu and S. Xiao contributed equally to this work.

Corresponding author: Xi Peng (pangsaai@gmail.com).

fast.

The rest of this paper is organized as follows. In Section II, we discuss the related works. In Section III, we briefly review the principal coefficient embedding algorithm. In Section IV, we present the objective function, the optimization as well as the whole algorithm of our method. In Section V, the extensive experimental results for image and document classifications are reported. Finally, the conclusions are given in Section VI.

II. Related Work

LLE was proposed in [4] to firstly learn the linear combination coefficients for data by minimizing the reconstruction error using its neighborhoods. It further embeds data by learning the low-dimensional features from reconstruction coefficients. Although LLE can capture the locally linear information of data, it cannot easily handle the out-of-sample case. To solve this problem, the neighborhood preserving embedding (NPE) [5] method was proposed to learn a projection matrix based on LLE. Instead of constructing the graph using data locally as in LEE and NPE, SPP [8], L1Graph [9], and L2Graph [11], [12] were proposed to construct the reconstruction combination coefficients globally by encouraging the sparsity via ℓ_1- or ℓ_2- norm regularizations. These approaches can automatically select the basis for the reconstruction of each data point, therefore the neighborhood number is not required.

RPCA [10] was proposed to recover the corrupted low-rank matrix by minimizing the nuclear norm of the recovered data matrix and the ℓ_1-norm of the error matrix. Following RPCA, the low-rank representation (LRR) [21] approach was proposed to learn the low-rank representation of data by using the data set itself as the basis for reconstruction. The nuclear-norm of the representation matrix and the group sparsity inducing $\ell_{2,1}$-norm regularization of the error matrix are simultaneously minimized. Similar to the ℓ_1-norm minimization problem, the nuclear norm based minimization problem also requires iterative solution and is quite time consuming.

Different from the existing works for subspace learning, PCE [18] was proposed to learn the linear combination coefficients matrix by removing possible errors from the original data. It utilized the Frobenius-norm regularization for both the combination coefficients matrix and the error matrix instead of the nuclear-norm and the group sparse inducing norm. It can not only eliminate the effect of noise from data, but also demonstrate satisfactory computation efficiency thanks to its closed-form solution. Our work further improves the PCE method by introducing an additional orthogonal constraint for the projection matrix. The orthogonal projection is beneficial for enhancing the discriminating ability.

III. Brief Review of Principal Coefficients Embedding (PCE)

In the following, we denote a matrix using a bold uppercase letter (*e.g.* $\mathbf{X}$) and a vector using a bold lowercase letter (*e.g.* $\mathbf{x}$). We use n and d to denote the number of training samples and the feature dimension of the input data. Besides, the transpose of a matrix/vector is represented as superscript "T", the Frobenius norm is represented as $\|\cdot\|_F$, and the pseudo-inverse of a matrix is denoted as superscript "†". A summary of the notations used in this work is shown in Table I.

TABLE I
The summary of the notations used in this work.

Notation	Definition
n	the number of training samples
d	the feature dimension of input data
$\mathbf{x}_i \in \mathbb{R}^d$	the i-th training sample
$\mathbf{X} \in \mathbb{R}^{d\times n}$	training data matrix with n samples
s	the rank of a given data matrix
$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$	Singular Value Decomposition (SVD)
$\mathbf{E} \in \mathbb{R}^{d\times n}$	the error matrix associated with $\mathbf{X}$
$\mathbf{Z} \in \mathbb{R}^{n\times n}$	the representation of $\mathbf{X}$
σ_i	the i-th singular value
$\mathbf{P} \in \mathbb{R}^{d\times k}$	the projection matrix
T	the transpose of a matrix/vector
†	the pseudoinverse of a matrix

PCE is a recently proposed method for robust subspace clustering, which is based on the minimization with the Frobenius norm instead of ℓ_1- or nuclear-norm. Therefore, it enjoys the fast closed-form solution for obtaining the representation coefficients, without the time-consuming iterative optimization procedure.

Suppose we are given a set of n training samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, PCE learns the linear combination coefficients $\mathbf{Z} \in \mathbb{R}^{n\times n}$ as well as the error matrix $\mathbf{E} \in \mathbb{R}^{d\times n}$ by solving the following minimizing problem:

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E}} \quad & \frac{1}{2}\|\mathbf{Z}\|_F^2 + \frac{\lambda}{2}\|\mathbf{E}\|_F^2 \\ \text{s.t.} \quad & (\mathbf{X} - \mathbf{E}) = (\mathbf{X} - \mathbf{E})\mathbf{Z}, \end{aligned} \tag{1}$$

where $\mathbf{X} - \mathbf{E}$ denotes the clean data, which is the difference between the given data matrix $\mathbf{X}$ and the error matrix $\mathbf{E}$. λ is the regularization parameter. Instead of directly using $\mathbf{X}$, the using of $\mathbf{X} - \mathbf{E}$ for the reconstruction can be more robust to the possible noise in the data. In this model, with Gaussian noise assumption (which is usually the case in real-world data), we adopt the Frobenius norm to penalize the error matrix. Moreover, a benefit thanks to the Frobenius norm is the fast closed-form solution and low rank property, as shown in the following Lemma [18], [22]:

Lemma 1. *Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ be the skinny SVD of the data matrix $\mathbf{X}$, where $\mathbf{U} \in \mathbb{R}^{d\times s}$ and $\mathbf{V} \in \mathbb{R}^{n\times s}$ contain the corresponding left and right singular vectors and $\mathbf{S} \in \mathbb{R}^{s\times s}$ is the diagonal matrix with the singular values $\{\sigma_i\}_{i=1}^s$ of $\mathbf{X}$ sorted in descending order. The unique solution for $\mathbf{Z}$ to problem (1) is given by $\mathbf{Z} = \mathbf{V}_k\mathbf{V}_k^T$, where $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{R}^{n\times k}$, and k can be obtained analytically as:*

$$k = \arg\min_{r\in\{1,2,\ldots,s\}} \left(r + \lambda \sum_{i=r+1}^{s} \sigma_i^2 \right). \tag{2}$$

We can get the optimal coefficient matrix $\mathbf{Z}$ based on the closed-form solution from (1). In the following stage, PCE further obtains a projection matrix $\mathbf{P} \in \mathbb{R}^{d\times k}$ given by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2686983, IEEE Transactions on Cognitive and Developmental Systems

the eigenvectors corresponding to the k largest eigenvalues of the generalized eigen-decomposition problem $\mathbf{XZX}^T\mathbf{p} = \sigma\mathbf{XX}^T\mathbf{p}$. As $(\mathbf{XZX}^T)^\dagger\mathbf{XX}^T$ is not generally symmetric, the basis of $\mathbf{P}$ are not guaranteed to be orthogonal.

IV. Orthogonal Principal Coefficients Embedding

A. The Objective Function

The projection matrix $\mathbf{P}$ in PCE is obtained based on the generalized eigen-decomposition problem. It is obvious that $\mathbf{P}$ is not orthogonal. Motivated by [7], [19], we propose the following optimization problem to learn an orthogonal projection matrix $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_k] \in \mathbb{R}^{d\times k}$ from a given $\mathbf{X}$ and $\mathbf{Z}$:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \frac{1}{2}\|\mathbf{P}^T\mathbf{X} - \mathbf{P}^T\mathbf{XZ}\|_F^2 \\ \text{s.t.,} \quad & \mathbf{p}_i^T\mathbf{XX}^T\mathbf{p}_i = 1, i = 1, \ldots, k, \\ & \mathbf{P}^T\mathbf{P} = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{I} \in \mathbb{R}^{k\times k}$ is the identity matrix, and we also enforce that the basis of $\mathbf{P}$ are orthogonal to each other using the additional orthogonal constraints.

B. The Solution for OPCE

To solve the optimization problem in (3), we obtain the columns $\mathbf{p}_1, \ldots, \mathbf{p}_k$ sequentially.

1) Obtain $\mathbf{p}_1$: From the optimization problem in (3), we can obtain the objective function with respect to $\mathbf{p}_1$ as follows:

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2}\|\mathbf{p}^T\mathbf{X} - \mathbf{p}^T\mathbf{XZ}\|_F^2 \\ \text{s.t.,} \quad & \mathbf{p}^T\mathbf{XX}^T\mathbf{p} = 1, \end{aligned} \tag{4}$$

where we ignore the constraint $\mathbf{p}^T\mathbf{p} = 1$ as it can be automatically satisfied in the solution as shown latter.

To obtain the solution $\mathbf{p}_1$ for the optimization problem in (4), we introduce the Lagrangian multiplier σ and get the following Lagrangian:

$$\mathcal{L} = \frac{1}{2}\|\mathbf{p}^T\mathbf{X} - \mathbf{p}^T\mathbf{XZ}\|_F^2 - \frac{\sigma}{2}(\mathbf{p}^T\mathbf{XX}^T\mathbf{p} - 1).$$

For simplicity, let us define a matrix $\mathbf{R}$ as $\mathbf{R} = \mathbf{X} - \mathbf{XZ}$. By setting the derivative of $\mathcal{L}$ with respect to $\mathbf{p}$ as zeros, we arrive at

$$\frac{\partial\mathcal{L}}{\partial\mathbf{p}} = \mathbf{RR}^T\mathbf{p} - \sigma\mathbf{XX}^T\mathbf{p} = 0,$$

based on which we can further obtain

$$\sigma = \frac{\mathbf{p}^T\mathbf{RR}^T\mathbf{p}}{\mathbf{p}^T\mathbf{XX}^T\mathbf{p}}.$$

Then, we have:

$$\mathcal{L} = \frac{1}{2}\mathbf{p}^T\mathbf{RR}^T\mathbf{p} - \frac{\mathbf{p}^T\mathbf{RR}^T\mathbf{p}}{2\mathbf{p}^T\mathbf{XX}^T\mathbf{p}}(\mathbf{p}^T\mathbf{XX}^T\mathbf{p} - 1) = \frac{1}{2}\sigma. \tag{5}$$

Thus, $\mathbf{p}_1$ is the eigenvector corresponding to the minimum nonzero eigenvalue of the following generalized eigen-decomposition problem:

$$\mathbf{RR}^T\mathbf{p} = \sigma\mathbf{XX}^T\mathbf{p}. \tag{6}$$

2) Obtain $\{\mathbf{p}_t\}_{t=2}^k$: Given $\mathbf{p}_1, \ldots, \mathbf{p}_{t-1}$, the optimization problem regarding the basis $\mathbf{p}_t$ is in the following form:

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2}\|\mathbf{p}^T\mathbf{X} - \mathbf{p}^T\mathbf{XZ}\|_F^2 \\ \text{s.t.,} \quad & \mathbf{p}^T\mathbf{XX}^T\mathbf{p} = 1, \\ & \mathbf{p}^T\mathbf{p}_1 = \mathbf{p}^T\mathbf{p}_2 = \ldots = \mathbf{p}^T\mathbf{p}_{t-1} = 0. \end{aligned} \tag{7}$$

where $t = 2, 3, \ldots, k$.

To solve (7), we introduce the Lagrange multipliers $\beta, \beta_1, \ldots, \beta_{t-1}$ and σ to the optimization problem and obtain the following Lagrange:

$$\mathcal{L} = \frac{1}{2}\|\mathbf{p}^T\mathbf{X} - \mathbf{p}^T\mathbf{XZ}\|_F^2 - \frac{\sigma}{2}(\mathbf{p}^T\mathbf{XX}^T\mathbf{p} - 1) - \sum_{i=1}^{t-1}\beta_i\mathbf{p}^T\mathbf{p}_i.$$

By setting the partial derivatives of $\mathcal{L}$ with respect to $\mathbf{p}$ to zeros, we further obtain the following Karush-Kuhn-Tucker condition,

$$\frac{\partial\mathcal{L}}{\partial\mathbf{p}} = \mathbf{RR}^T\mathbf{p} - \sigma\mathbf{XX}^T\mathbf{p} - \sum_{i=1}^{t-1}\beta_i\mathbf{p}_i = \mathbf{0}. \tag{8}$$

Besides, by multiplying $\mathbf{p}^T$ to the left side of (8) and using the orthogonal constraints, we can have

$$\mathbf{p}^T\mathbf{RR}^T\mathbf{p} = \sigma\mathbf{p}^T\mathbf{XX}^T\mathbf{p}, \tag{9}$$

which leads to

$$\sigma = \frac{\mathbf{pRR}^T\mathbf{p}}{\mathbf{pXX}^T\mathbf{p}}. \tag{10}$$

By left-multiplying $\mathbf{p}_1^T(\mathbf{XX}^T)^\dagger, \ldots, \mathbf{p}_{t-1}^T(\mathbf{XX}^T)^\dagger$ sequentially to both sides of (8), we can obtain the following $(t-1)$ equations:

$$\begin{aligned} \mathbf{p}_1^T(\mathbf{XX}^T)^\dagger\mathbf{RR}^T\mathbf{p} &= \mathbf{p}_1^T(\mathbf{XX}^T)^\dagger\sum_{i=1}^{t-1}\beta_i\mathbf{p}_i, \\ \mathbf{p}_2^T(\mathbf{XX}^T)^\dagger\mathbf{RR}^T\mathbf{p} &= \mathbf{p}_2^T(\mathbf{XX}^T)^\dagger\sum_{i=1}^{t-1}\beta_i\mathbf{p}_i, \\ &\vdots \\ \mathbf{p}_{t-1}^T(\mathbf{XX}^T)^\dagger\mathbf{RR}^T\mathbf{p} &= \mathbf{p}_{t-1}^T(\mathbf{XX}^T)^\dagger\sum_{i=1}^{t-1}\beta_i\mathbf{p}_i. \end{aligned}$$

The above $(t-1)$ equations are the linear equations for the $(t-1)$ Lagrangian multipliers $\beta_1, \ldots, \beta_{t-1}$. To obtain the solution for these Lagrangian multipliers, we define the following notations for ease of presentation:

$$\begin{aligned} &\mathbf{Q} = [\mathbf{Q}_{ij}] = [\mathbf{p}_i^T(\mathbf{XX}^T)^\dagger\mathbf{p}_j] \in \mathbb{R}^{(t-1)\times(t-1)}, \\ &\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{t-1}]^T \in \mathbb{R}^{(t-1)}, \\ &\mathbf{P}_{(t-1)} = [\mathbf{p}_1, \ldots, \mathbf{p}_{(t-1)}] \in \mathbb{R}^{d\times(t-1)}. \end{aligned} \tag{11}$$

With these definitions, we can rewrite the $(t-1)$ equations into the following matrix form:

$$\mathbf{P}_{(t-1)}^T(\mathbf{XX}^T)^\dagger\mathbf{RR}^T\mathbf{p} = \mathbf{Q}\boldsymbol{\beta}. \tag{12}$$

Therefore, we can obtain the optimal solution to $\boldsymbol{\beta}$ in the following closed-form:

$$\boldsymbol{\beta} = \mathbf{Q}^\dagger\mathbf{P}_{(t-1)}^T(\mathbf{XX}^T)^\dagger\mathbf{RR}^T\mathbf{p}. \tag{13}$$

By left-multiplying $(\mathbf{X}\mathbf{X}^T)^\dagger$ to both sides of (8), we obtain

$$(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p} - (\mathbf{X}\mathbf{X}^T)^\dagger\sum_{i=1}^{t-1}\beta_i\mathbf{p}_i = \sigma\mathbf{p},$$

and together with (13), we can further get:

$$\begin{aligned}&(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p} - (\mathbf{X}\mathbf{X}^T)^\dagger\sum_{i=1}^{t-1}\beta_i\mathbf{p}_i\\ =\ &(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p} - (\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{P}_{(t-1)}\boldsymbol{\beta}\\ =\ &(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p} -\\ &(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{P}_{(t-1)}\mathbf{Q}^\dagger\mathbf{P}_{(t-1)}^T(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p}.\end{aligned}$$

Thus, we have

$$\left(\mathbf{I}-(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{P}_{(t-1)}\mathbf{Q}^\dagger\mathbf{P}_{(t-1)}^T\right)(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T\mathbf{p}=\sigma\mathbf{p}, \quad (14)$$

which is a generalized eigen-decomposition problem.

As our objective is to minimize σ with the optimal $\mathbf{p}$, the optimal solution to (7) is given by the eigenvectors corresponding to smallest nonzero eigenvalues of the following matrix:

$$\left(\mathbf{I} - (\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{P}_{(t-1)}\mathbf{Q}^\dagger\mathbf{P}_{(t-1)}^T\right)(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T. \quad (15)$$

C. Automatic feature dimension estimation

In this section, we show that OPCE can automatically estimate the feature dimension with the following theorem.

Theorem 1. *For a given data set* $\mathbf{X}$*, the feature dimension* m' *is upper bounded by the rank of* $\mathbf{Z}^*$*, i.e.,*

$$m' \leq k. \quad (16)$$

Proof. It is easy to see that the optimal solution to (14) is also given by the leading eigenvectors of the following matrix:

$$(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{P}_{(t-1)}\mathbf{Q}^\dagger\mathbf{P}_{(t-1)}^T(\mathbf{X}\mathbf{X}^T)^\dagger\mathbf{R}\mathbf{R}^T. \quad (17)$$

Clearly, the rank of the above matrix (denoted by m') is upper bounded by

$$m' = \min(rank(\mathbf{X}), k). \quad (18)$$

From Lemma 1, ones could see that $\mathbf{Z}^* = \mathbf{V}_k\mathbf{V}_k^T$ and $\mathbf{V}_k$ consists of the first k right singular vectors of $\mathbf{X}$. Thus, we have $rank(\mathbf{X}) = s > k$ and the result as desired. □

D. Algorithm description and complexity analysis

The whole optimization procedure for the OPCE is summarized in Algorithm 1. After obtaining the projection $\mathbf{P}$, we can get the low-dimensional representation for any given data point $\mathbf{x}$ as $\mathbf{P}^T\mathbf{x}$.

Algorithm 1 is composed of obtaining $\mathbf{Z}$ and the projection matrix $\mathbf{P}$, so the corresponding time complexity depends on the complexity of these two parts. For the optimization process *w.r.t.* $\mathbf{Z}$, given the training data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, OPCE usually conducts the skinny SVD in $O(d^2n+dn^2+n^3)$. Using Brand's method [23], the complexity of the skinny SVD can be reduced to $O(dnk)$ with k being the rank of $\mathbf{X}$. Moreover, OPCE determines the feature dimension k with $O(s\log s)$ complexity. On the other hand, to calculate the projection matrix $\mathbf{P}$, the complexity of calculating each $\mathbf{p}_t$ is $O(dn+dn^2)$ as it is a generalized eigen-decomposition problem. Therefore, the complexity of obtaining $\mathbf{P}$ is $O(d(n+n^2)k)$. Considering that $k < \min(d, n)$, the whole time complexity for OPCE is $O(d(n+n^2)k)$.

Algorithm 1 The algorithm for our proposed Orthogonal Principal Coefficients Embedding (OPCE)

Input: n training samples $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ and the regularization parameter $\lambda > 0$.

1: Obtain $\mathbf{Z}$ using Lemma 1.
2: Obtain the solution for $\mathbf{p}_1$ via (6).
3: Obtain $\{\mathbf{p}_t\}_{t=2}^k$ sequentially using (15).

Output: The projection matrix $\mathbf{P}$. (For any new sample $\mathbf{x} \in \mathbb{R}^d$, the low-dimensional representation is given by $\mathbf{P}^T\mathbf{x}$.)

V. Experiments

In this section, to evaluate the performance of our proposed OPCE in comparison with six state-of-the-art subspace learning methods, we perform experiments on five real-world data sets. The used data sets cover different sources, *i.e.* text corpus, handwritten digital images, object images, and the facial images captured under controlled environment.

A. Experimental settings

We implement our method in MATLAB[1] and carry out experiments on a MacBook with a 2.6GHz Intel Core i5 CPU and 8GB memory. To examine the efficacy of our method, we investigate the performance of the proposed method in the context of classification. More specifically, we split each data set into two partitions for training and testing. The training data is used to learn the projection matrix for dimension reduction and train a classifier for classification. In other words, the label information is only used in training of classification phase. With the learned projection matrix and classifier, we obtain the classification accuracy on the testing data. In the experiments, we use four different classifiers including sparse representation based classifier (SRC) [24], linear regression classifier (LRC) [25], linear support vector machine (SVM) [26], and the k nearest neighbor classifier (KNN).

For fair comparisons, we seek optimal parameters for each algorithm to achieve their best performance on each data set by following the settings in [18]. Morever, we repeat each algorithm 10 times and report their mean and standard deviation of classification accuracy and time cost.

1) Baseline Algorithms: To show the effectiveness of our method, we compare the proposed OPCE with PCE. Moreover, we also use six state-of-the-art unsupervised subspace learning methods as baselines, including LPP [6], [27], NPE [5], L1Graph [9], nonnegative matrix decomposition (NMF) [28], [29], robust principal component analysis (RPCA) with PCA, and RPCA with Gradient Descent (RPCAG) [30]. In all

[1]The MATLAB codes of tested algorithms are provided by the authors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2686983, IEEE Transactions on Cognitive and Developmental Systems

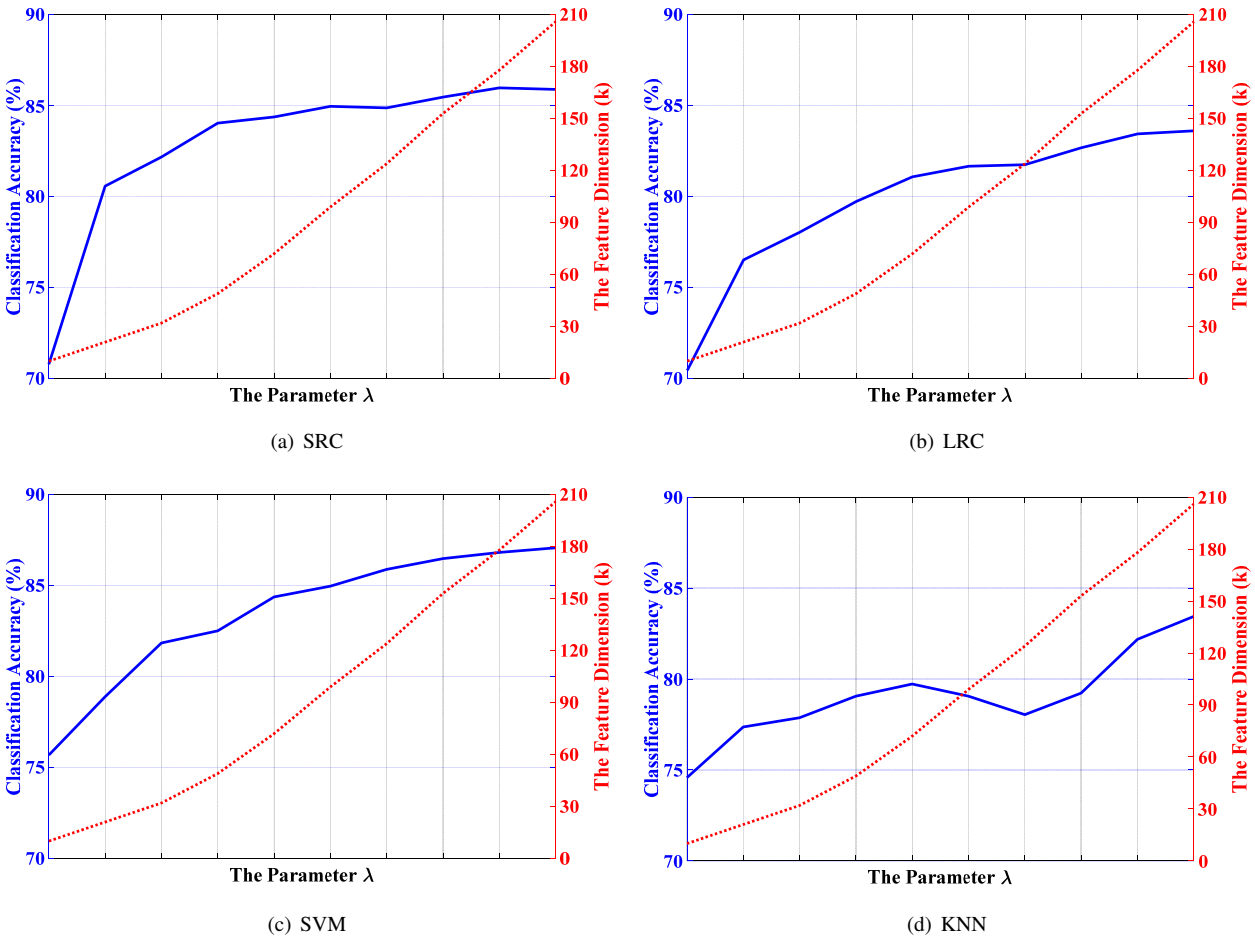


Fig. 1. The influence of the parameter λ. The solid and dotted lines denote the classification accuracy and the estimated feature dimension m' (*i.e.*, k), respectively.

evaluations, we specify the feature dimension as 300 for all tested methods excepted OPCE and PCE since these two methods can automatically estimate the dimension of feature space.

2) Data sets: We use five different data sets in our experiments, including AR facial images [31], COIL100 object images [32], USPS handwritten digital database, Extended Yale Database B (ExYaleB) [33] and Reuters21578 text corpus [34].

The used AR database [35] contains 1,400 clean images, 600 face images with sunglasses, and 600 faces with scarves that evenly distributed over 50 male and 50 female. For computational efficiency, we resize all samples from 165×120 to 55×40. The used COIL100 consists of 1,000 randomly chosen samples that are drawn from 10 different objects. Each image is resized from 128×128 to 64×64. The USPS contains 11,000 digital images, from zero through nine. Each image is with the size of 16×16. The used ExYaleB contains 2024 images that evenly distribute over 38 subjects, of which each image is resized from 192×168 to 54×48. The Reuters21578 corpus includes 21578 documents from 135 categories. In our experiments, we use a subset which contains 2,347 documents in 54 subjects. Each document is represented by a 18,993 dimensional vector.

B. Influence of parameters

We first investigate the influence of parameter of OPCE, *i.e.* the regularization parameter λ. To this end, we perform experiment using the Reuters21578 data set, where 1,160 documents are randomly chosen for training and the remaining 1,187 samples are used for testing.

From Fig. 1, we can see that the estimated feature dimension m' (*i.e.* k) increases from 10 to 206 when λ increases from 0.1 to 1.0. Meanwhile, the classification accuracy increases from 70.78% to 85.90% with SRC, 70.44% to 83.61% with LRC, 75.68% to 87.08% with SVM, and 74.58% to 83.45% with KNN. The classification accuracy increases with λ within a certain range because larger λ leads to larger k (see red curve) so that more energy is preserved.

C. Performance with varying training samples

In this section, we report the performance of OPCE with increasing training samples on the AR data set. We randomly select p clean faces from each category for training and use the remaining $14 - p$ samples for testing, where p increases from 3 to 10 with an interval of 1.

Table II shows the classification accuracy achieved by different feature extraction methods with the LRC classifier,

TABLE II
PERFORMANCE VERSUS INCREASING TRAINING SAMPLES ON THE **AR** DATA SET, WHERE p DENOTES TRAINING SAMPLES PER SUBJECT. RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, ACCORDING TO THE T-TEST WITH A SIGNIFICANCE LEVEL AT 0.05.

p	OPCE	PCE	LPP	NPE	L1Graph	NMF	RPCA	RPCAG
3	79.13±1.08	**79.65±1.39**	75.63±1.98	72.14±2.62	71.45±2.73	25.32±1.83	79.48±2.40	56.88±0.86
4	**86.89±1.16**	84.35±1.04	76.86±2.03	80.79±1.48	80.96±1.91	23.50±1.56	83.56±1.90	64.80±1.83
5	**90.70±1.31**	88.49±1.54	80.22±1.20	85.99±1.54	84.76±0.92	36.78±1.50	86.36±0.74	71.64±1.10
6	**94.57±0.94**	92.84±1.25	84.39±1.65	88.40±2.05	85.38±1.08	57.15±0.91	91.60±0.90	76.95±2.48
7	**96.80±1.07**	94.43±0.84	85.11±2.02	89.10±1.28	84.10±1.36	58.82±1.36	92.79±0.73	76.76±1.83
8	**96.57±0.88**	95.58±1.60	87.89±1.78	89.22±1.53	84.96±1.18	62.56±1.16	95.00±0.62	80.17±1.42
9	**98.03±0.73**	96.36±0.94	91.87±2.07	88.76±2.16	86.10±1.78	62.39±1.04	97.12±0.78	83.74±2.21
10	**99.25±0.91**	96.62±1.23	90.25±1.96	88.33±1.48	85.67±1.28	52.00±0.82	97.43±0.97	84.11±0.83

TABLE III
ROBUSTNESS OF DIFFERENT SUBSPACE LEARNING METHODS WITH THE LRC CLASSIFIER ON THE **DISGUISED AR** IMAGES. RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, ACCORDING TO THE T-TEST WITH A SIGNIFICANCE LEVEL AT 0.05.

Methods	sunglasses		scarves	
	Accuracy	Time	Accuracy	Time
OPCE	**91.07±1.37**	2.48±0.03	**91.35±1.25**	3.54±0.01
PCE	86.78±1.21	1.25±0.02	87.08±1.97	1.26±0.01
LPP	67.61±3.08	2.99±0.48	65.95±2.31	2.56±0.54
NPE	81.73±1.43	5.35±0.04	78.72±1.73	4.89±0.03
L1graph	63.85±2.49	307.57±14.25	58.00±1.57	234.42±33.75
NMF	78.28±2.96	23.52±1.10	76.82±2.52	21.70±0.65
RPCA	87.43±2.04	978.95±79.97	89.57±0.98	193.44±4.11
RPCAG	53.23±2.72	38.64±1.00	40.73±2.22	38.96±1.15

TABLE IV
ROBUSTNESS OF DIFFERENT SUBSPACE LEARNING METHODS WITH THE SRC CLASSIFIER ON THE **CORRUPTED EXYALEB** IMAGES. RESULTS IN BOLDFACE ARE SIGNIFICANTLY BETTER THAN THE OTHERS, ACCORDING TO THE T-TEST WITH A SIGNIFICANCE LEVEL AT 0.05.

Methods	Accuracy	Time (seconds)
OPCE	**82.67±0.76**	31.91±3.21
PCE	80.89±1.06	22.84±0.91
LPP	44.56±1.82	50.23±5.38
NPE	62.30±2.12	34.57±0.57
L1graph	39.13±3.24	548.26±15.95
NMF	65.34±1.66	208.84±8.05
RPCA	80.85±0.61	415.13±7.89
RPCAG	77.40±1.28	203.39±9.88

from which we can see that: (1) OPCE consistently achieves the highest accuracy and PCE usually achieves the second best performance; (2) OPCE is more competitive than the baselines when more training samples are available. For example, with the increasing p, the performance gain of OPCE over NPE increases from 6.99% to 9.27%.

D. Performance comparison on clean data

In this section, we compare our method with several state-of-the-art subspace learning approaches on the COIL100, USPS, and Reuters21578 data sets. For COIL100 and USPS, we randomly select half of samples for training and use the rest for testing. For the Reuters21578 data set, we use 1,740 documents for training and 607 documents for testing.

Figures 2–4 show the results from which we can see:

- In most cases, our OPCE achieves the best performance on these three data sets. On COIL100, the best performance is achieved by OPCE with the SRC and KNN classifiers, *i.e.* 64.40%, which is at least 4.80% higher than the other methods in all settings.
- On the USPS data set, the highest accuracy rate is also achieved by OPCE with SRC, which is about 97.65% and is 1.05% higher than the second best method. Note that, OPCE and PCE achieve similar recognition rate when LRC is used as classifier.
- On the Reuters21578 data set, OPCE outperforms all the evaluated algorithms by a considerable margin. The highest accuracy of OPCE is achieved by SVM, which is 1.70% higher than the best result of baseline approaches.

E. Robustness to real disguises and random pixel corruption

In this section, we first investigate the robustness of our method with LRC using two subsets of the AR data set. The first subset contains 600 clean images and 600 faces disguised by sunglasses. The second subset contains 600 clean images and 600 faces disguised by scarves. Moreover, we compare the computational efficiency of these methods.

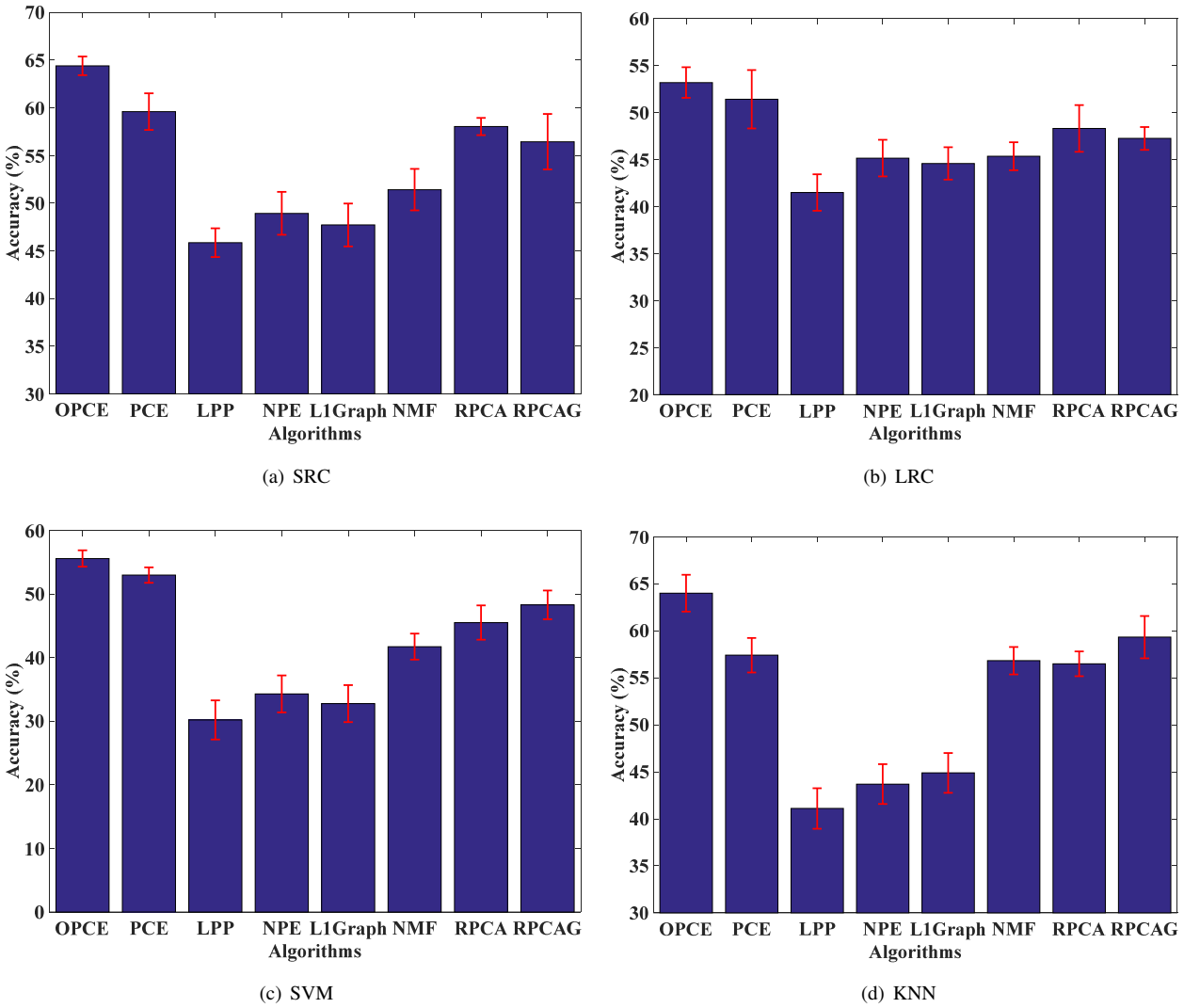
Table III shows that our method is remarkably superior to the other approaches. When the faces are disguised by sunglasses (occlusion rate is about 20%), OPCE exceeds the best baseline method of 3.64% and PCE of 4.29% in accuracy. In the case of scarves disguise, the corresponding gains are 1.78% and 4.27%. OPCE not only achieves the highest accuracy, but also is the one of the fastest algorithms. It is over 435 and 54 times faster than RPCA *w.r.t.* the sunglasses and the scarves disguises.

To examine the robustness of our method against missing entries, we conduct experiments on the ExYaleB data set with random pixel corruption. Specifically, we randomly select a half of images to remove some entries by replacing the value of 30% pixels with 0 or $p_{\max}$, where $p_{\max}$ is the largest pixel value of the current image. The results are summarized in Table IV which show that OPCE is significantly prior to the other tested methods in terms of classification accuracy and efficiency. Our method is 1.78% higher than PCE, which shows the effectiveness of our orthogonal constraint.

VI. CONCLUSION

In this paper, we proposed an unsupervised subspace learning algorithm, termed orthogonal principal coefficients embedding (OPCE). With a novel objective function, OPCE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2686983, IEEE Transactions on Cognitive and Developmental Systems



(a) SRC (b) LRC

(c) SVM (d) KNN

Fig. 2. Performance comparison on the **COIL100** data set.

enforces the projection matrix to be mutually orthogonal in column space, thus resulting in more discriminative features from raw data. Moreover, OPCE automatically determines this parameter based on the data distribution, without requiring the dimension of feature space to be specified in advance. Extensive experimental results show the effectiveness and efficiency of our proposed method on face, object, handwritten digit image, and text corpus classification in comparison with several baseline methods.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that significantly improve the quality of this paper.

REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991. 1

[2] 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, 1997. 1

[3] W. Zuo, D. Zhang, J. Yang, and K. Wang, "Bdpca plus lda: a novel fast feature extraction technique for face recognition," *IEEE Trans Syst. Man Cybern. B Cybern.*, vol. 36, no. 4, pp. 946–953, Aug. 2006. 1

[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000. 1, 2

[5] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. of 10th IEEE Conf. Comput. Vis.*, Beijing, China, Oct. 2005, pp. 1208–1213. 1, 2, 4

[6] 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, 2005. 1, 4

[7] D. Cai, X. He, J. Han, and H. J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006. 1, 3

[8] L. S. Qiao, S. C. Chen, and X. Y. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010. 1, 2

[9] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, "Learning with L1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, 2010. 1, 2, 4

[10] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, p. 11, 2011. 1, 2

[11] X. Peng, Z. Yi, and H. Tang, "Robust subspace clustering via thresholding ridge regression," in *Proc. of 29th AAAI Conf. Artif. Intell.*, Jan. 2015. 1, 2

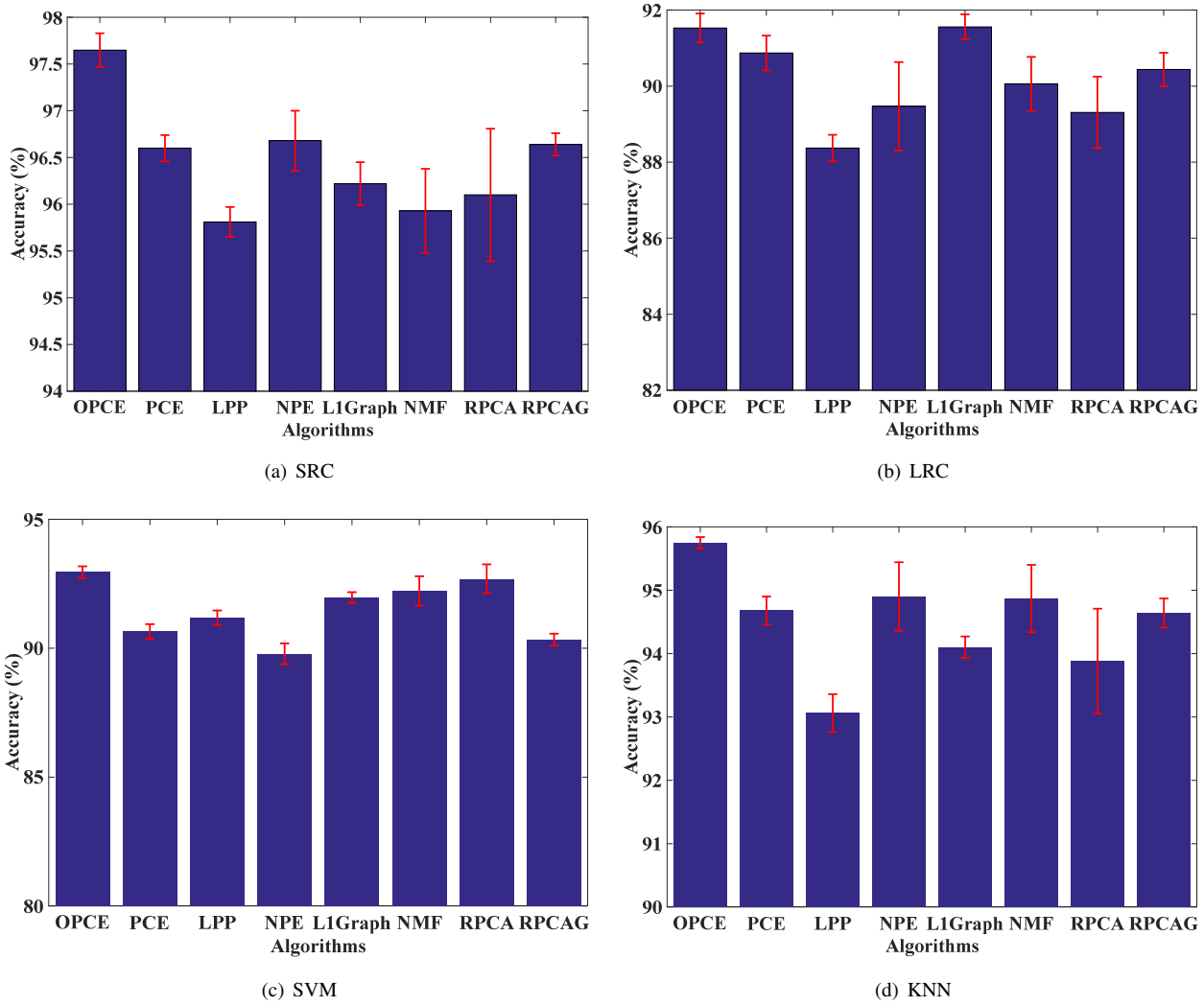


Fig. 3. Performance comparison on the **USPS** data set.

[12] X. Peng, C. Lu, Y. Zhang, and H. Tang, "Connections between nuclear norm and frobenius norm based representation," *IEEE Trans Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–7, 2016. 1, 2

[13] X. Peng, M. Yuan, Z. Yu, W.-Y. Yau, and L. Zhang, "Semi-supervised subspace learning with l2graph," *Neurocomputing*, vol. PP, no. 99, 2016. 1

[14] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016. 1

[15] W. Zuo, D. Ren, D. Zhang, S. Gu, and L. Zhang, "Learning iteration-wise generalized shrinkage-thresholding operators for blind deconvolution," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1751–1764, Apr. 2016. 1

[16] X. Peng, H. Tang, L. Zhang, Z. Yi, and S. Xiao, "A unified framework for representation-based subspace clustering of out-of-sample and large-scale data," *IEEE Trans Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2499–2512, 2016. 1

[17] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *Proc. of 13th Eur. Conf. Comput. Vis.*, 2014, pp. 123–138. 1

[18] X. Peng, J. Lu, Z. Yi, and Y. Rui, "Automatic subspace learning via principal coefficients embedding," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–14, 2016. 1, 2, 4

[19] P. M. Knutsen and E. Ahissar, "Orthogonal coding of object location." *Trends Neurosci.*, vol. 32, no. 2, pp. 101–109, Feb. 2009. 1, 3

[20] Q. Yu, R. Yan, H. Tang, K. C. Tan, and H. Li, "A spiking neural network system for robust sequence recognition," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 27, no. 3, pp. 621–635, Mar. 2016. 1

[21] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013. 2

[22] X. Peng, C. Lu, Z. Yi, and H. Tang, "Connections between nuclear norm and frobenius norm based representation," *arXiv:1502.07423*, 2015. 2

[23] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra Appl.*, vol. 415, no. 1, pp. 20–30, 2006. 4

[24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009. 4

[25] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010. 4

[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008. 4

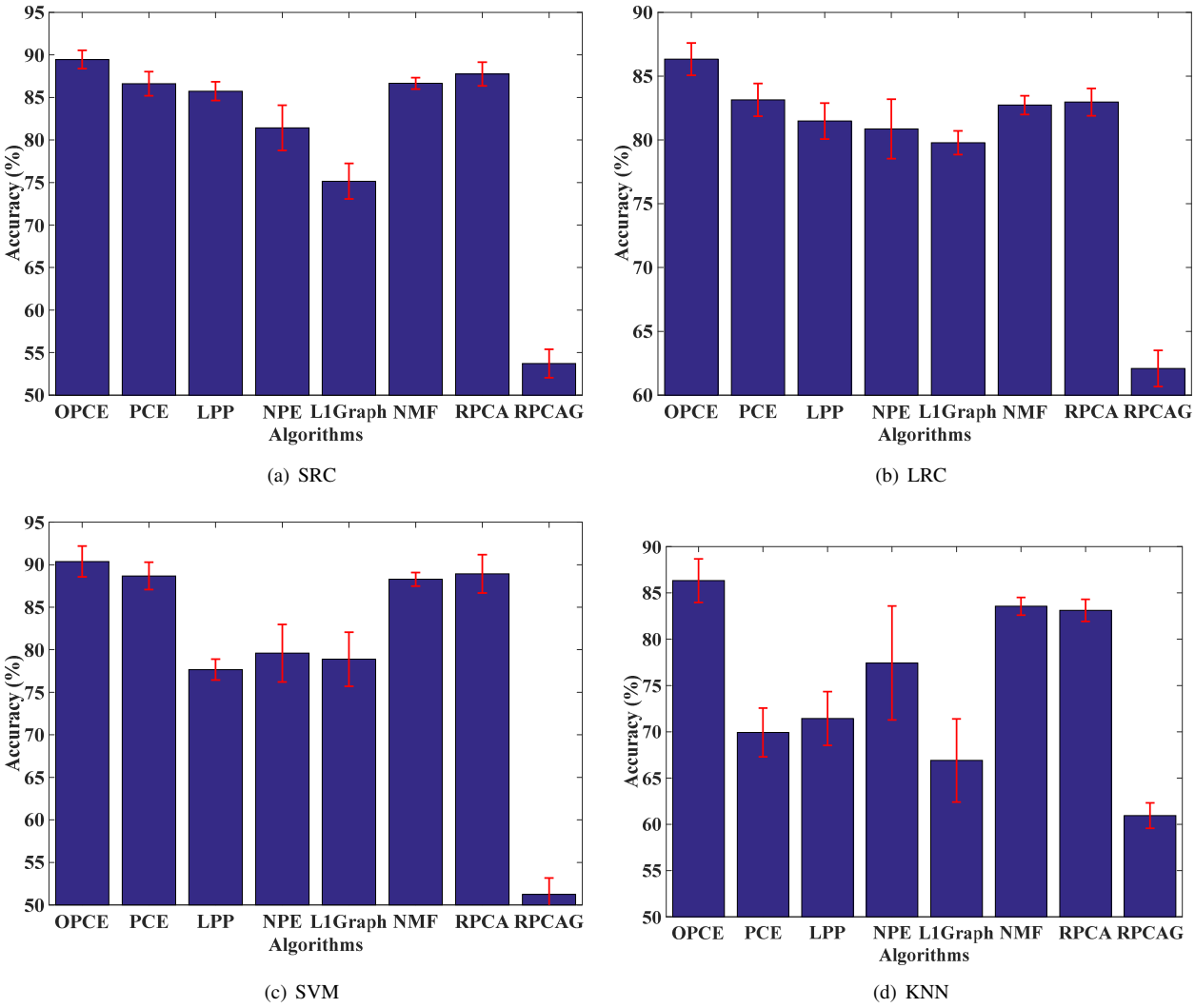
[27] X. He and P. Niyogi, "Locality preserving projections," in *Proc. of 17th Adv. in Neural Inf. Process. Syst.*, Vancouver, British Columbia, Canada, Dec. 2004, pp. 153–160. 4

[28] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of Adv. in Neural Inf. Process. Syst.*, 2001, pp. 556–562. 4

[29] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004. 4

[30] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2017.2686983, IEEE Transactions on Cognitive and Developmental Systems



(a) SRC (b) LRC

(c) SVM (d) KNN

Fig. 4. Performance comparison on the **Reuters21578** data set.

pca via gradient descent," in *Proc. of Adv. in Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 4152–4160. 4

[31] A. Martinez and R. Benavente, "The AR face database," 1998. 5

[32] S. Nayar, S. A. Nene, and H. Murase, "Columbia object image library (COIL 100)," Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96, Tech. Rep., Feb. 1996. 5

[33] A. S. Georghiades, P. N. Belhumeur, and D. J. 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, 2001. 5

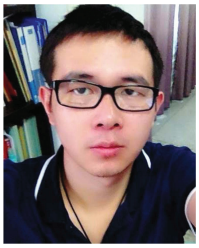
[34] D. Cai, X. F. He, and J. W. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, 2005. 5

[35] X. Peng, L. Zhang, Z. Yi, and K. K. Tan, "Learning locality-constrained collaborative representation for robust face recognition," *Pattern Recognit.*, vol. 47, no. 9, pp. 2794–2806, 2014. 5

Xinxing Xu received the BE degree from the University of Science and Technology of China, Hefei, China, in 2009. He received the PhD degree in computer engineering from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2015. He is currently a scientist with the Institute of High Performance Computing (IHPC), the Agency for Science, Technology and Research, Singapore. His current research interests include artificial intelligence, machine learning and their applications to computer vision. He has published a few research papers in refereed international journals and conference proceedings. He received the 2016 Best Paper Award in the first International Workshop on *BeyondLabeler - Human is More Than a Labeler* at the International Joint Conference on Artificial Intelligence (IJCAI).

Shijie Xiao received the B.E. degree from the Harbin Institute of Technology, Harbin, China, in 2011, and the Ph.D. degree from the Nanyang Technological University, Singapore, in 2016. Since 2015 he is a senior algorithm development engineer at OmniVision Technologies Singapore Pte. Ltd. His current research interests include machine learning and computer vision.

Zhang Yi (SM'10–F'15) received the Ph.D. degree in mathematics from the Institute of Mathematics, The Chinese Academy of Science, Beijing, China, in 1994. Currently, he is a Professor at the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu, China. He is the co-author of three books: Convergence Analysis of Recurrent Neural Networks (Kluwer Academic Publishers, 2004), Neural Networks: Computational Models and Applications (Springer, 2007), and Subspace Learning of Neural Networks (CRC Press, 2010). He was an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (2009–2012) and is an Associate Editor of IEEE Transactions on Cybernetics (2014). His current research interests include Neural Networks and Big Data. He is a fellow of IEEE.



Xi Peng received the BEng degree in Electronic Engineering and MEng degree in Computer Science from Chongqing University of Posts and Telecommunications, and the Ph.D. degree from Sichuan University, China, respectively. His current research interests include machine intelligence and computer vision and has authored more than 20 articles in these areas.

Dr. Peng has served as a Guest Editor for *IEEE Trans. on Neural Network and Learning Systems*, and *Image and Vision Computing*, a Session Chair for *AAAI Conference on Artificial Intelligence* 2017, a Senior Program Committee Member, Program Committee Member and reviewer for over 20 international conferences and international journals. He has given a tutorial at *European Conference on Computer Vision* (ECCV'16).

Yong Liu is working as Capability Group Manager for Artificial Intelligence Group and Scientist in Institute of High Performance Computing (IHPC) at A*STAR, Singapore. He has worked as Principle Investigator to lead several projects on artificial intelligence and machine learning. Dr. Liu Yong holds a PhD degree from National University of Singapore. After PhD study, he has worked as Post-Doc researcher at Royal Swedish Academic of Science. He has received multiple awards and research grants from Singapore government agencies such as EDB and SPRING Singapore. His research areas include artificial intelligence, large scale machine learning, recommender system, cloud computing and computer networks. He is also the co-author of two books.