Dual Pursuit for Subspace Learning

Shuangyan Yi, *Member, IEEE*, Yingyi Liang, Zhenyu He, *Senior Member, IEEE*, Yi Li, and Yiu-Ming Cheung, *Fellow, IEEE*

***Abstract*—In general, low-rank representation (LRR) aims to find the lowest rank representation with respect to a dictionary. In fact, the dictionary is a key aspect of low-rank representation. However, a lot of low-rank representation methods usually use the data itself as a dictionary (i.e., a fixed dictionary), which may degrade their performances due to the lack of clustering ability of a fixed dictionary. To this end, we propose learning a locality-preserving dictionary instead of the fixed dictionary for low-rank representation, where the locality-preserving dictionary is constructed by using a graph regularization technique to capture the intrinsic geometric structure of the dictionary and, hence, the locality-preserving dictionary has an underlying clustering ability. In this way, the obtained low-rank representation via the locality-preserving dictionary has a better grouping-effect representation. Inversely, a better grouping-effect representation can help to learn a good dictionary. The locality-preserving dictionary and the grouping-effect representation interact with each other, where dual pursuit is called. The proposed method, namely, Dual Pursuit for Subspace Learning, provides us with a robust method for clustering and classification simultaneously, and compares favorably with the other state-of-the-art methods.**

***Index Terms*—Low-rank representation, dual pursuit, graph-regularization technique.**

I. Introduction

LOW-RANK Representation (LRR) [1]–[3], as a promising subspace clustering method [4]–[6], aims to capture the underlying data structure from a global perspective, which has been reported to be superior to similar methods [7]. Due to the effectiveness of LRR, various methods based on it [8]–[12] have been proposed and widely used in motion segmentation [13]–[15], face recognition [16]–[18], visual tracking [19], [20], saliency detection [21], [22], and recommendation system [23], [24]. Since multimedia data includes texts, images, and videos, etc. Recently, many machine learning algorithms including low-rank methods have appeared in multimedia retrieval, such as image retrieval [25]–[27] and image classification [28]–[30] to improve the multimedia retrieval performance.

LRR exploits the self-expressive ability of the data itself via low-rank and finds the underlying low-rank structure. Although LRR provides us with an efficient way to automatically correct the corruptions lying in the original data, it only considers the global structure. Therefore, it is a natural idea to take the local manifold structure into consideration [2], [20], [31]. Lu *et al.* [2] proposed to incorporate a graph Laplacian into LRR and to enforce the desired low-rank subspace structures. Liu *et al.* [31] introduced a manifold regularization into LRR and formed a non-negative low-rank representation method.

All of the aforementioned methods use the observation data itself as a dictionary (i.e., a fixed dictionary), however, such a strategy may degrade their performances, especially when the intrinsic geometric structures of observations are hidden in observations. To this end, the idea of constructing a novel dictionary for low-rank representation has appeared in the literature [9], [32]. For example, latent low-rank representation (LatLRR) [9] is proposed to construct a dictionary by using both observed and unobserved data. Similar to LatLRR, the dual low-rank method in [33] is proposed to learn a set of low-rank features as a dictionary for detecting the low-rank salient regions. Moreover, a novel low-rank version is constructed in the transformed data space [34], [35], which uses the transformed data as a dictionary to recover the transformed data itself. In contrast, we argue that learning a locality-preserving dictionary is necessary for low-rank representation, because the locality-preserving dictionary has an underlying clustering ability and this will help learn a good grouping-effect representation. In this paper, the locality-preserving dictionary is therefore constructed by using a graph regularization technique, which is able to capture the intrinsic geometry structure of data and implicitly has a clustering ability [36]. Based on such a locality-preserving dictionary, the grouping-effect representation is likely to be in the dense diagonal blocks due to the stronger discriminative features extracted by locality-preserving dictionary. Here, the proposed method, namely dual pursuit, interacts locality-preserving dictionary and grouping-effect representation each other to simultaneously obtain the most optimal dictionary and representation.

Manuscript received March 21, 2018; revised June 27, 2018 and August 22, 2018; accepted September 28, 2018. Date of publication October 24, 2018; date of current version May 22, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61672183, Grant 61272366, Grant 61672444, Grant 61772141, in part by the Natural Science Foundation of Guangdong Province under Grant 2016B090918047, in part by the Shenzhen Research Council under Grant JCYJ20170413104556946, Grant JCYJ20160406161948211, Grant JCYJ20160226201453085, Grant JSGG20150331152017052, Grant JCYJ20160531194006833, in part by the Shenzhen science and technology plan under Grant KJYY20170724152625446, in part by the Scientific Research Platform Cultivation Project of SZIIT (PT201704), and in part by the Faculty Research Grant of Hong Kong Baptist University under Grant FRG2/17-18/082. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dong Xu. *(Shuangyan Yi and Yingyi Liang contributed equally to this work.) (Corresponding author: Zhenyu He.)*

S. Yi, Y. Liang, Z. He, and Y. Li are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: shuangyanshuangfei@163.com; liangyingyi002@foxmail.com; zhenyuhe@hit.edu.cn; ly_res@163.com).

Y.-M. Cheung is with the Department of Computer Science and the Institute of Research and Continuing Education, Hong Kong Baptist University Hong Kong, and also with the United International College, Beijing Normal University-HKBU, Zhuhai 519000, China (e-mail: ymc@comp.hkbu.edu.hk).

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/TMM.2018.2877888

Therefore, unlike the traditional subspace learning methods that usually consider clustering and classification independently, the proposed method can perform both clustering and classification simultaneously.

The main contributions of this paper are summarized as follows:

- We propose to learn a locality-preserving dictionary for low-rank representation instead of a fixed dictionary. The learned locality-preserving dictionary can capture the intrinsic geometry structure of data, and thus the learned dictionary has an underlying clustering ability to some extent.
- We propose a novel constraint term to unify the locality-preserving dictionary and grouping-effect representation. Such a strategy can make the locality-preserving dictionary and grouping-effect representation interact with each other and achieves the most optimal locality-preserving dictionary and grouping-effect representation.

The remainder of this paper is organized as follows. In Section II, low-rank representation and the graph regularization technique are reviewed. In Section III, the proposed method and its optimization algorithm are presented. In Section IV, the differences between our method and the related work are discussed. In Section V, the experiments using the proposed method are designed to demonstrate its effectiveness. Finally, a conclusion is drawn in Section VI.

II. Brief Review of Low-Rank Representation and Graph Regularization Technique

In this section, we briefly review low-rank representation and graph regularization technique.

A. Low-Rank Representation

Given the observed data $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ (n is the number of samples and d is the dimension of features), which is approximately drawn from a mixture of multiple low-rank subspaces, the LRR method [2] uses the observed data itself to find the lowest-rank representation matrix Z of all data jointly as follows:

$$\begin{aligned} \arg\min_{Z,E} \quad & \|Z\|_* + \lambda \|E\|_{2,1}, \\ \text{s.t.} \quad & X = XZ + E, \end{aligned} \tag{1}$$

where $\|Z\|_*$, defined as the sum of all singular values of Z, is the so-called nuclear norm [37], and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-norm to characterize the error E.

B. Graph Regularization Technique

The graph regularization technique is built on a graph and the graph is constructed from the data samples. In the graph, the weight of the edge between data samples x_i and x_j is usually defined as:

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in \aleph_K(x_j) \text{ or } x_j \in \aleph_K(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\aleph_K(x_i)$ indicates the set of K nearest neighbours of x_i. The set of W_{ij} is denoted as W, which is a sparse symmetric $n \times n$ matrix.

The goal of the graph regularization technique is to preserve the locality relation among samples in the low-dimensional subspace. Assuming that Φ_i and Φ_j are any two low-dimensional subspace points produced by the original samples x_i and x_j, where Φ is the low-dimensional data of original data X. The formulation of the graph regularization technique can be generalized as follows:

$$\arg\min_{\Phi} \sum_{ij} \| \Phi_i - \Phi_j \|_2^2 W_{ij}. \tag{3}$$

The graph regularization technique is able to weaken unnecessary connections and strengthen the necessary connections. Generally, $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ and $\sum_{ij} \| z_i - z_j \|_2^2 W_{ij}$ are two common graph regularization terms, where the former imposes the graph regularization technique on the projection matrix P while the latter imposes the graph regularization technique on the representation coefficient matrix Z.

III. Dual Pursuit for Subspace Learning

In this section, the objective function of dual pursuit is first constructed and then its optimization algorithm is given.

A. Objective Function of the Proposed Method

Low-rank representation uses the data itself as a fixed dictionary to recover the original data. However, using the data directly itself as a dictionary may be blind and may degrade the performance of grouping-effect representation. To this end, it is necessary to improve the performance of grouping-effect representation by learning an effective dictionary. In fact, a priori information (i.e., the grouping-effect) can be used to guide the learning of the effective dictionary. Fortunately, the graph regularization term $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ can make the transformed data $P^T X$ have the grouping-effect property. This is because the transformed data, by preserving the neighborhood structure in the transformed space, implicitly emphasizes the data groups that are more correct than the original data groups. More specifically, the graph regularization term $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ can make the neighboring points in the original space nearer in the transformed data space and the faraway points in the original space further in the transformed data space [36]. Taking Fig. 1 as an example, node3 and node1 are usually classified as one class due to their nearest Euclidean distance. In fact, node3 and node1 are not in the same class. However, the graph regularization term $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ will make node4 and node3 nearer while make node4 and node1 farther. That is, in the transformed space, node4 and node3 are classified as one class while node4 and node1 are not classified as one class. Naturally, node3 and node1 are not classified as the same class. Therefore, the learned data transformed by the graph regularization technique, which can correctly capture the intrinsic geometric structure of the data, is called the locality-preserving dictionary. Based on such

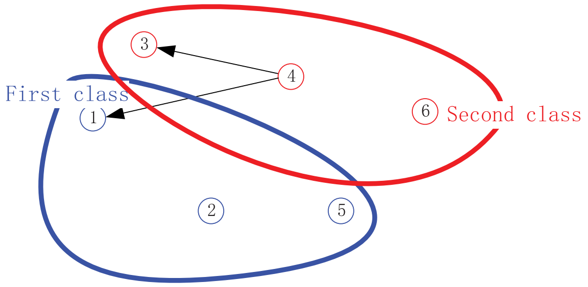

Fig. 1. Illustration of the locality-preserving dictionary. There are a total of six nodes in the original space, where the red curve line is the neighbourhood of node4 and the blue curved line is the neighbourhood of node2. Usually, node3 and node1 would be classified as one class due to their nearest Euclidean distance. In fact, node3 and node1 are not in the same class. However, the graph regularization technique can correctly capture the neighbourhood structures of data.

a locality-preserving dictionary, a more optimal grouping-effect representation coefficient matrix is expected.

For the noiseless data, the proposed dual pursuit is written as follows:

$$\begin{aligned} \arg\min_{Z,P} \quad & \|Z\|_* + \frac{\lambda}{2}\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}, \\ \text{s.t.} \quad & X = P^T X Z. \end{aligned} \tag{4}$$

Furthermore, when the data is corrupted by noise (i.e., illumination corruptions or random pixel corruptions), the objective function of dual pursuit is written as follows:

$$\begin{aligned} \arg\min_{Z,P,E} \quad & \|Z\|_* + \frac{\lambda}{2}\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij} + \gamma\|E\|_{2,1}, \\ \text{s.t.} \quad & X = P^T X Z + E, \end{aligned} \tag{5}$$

where $\lambda \geq 0$ and $\gamma \geq 0$ are two balance parameters, Z is the low-rank representation coefficient matrix, $P \in \mathbb{R}^{m\times m}$ is the transformation matrix, and E is an error term. Here, W_{ij} defined in Eq. (2) is used to constrain $P^T x_i$ and $P^T x_j$. The graph regularization term $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ is used to find a transformation matrix P such that, under this transformation matrix, the sum of Euclidean distances between data pairs that are local to each other is minimized.

There are two main distinctive aspects of our method: the graph regularization term (i.e. $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$) and the constraint (i.e. $X = P^T X Z + E$). In terms of the graph regularization term, the graph weight W is used to constrain the transformed data pairs (e.g., $P^T x_i$ and $P^T x_j$) such that they preserve the local geometric structure well. One important reason for adding $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ into our objective function is to let the learned dictionary $P^T X$ have the grouping-effect property. Based on such a locality-preserving dictionary, the grouping-effect representation will bring a better clustering result than that based on a fixed dictionary. On the other hand, through the constraint $X = P^T X Z + E$, the dual pursuit is produced where the optimized variables Z and P interact with each other. Finally, the optimal Z^* can be used for subspace clustering and the optimal P^* can be used for feature extraction.

B. Optimization

Many algorithms have been developed to optimize the low-rank representation methods, such as Singular Value Thresholding (SVT) [38], Augmented Lagrange Multiplier (ALM) [39], Alternating Direction Method (ADM) [40], and Linearized Alternating Direction Method with Adaptive Penalty (LADMAP) [41]. Recently, in order to deal with multi-block variables, the Linearized Alternating Direction Method with Parallel Splitting and Adaptive Penalty (LADMPSAP) [42] has been proposed with a convergence guarantee.

Using some simple algebraic formulations [31], [43]–[45] for $\frac{\lambda}{2}\| P^T x_i - P^T x_j \|_2^2 W_{ij}$, Eq. (5) is converted into the equivalent optimization formulation as follows:

$$\begin{aligned} \arg\min_{Z,P,E} \quad & \|Z\|_* + \lambda tr(P^T X L X^T P) + \gamma\|E\|_{2,1}, \\ \text{s.t.} \quad & X = P^T X Z + E, \end{aligned} \tag{6}$$

where $L = D - W$ is the graph Laplacian matrix [46], D is a diagonal matrix whose entries are column sums of W, and W is the weight matrix with W_{ij}.

The main part of the objective function in Eq. (6) is convex, and the constraint in Eq. (6) is convex only when the variable P or Z is fixed. For efficiency, we utilize the ALM method [39] to solve the optimization problem. One trick in applying ALM is to make the objective function separable. Therefore, we introduce one auxiliary variable A and reformulate Eq. (6) into the following equivalent formulation:

$$\begin{aligned} \arg\min_{Z,A,P,E} \quad & \|A\|_* + \lambda tr(P^T X L X^T P) + \gamma\|E\|_{2,1}, \\ \text{s.t.} \quad & X = P^T X Z + E, Z = A. \end{aligned} \tag{7}$$

Then, using the ALM method, we can minimize the following augmented Lagrange function:

$$\begin{aligned} \arg\min_{Z,A,P,E} \quad & \|A\|_* + \lambda tr(P^T X L X^T P) + \gamma\|E\|_{2,1} \\ & + tr(Y_1^T(X - P^T X Z - E)) + tr(Y_2^T(Z - A)) \\ & + \frac{\mu}{2}\left(\|X - P^T X Z - E\|_F^2 + \|Z - A\|_F^2\right), \end{aligned} \tag{8}$$

where $tr(\cdot)$ and $\|\cdot\|_F$ denote the trace and Frobenious norm of a matrix respectively, and $\mu > 0$ is a penalty parameter. This augmented Lagrange function is unconstrained, and hence it can be minimized with respect to Z, A, P and E respectively by fixing other variables. Finally, we update the Lagrange multipliers Y_1 and Y_2. The main solving process is given as follows.

Step 1: Fix other variables and update Z. Eq. (8) is reduced to the following formulation:

$$\begin{aligned} \arg\min_{Z} \quad & \left\| X - P_k^T X Z_k - E_k + \frac{Y_1^k}{\mu} \right\|_F^2 \\ & + \left\| Z_k - A_k + \frac{Y_2^k}{\mu} \right\|_F^2. \end{aligned} \tag{9}$$

By derivating Eq. (9) with respect to Z be 0, this optimization solution $Z_{k+1} = (\mu I + \mu X^T P_k P_k^T X)^{-1}$

$(X^TP_kY_1^k + \mu X^TP_k(X - E_k) - Y_2^k + \mu A_k)$ is got.

Step 2: Fix other variables and update A. Eq. (8) is reduced to the following formulation:

$$\arg\min_{A} \|A\|_* + \frac{\mu}{2}\left\|A - (Z + \frac{Y_2^k}{\mu})\right\|_F^2. \tag{10}$$

This optimization solution $A_{k+1} = J_{\frac{1}{\mu}}(Z_{k+1} + \frac{Y_2^k}{\mu})$, where J is the thresholding operator with respect to the singular value $\frac{1}{\mu}$, can be obtained via SVT operator [38].

Step 3: Fix other variables and update P. Eq. (8) is reduced to the following formulation:

$$\begin{aligned} \arg\min_{P} \quad & \lambda tr(P^TXLX^TP) \\ & + \frac{\mu}{2}\left\|X - P^TXZ - E_k + \frac{Y_1^k}{\mu}\right\|_F^2. \end{aligned} \tag{11}$$

By derivating Eq. (11) with respect to P be 0, this optimization solution $P_{k+1} = (2\lambda XLX^T + \mu XZ_{k+1}Z_{k+1}^TX^T)^{-1}\mu XZ_{k+1}(X - E_k + Y_1^k/\mu)^T$ is got.

Step 4: Fix other variables and update E. Eq. (8) is reduced to the following formulation:

$$\arg\min_{E} \gamma\|E\|_{2,1} + \frac{\mu}{2}\left\|X - P^TXZ - E + \frac{Y_1^k}{\mu}\right\|_F^2, \tag{12}$$

which has the closed-form solution $E_{k+1} = S_{\frac{\gamma}{\mu}}(X - P^T{}_{k+1}XZ_{k+1} + \frac{Y1^k}{\mu})$ using the shrinkage operator [39], where S is the ℓ_{21} minimization operator [10].

The complete solving process is shown in Algorithm 1. The major computation of Algorithm 1 is at Step 2, which requires computation of the Singular Value Decomposition (SVD) of the matrix $Z_{k+1} + Y_2^k/\mu_k \in \mathbb{R}^{n\times n}$, where k is the number of iterations. Therefore, the complexity of this algorithm is $O(kn^3)$. When n is large, its computational cost is very high. Some references [3], [47] provide a way to reduce the computational cost. Moreover, the most optimized solutions Z_{k+1}, A_{k+1}, P_{k+1}, and E_{k+1} are denoted as Z^*, A^*, P^*, and E^*.

IV. Comparisons to Related Work

A. Comparison to LRR

LRR (See Eq. (1)) uses the observation data X itself as a dictionary to globally represent the original data. When the observation data is embedded in a low-dimensional manifold, LRR may degrade its performance [45]. In contrast, our method (see Eq. (5)) learns the transformed data P^TX as a locality-preserving dictionary to represent the original data. Moreover, LRR only provides a way to cluster the given data but can not directly classify the new coming data samples. For a new datum, LRR needs to recalculate over all the data, and results in high computational cost. Therefore, it is limited in application that require fast online computation. However, our method can simultaneously cluster the given data and classify the new coming data samples via the learned transformation matrix P.

Algorithm 1: Optimization Algorithm

```
Initialize: $Z_0 = A_0 = 0, E_0 = 0, Y_1^0 = 0, Y_2^0 = 0,$
            $P_0 = I$, $\mu_0 = 10^{-3}, \max_\mu = 10^6, \rho = 1.1,$
            $\varepsilon = 10^{-6}$, $k = 0$;
While not converged do
  1: Fix the others and update $Z$ by setting:
     $Z_{k+1} = (\mu_k I + \mu_k X^TP_kP_k^TX)^{-1}$
         $\times (X^TP_kY_1^k + \mu_k X^TP_k(X - E_k) - Y_2^k + \mu_k A_k)$;
  2: Fix the others and update $A$ by setting:
     $A_{k+1} = J_{\frac{1}{\mu_k}}\left(Z_{k+1} + \frac{Y_2^k}{\mu_k}\right)$;
     where $J$ is the thresholding operator [38] with respect to
     the singular value $\frac{1}{\mu_k}$.
  3: Fix the others and update $P$ by setting:
     $P_{k+1} = (2\lambda XLX^T + \mu_k XZ_{k+1}Z_{k+1}^TX^T)^{-1}$
         $\times \mu_k XZ_{k+1}(X - E_k + Y_1^k/\mu_k)^T$;
  4: Fix the others and update $E$ by setting:
     $E_{k+1} = S_{\frac{\gamma}{\mu_k}}\left(X - P_{k+1}^TXZ_{k+1} + \frac{Y1^k}{\mu_k}\right)$;
     where $S$ is the $\ell_{21}$ minimization operator [10].
  5: Update the multipliers:
     $Y_1^{k+1} = Y_1^k + \mu_k(X - P_{k+1}^TXZ_{k+1} - E_{k+1})$
     $Y_2^{k+1} = Y_2^k + \mu_k(Z_{k+1} - A_{k+1})$;
  6: Update the parameter $\mu_{k+1}$: $\mu_{k+1} = \min(\rho\mu_k, \max_\mu)$
  7: Check the convergence conditions:
     $\|X - P_{k+1}^TXZ_{k+1} - E_{k+1}\|_\infty < \varepsilon$ and
     $\|Z_{k+1} - A_{k+1}\|_\infty < \varepsilon$;
  8: Update $k : k \leftarrow k + 1$;
End while
```

B. Comparison to LatLRR

LatLRR proposes the dual low-rank optimization problem as follows:

$$\begin{aligned} \arg\min_{Z,G,E} \quad & \|Z\|_* + \|G\|_* + \lambda\|E\|_1, \\ \text{s.t.} \quad & X = XZ + GX + E, \end{aligned} \tag{13}$$

where $\|\cdot\|_1$ is the ℓ_1-norm to characterize the error E, Z is a low-rank representation coefficient matrix and G is a low-rank transformation matrix.

LatLRR (See Eq. (13)) is proposed to construct a novel dictionary by using the observed and unobserved data, where the effect of unobserved data is reflected by a low-rank transformation matrix G. From Eq. (13), the data recovered by LatLRR is decomposed into two terms; that is, the recovered data is the sum of XZ and GX. Our method (see Eq. (5)), in contrast, integrates the transformed data $P^T X$ and the representation coefficient matrix Z into one united term; that is, the recovered data is $P^T XZ$. Through $P^T XZ$, the dual pursuit is formed and hence our method can be regarded as learning a locality-preserving dictionary for grouping-effect representation.

C. Comparison to Graph-Regularized LRR or LatLRR

In recent years, the graph regularization technique has been frequently introduced into the low-rank representation [7], [31], [44], [45], [48]. For example, Lu *et al.* [44] proposed a novel graph-regularized LRR approach by incorporating a graph Laplacian into LRR. Yin *et al.* [45] proposed a general Laplacian regularized low-rank representation method by using both a pairwise graph and hypergraph regularizers. These methods aim to incorporate a graph regularization technique into LRR, which can be generalized as follows:

$$\begin{aligned} &\arg\min_{Z,E} \quad \|Z\|_* + \frac{\lambda}{2}\sum_{ij} \| z_i - z_j \|_2^2 W_{ij} + \gamma\|E\|_1, \\ &\text{s.t.} \qquad X = XZ + E, \end{aligned} \tag{14}$$

where λ and γ are two balance parameters, z_i and z_j are the i-th column and j-th column of Z, respectively. Here, W_{ij}, defined in Eq. (2), is imposed on representation coefficient z_i and z_j.

However, these graph-regularized low-rank representation methods can not deal with the new coming data samples and limit their applications in classification. To this end, the graph regularization technique is introduced into the LatLRR method as follows:

$$\begin{aligned} &\arg\min_{Z,G,E} \quad \|Z\|_* + \|G\|_* + \gamma\|E\|_1 \\ &\qquad\qquad + \frac{\beta}{2} tr(ZL_Z Z^T) + \frac{\lambda}{2} tr(GL_G G^T), \\ &\text{s.t.} \qquad X = GX + XZ + E. \end{aligned} \tag{15}$$

Both Eq. (14) and Eq. (15) focus on imposing a graph regularization constraint of Z to expect to obtain an effective Z. In contrast, our method uses the graph regularization term $\sum_{ij} \| P^T x_i - P^T x_j \|_2^2 W_{ij}$ to learn a dictionary, and then expect to obtain an effective Z.

V. Experiments

Databases: In our experiments, four databases are adopted: PIE, Extended Yale B, COIL20, and USPS. Some image examples from Extended Yale B and COIL20 are displayed in Fig. 2.

Implementation Details: For each database, all the data is randomly divided into two groups. More specifically, for the PIE database, 15 images per class, that is, a total of 15 × 68 = 1020 images, are randomly selected as the first group of data and the remaining 612 images are randomly selected as the second group of data; for Extended Yale B, 30 images per class, that is, a total of 30 × 38 = 1140 images, are randomly selected as the first group of data and the remaining 1274 images are randomly selected as the second group of data; for COIL20, 5 images per class, that is, a total of 5 × 20 = 100 images, are randomly selected as the first group of data and the remaining 1340 images are randomly selected as the second group of data; for USPS, 20 images per class, that is, a total of 20 × 10 = 200 images, are randomly selected as the first group of data and the remaining 9098 images are randomly selected as the second group of data.

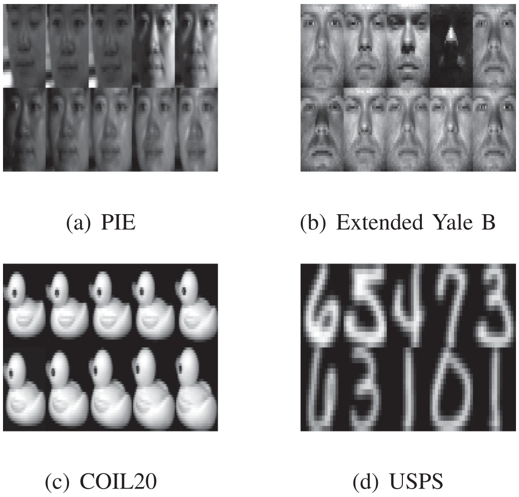
(a) PIE
(b) Extended Yale B
(c) COIL20
(d) USPS

Fig. 2. Some image examples from four databases.

In our experiments, the proposed method is optimized on the first group of data of each database, and then, the optimized variables Z^* and P^* are simultaneously obtained for the same database and the same parameters. After the optimized variables Z^* and P^* have been simultaneously obtained under the same parameters, Z^* is used to perform data clustering on the first group of data while P^* is used to perform data classification on the second group of data. Before that, we first discuss the effectiveness of the proposed dual pursuit.

A. Discussion About Dual Pursuit

In order to show the effectiveness of the proposed dual pursuit, it is compared with LatLRR and LRR in Fig. 3, where the optimal solution (L^*, Z^*) of LatLRR is visualized in the first row, the optimal solution (P^*, Z^*) of our method is visualized in the second row, and the optimal solution (Z^*) of LRR is visualized in the third row. From both the second and third rows, we can see that our method uses the transformed data $(P^*)^T X$ as a dictionary while LRR uses the data X itself as a dictionary (see the region marked by a dashed line). Since the three faces in X marked by yellow, blue, and green boxes are heavily illuminated, they are easily classified as the same group. In fact, the faces marked by yellow and blue boxes are in the same class while the face marked in the green box is in another class. Obviously, the dictionary used by LRR does not have the grouping-effect property. However, in our method,

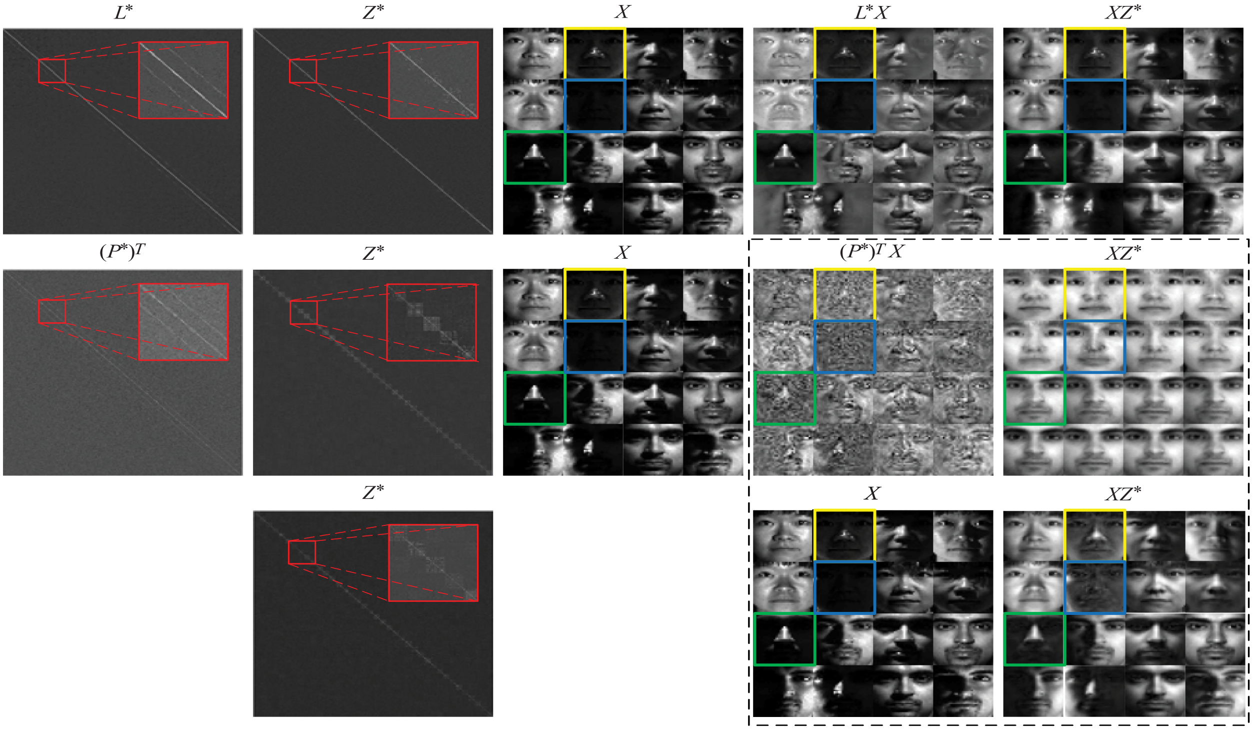

Fig. 3. Visualization of LatLRR and our method on Extended Yale B with 30 training samples. The optimal solution (L^*, Z^*) of LatLRR is shown in the first row and the optimal solution (P^*, Z^*) of our method is shown in the second row. The regions marked by red boxes are the enlarged areas. Moreover, a total of sixteen training samples from two classes are selected as X, where each class includes eight images. Based on such X, the corresponding (L^*X, XZ^*) of LatLRR and $((P^*)^T X, XZ^*)$ of our method are visualized to show the superiority of our method compared to LatLRR.

the intrinsic geometric structure of these three faces is captured by $(P^*)^T X$ and they are implicitly grouped into their respective classes. Therefore, we say that the dictionary $(P^*)^T X$ in our method has the grouping-effect property. Based on such a locality-preserving dictionary, our method will have an strong power to push the illumination corruptions into its error term while LRR will have a weak power to push the illumination corruptions into its error term. Hence, our method has a better grouping-effect representation than LRR based on a fixed dictionary, and this can also be observed from their optimized term XZ^*.

From both the first and second rows, we can see the two transformation matrices L^* and $(P^*)^T$ and their corresponding performance on feature extraction, as can be seen from $(P^*)^T X$ and L^*X in Fig. 3. Comparing the transformed data $(P^*)^T X$ with L^*X, it can be seen that our method can capture the important features, such as the eyes, nose, and mouth of a face, while LatLRR captures many redundant features of the face. This is because our method imposes a graph regularization technique on P while LatLRR enforces a low-rank criterion on L. Moreover, the performance of XZ^* in our method is also better than that of LatLRR. Therefore, our method with the dual pursuit compares favourably with LatLRR.

B. Experiment on Subspace Clustering

Data clustering aims to group samples into different groups. In this paper, after the low-rank representation coefficient matrix Z is obtained, K-means is used to cluster data points and then the values of two criteria, i.e., accuracy (AC) and normalized mutual information (NMI) are obtained to evaluate the clustering performance.

Given a data point x_i, let F and $\hat{F}$ be the ground truth label and the label produced by a clustering approach, respectively. Then the AC measure is defined by

$$AC = \frac{\sum_{i=1}^{n} \delta(\hat{F}(i), Match_{(\hat{F},F)}(i))}{n},$$

where n is the total number of samples and $\delta(a, b)$ is equal to 1 if $a = b$ and 0 otherwise. The $Match_{(\hat{F},F)}(i)$ is the best permutation mapping function that maps each cluster label $F(i)$ to the equivalent label from the database, which is fulfilled by the Kuhn-Munkres algorithm [49].

The NMI measure between two index sets K and $K^{'}$ is defined as

$$NMI(K, K^{'}) = \frac{MI(K, K^{'})}{\max(H(K), H(K^{'}))},$$

where $H(K)$ and $H(K^{'})$ denote the entropy of K and $K^{'}$, respectively. MI is defined as

$$MI(K, K^{'}) = \sum_{y \in K} \sum_{y \in K^{'}} p(x, y) \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right),$$

where $p(y)$ and $p(x)$ denote the marginal probability distribution functions of K and $K^{'}$, respectively, and $p(x, y)$ is the joint probability distribution function of K and $K^{'}$. Usually, $NMI(K, K^{'})$ ranges from 0 to 1, where the value 1 means that the two clusters are identical and the value 0 means that the two clusters are independent.

TABLE I
Clustering Accuracy (%) on the First Group of Data of the PIE Database, Where Num.# is the Number of Clusters

Cluster	AC(%)					NMI(%)				
Num.#	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*
4	95.00	86.67	71.67	93.33	**95.00**	88.10	73.48	71.36	83.24	**88.10**
30	63.33	60.89	57.11	**72.67**	71.78	79.49	76.32	75.47	**83.09**	81.12
56	60.71	60.48	59.64	65.24	**68.36**	75.51	75.19	78.56	79.66	**81.26**
68	60.85	61.57	63.04	62.84	**68.33**	75.43	79.09	80.88	80.36	**81.61**

TABLE II
Clustering Accuracy (%) on the First Group of Data of the Extended Yale B Database, Where Num.# is the Number of Clusters

Cluster	AC(%)					NMI(%)				
Num.#	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*
10	63.67	77.07	56.67	70.33	**77.33**	67.17	73.19	60.11	69.56	**74.38**
20	59.50	63.67	52.17	67.17	**68.00**	68.72	70.66	63.10	69.49	**70.82**
30	46.56	50.44	55.44	59.78	**63.44**	66.74	67.70	68.19	67.44	**68.37**
38	56.14	57.02	50.79	**58.07**	57.28	64.07	61.86	**65.22**	63.49	62.02

TABLE III
Clustering accuracy (%) on the First Group of Data of the COIL20 Database, where Num.# is the Number of Clusters

Cluster	AC(%)					NMI(%)				
Num.#	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*
5	60.00	64.00	72.00	56.00	**80.00**	53.57	58.66	66.10	62.87	**69.52**
10	60.00	62.00	60.00	70.00	**72.00**	69.65	72.11	66.04	79.53	**76.85**
15	62.67	65.33	61.33	66.67	**72.00**	74.02	77.40	74.58	77.42	**79.33**
20	65.00	62.00	68.00	**73.00**	71.00	79.51	75.90	79.00	**83.21**	80.02

Experimental Results: The proposed method is compared with four state-of-the-art subspace clustering methods based on low-rank representation LRR, LatLRR, smooth representation clustering (SMR) [48], and dual graph regularized latent low-rank representation (DGLRR) [7]. Here, SMR provides two ways to compute the edge weights of an undirected graph, and here we use the popular $J1$ for a fair comparison.

In order to show the robustness of the proposed method, the clustering experiments are implemented with different number of clusters. More specifically, on the first group data of each database, we use its first k classes for the corresponding data clustering experiments. The detailed clustering results are reported in four tables, where the red typeface indicates the best clustering result. From Tables I and III, it can be seen that our method basically outperforms the other state-of-the-art methods. From Table II, it can be seen that our method obviously outperforms LRR and LatLRR, but gives a comparable result to SMR and DGLRR when the number of clusters is $k = 38$. From Table IV, it can be seen that our method compares favorably with LRR. To sum up, our method gives the better clustering results than the other methods overall.

Generally, the clustering performance decreases as the number of clusters increases. As can be seen, the AC result obtained by the proposed method decreases as the number of clusters increases. However, it can be observed that there exists a fluctuation in terms of NMI in our proposed method. The phenomenon could be interpreted by the use of K-means, whose result may be different under different initializations of cluster centers. A similar phenomenon occurred strongly in the other methods in terms of both AC and NMI. This also shows that our method is more robust than the other methods.

Furthermore, the data decomposition abilities of the proposed method and LRR are compared in Fig. 4. By comparing the recovered data of our method (i.e., $(P^*)^T XZ^*$) with that of LRR (i.e., XZ^*), it can be seen that both two methods have good performance and our method is slightly better than LRR (see the red and yellow boxes). This is because both LRR and our method use the same low-rank representation, which can well separate the noise image from the original image to obtain a good recovery. Since the result of subspace clustering depends on the optimized representation coefficient matrix Z^*, we additionally display our XZ^* under the $(P^*)^T XZ^*$ (see the grey arrow). By comparing our XZ^* with that of LRR, it can be seen that our XZ^* can capture the generality characteristic of the intraclass faces/objects and achieves a high clustering accuracy. This is because, being based on a learned locality-preserving dictionary, our dictionary can capture the neighbor relation existing in original data (e.g., Extended Yale B and COIL20 databases)

TABLE IV
CLUSTERING ACCURACY (%) ON THE FIRST GROUP OF DATA OF THE USPS DATABASE, WHERE NUM.# IS THE NUMBER OF CLUSTERS

Cluster	AC(%)					NMI(%)				
Num.#	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*	LRR	LatLRR	SMR(J1)	DGLRR	*Ours*
4	**92.50**	65.00	85.00	88.75	91.25	**81.17**	63.20	67.71	77.52	80.51
6	78.33	81.67	83.33	81.67	**85.00**	68.26	73.78	73.60	71.49	**77.35**
8	81.87	72.50	73.12	83.13	**84.38**	73.65	67.70	70.18	75.24	**78.21**
10	**77.00**	63.00	67.50	68.50	76.00	73.79	60.41	70.34	68.07	**74.36**

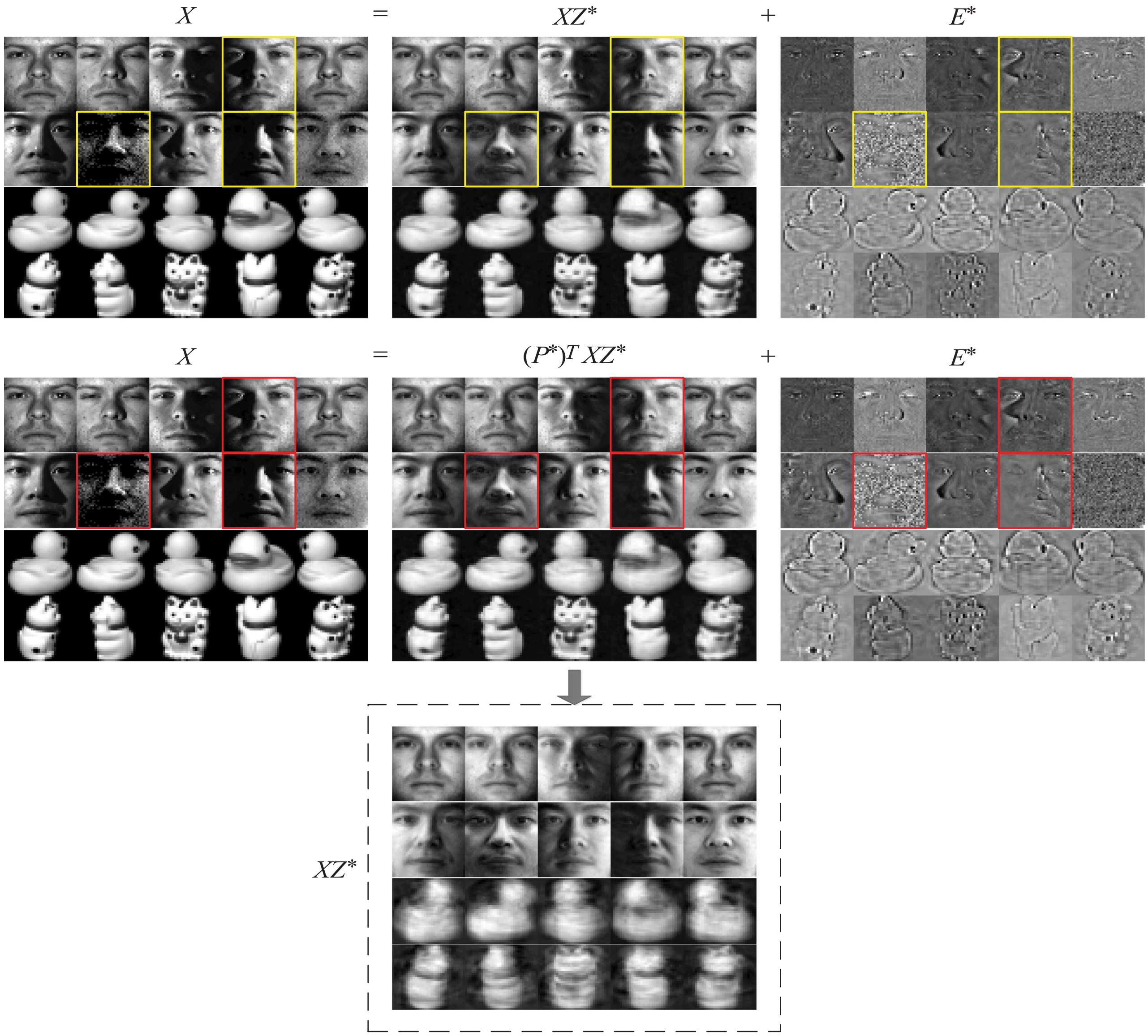

Fig. 4. Visualization of data decomposition of some samples from both the Extended Yale B and the COIL20 databases. A total of twenty samples from two databases (each database includes two classes and each class includes five images) are selected to display the performance of our method and LRR. For this total of twenty samples, the data decomposition of LRR is shown in the first row while ours is shown in the second row.

and thus our method may have a strong power to obtain the better global recovery than LRR based on a fixed dictionary (see the term XZ^*). Therefore, our method can obtain the better clustering result than LRR.

C. Experiment on Feature Extraction

In this section, we use P^* to perform classification for the second group of data. That is, for a training sample or testing sample x, the transformed feature vector y can be calculated as

TABLE V
CLASSIFICATION ACCURACY ON FOUR DATABASES. THE OPTIMAL DIMENSION IS SHOWN IN BRACKETS

Database	LPP	NPE	LatLRR	***Ours***
PIE	89.38(300D)	90.36(358D)	91.67(1024D)	**92.81**(1024D)
Extended Yale B	90.42(350D)	87.44(415D)	85.95(1024D)	**90.74**(1024D)
COIL20	74.18(35D)	73.06(47D)	**83.06**(1024D)	81.04(1024D)
USPS	76.28(50D)	76.42(68D)	**88.70**(256D)	84.03(256D)

follows:

$$y = (P^*)^T x. \tag{16}$$

It is worth noting that y has the same dimension as x, unlike in the general dimension reduction methods. After all the training samples (i.e., the first group data) and testing samples (i.e., the second group of data) have been transformed by P^*, the 1-nearest neighbour classifier is used to classify these testing samples in the transformed space.

Experimental Results: The proposed method is compared with LatLRR and two popular dimension reduction methods, including Locality Preserving Projection (LPP) [36] and Neighbourhood Preserving Embedding (NPE) [50]. For each database, the first group of data is used as the training samples and the second group of data is used as the testing samples. More specifically, for the PIE database, the selected 15 images per class are used as the training samples and the remaining images are used as the testing samples; on the Extended Yale B database, the selected 30 images per class are used as the training samples and the remaining images are used as the testing samples; for the COIL20 database, the selected five images per class are used as the training samples and the remaining images are used as the testing samples; for the USPS database, the selected 20 images per class are used as the training samples and the remaining images are used as the testing samples.

Table V shows that our method gives the best classification accuracy on the Extended Yale B and PIE databases. As is well known, the face images may reside on a nonlinear submanifold [51], [52] and the local feature extraction method may achieve a good classification performance. Therefore, LPP and NPE have achieved a good classification accuracy. However, our method outperforms them by 0.32–3.45%. This is because our method not only uses P to capture the local geometric structure but also uses Z to capture the global structures, and hence our method is likely to achieve the best performance on these two facial databases. On the COIL20 and USPS databases, our method has approximately the same performance to LatLRR. This is because both the COIL20 and USPS databases have a big distance between inter-classes and LatLRR is likely to achieve the best performance as a global feature extraction method.

Robustness to Noise: In order to test the robustness of our method to noise, which possibly appears in testing data, we design an additional experiment on the Extended Yale B database with 30 training samples per class. For each testing sample, we randomly corrupt some pixels. The value of a corrupted pixel is replaced by a random value that ranges uniformly from 0 to 1. We implement all methods on this noisy testing data and record the best result of each method. Fig. 5 shows that both our method and LatLRR are robust to noise, but our method outperforms LatLRR.

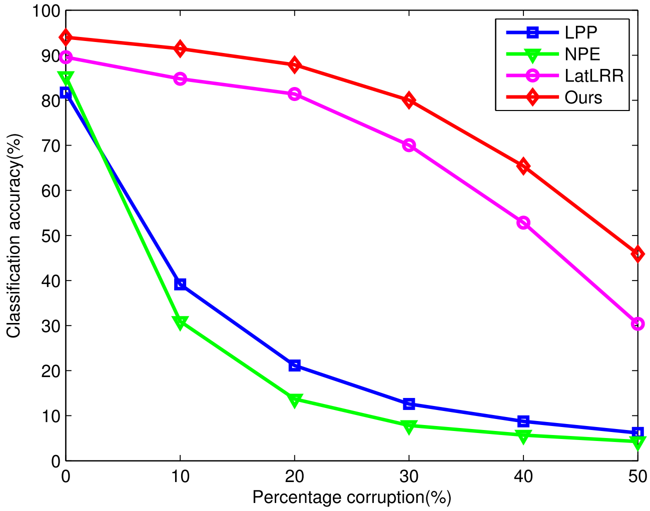

Fig. 5. Testing the robustness of all methods on the Extended Yale B database with 30 training samples. The classification accuracy (averaged over 20 runs) of all methods is drawn with different percentage corruption.

D. Parameter Settings and Convergence

There are two parameters in our objective function, namely λ and γ. To demonstrate the effects of these two parameters for experiments, different combinations of these values selected from a reasonable discrete set $\{1e^{-6}, 1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}, 1e^{2}, 1e^{3}, 1e^{4}\}$ are evaluated on each database. Specifically, the classification accuracy of each combination of parameter values is shown in Fig. 6, in which both the first and second columns represent the clustering performance of subspace clustering, and the third column represents the classification performance of feature extraction. In terms of subspace clustering experiments, there are different optimal parameters with different number of clusters k on the first group of data. Here, we give the clustering performance of each database with the largest number of clusters, that is, the PIE database with $k = 68$, the Extended Yale B database with $k = 38$, the COIL20 database with $k = 20$, and the USPS database with $k = 10$. In the case of the largest number of clusters, the AC performance is given in the first column of Fig. 6, and the NMI performance is given in the second column of Fig. 6. Comparing the first

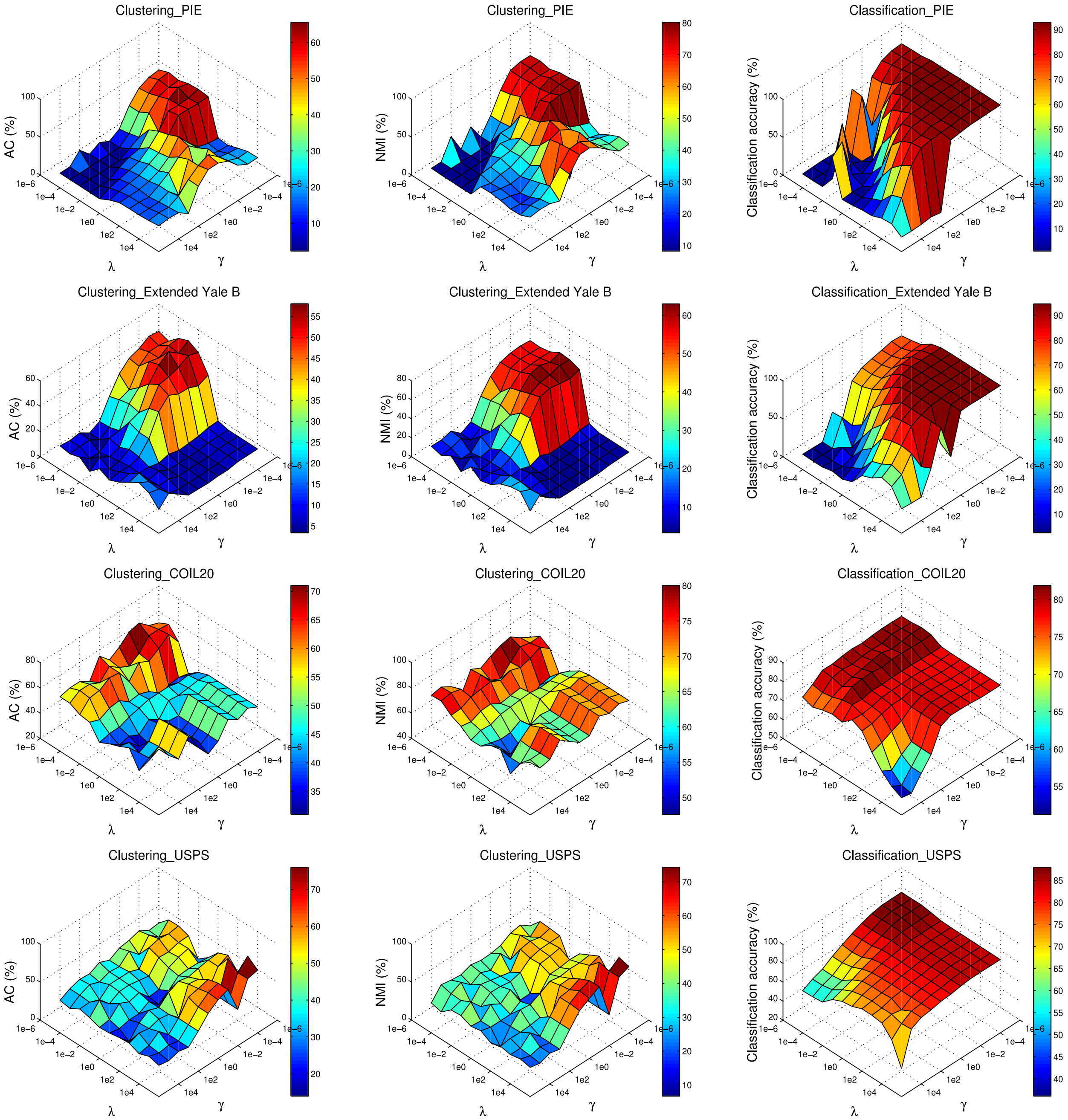

Fig. 6. The clustering and classification performances of our method versus the parameters λ and γ on four databases.

column with the second column, we can see that the AC performance and NMI performance have almost the same optimal parameters. Moreover, we give the classification performance of each database in the third column of Fig. 6. From the third column of Fig. 6, we can see that the classification performance is roughly consistent over a wide parameter range, which overlaps with the optimal parameter for the clustering performance. Therefore, we use the same parameter for subspace clustering and feature extraction. For example, for the PIE database, when $k = 68$, the parameters used are $\lambda = 0.032$ and $\gamma = 10^{-5}$. For the Extended Yale B database, when $k = 38$, the parameters used are $\lambda = 0.0007$ and $\gamma = 0.00003$. On the COIL20 database, when $k = 20$, the parameters used are $\lambda = 10^{-6}$ and $\gamma = 10^{-4}$. For the USPS database, when $k = 10$, the parameters used are $\lambda = 10^{3}$ and $\gamma = 10^{-6}$.

The convergence curves of our method are visualized in Fig. 7, where the maximum iteration number is 250. Generally speaking, the objective function value decreases as the number of iterations increases. As can be observed, our method achieves a fast convergence on both the Extended Yale B and COIL20 databases. However, on the PIE and USPS databases, the objective function value has a violent vibration. This phenomenon can be interpreted as the consequence of the inexact solution of Eq. (11), that is, the exact solution is permutated a little in our method by adding a Tikhonov regularization ηI to the inverse of the matrix $2\lambda X L X^T + \mu X Z_{k+1} Z_{k+1}^T X^T$. In this

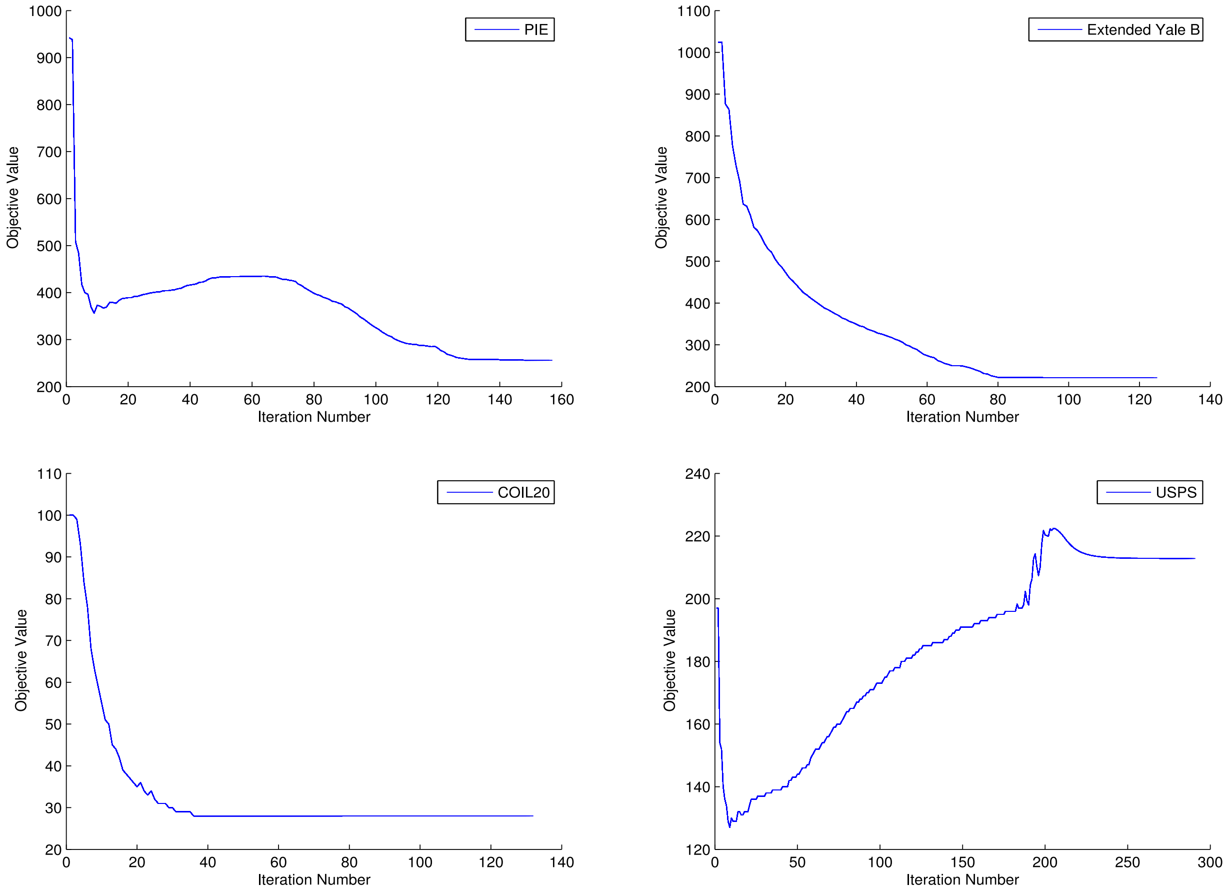

Fig. 7. Convergence curves of our method on four databases.

paper, $\eta = 0.001$ is used. In fact, the larger η is, the more violent the vibration is. But eventually, we can observe that the objective function value decreases steadily as the number of iterations continues to increase. This indicates that our method may achieve the final convergence after a long time.

VI. Conclusion

In this paper, from the view of learning a dictionary, we integrate the traditional subspace learning method into the low-rank representation and produce a dual pursuit for clustering and classification simultaneously. The proposed method provides us a robust unsupervised subspace clustering algorithm as well as a robust unsupervised feature extraction algorithm simultaneously. As an unsupervised feature extraction algorithm, our method shows the robustness to the noise and produces the compared classification results with the previous methods. As an unsupervised subspace clustering algorithm, our method shows a better clustering results than previous methods.

References

[1] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," *Proc. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[2] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[3] G. Liu *et al.*, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[4] X. Chen and Q. Wu, "Subspace weighting co-clustering of gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 52, no. 1, pp. 2–5, 2017.

[5] Y. Yan *et al.*, "Learning discriminative correlation subspace for heterogeneous domain adaptation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3252–3258.

[6] Y. Ye, Q. Wu, H. J. Zhexue, M. Ng, and X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognit.*, vol. 46, no. 3, pp. 769–787, 2013.

[7] M. Yin, J. Gao, Z. Lin, Q. Shi, and Y. Guo, "Dual graph regularized latent low-rank representation for subspace clustering," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4918–4933, Dec. 2015.

[8] A. Talwalkar *et al.*, "Distributed low-rank subspace segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3543–3550.

[9] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1615–1622.

[10] L. Zhuang *et al.*, "Non-negative low rank and sparse graph for semi-supervised learning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2328–2335.

[11] 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.

[12] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 4, pp. 831–849, 2018.

[13] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[14] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1832–1845, Oct. 2010.

[15] J. Yan and M. Pollefeys, “A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate,” in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 94–106.

[16] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, “Low-rank matrix recovery with structural incoherence for robust face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2618–2625.

[17] L. Ma, C. Wang, B. Xiao, and W. Zhou, “Sparse representation for face recognition based on discriminative low-rank dictionary learning,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2586–2593.

[18] D. Huang, R. S. Cabral, and F. De la Torre, “Robust regression,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 616–630.

[19] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, “Robust visual tracking via consistent low-rank sparse learning,” *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 171–190, 2014.

[20] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, “Low-rank sparse coding for image classification,” in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 281–288.

[21] C. Lang, G. Liu, J. Yu, and S. Yan, “Saliency detection by multitask sparsity pursuit,” *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, Mar. 2012.

[22] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, “Self-adaptively weighted co-saliency detection via rank constraint,” *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.

[23] Q. Gu and J. Zhou, “Co-clustering on manifolds,” in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 359–368.

[24] Q. Gu, J. Zhou, and D. Chris, “Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs,” in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 199–210.

[25] Y. Yang, F. Nie, D. Xu, and J. Luo, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 34, no. 4, 723–742, Apr. 2012.

[26] Y. Chen, H. Zhang, X. Zhang, and R. Liu, “Regularized semi-nonnegative matrix factorization for hashing,” *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1823–1836, Jul. 2018.

[27] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, “Generalized semi-supervised and structured subspace learning for cross-modal retrieval,” *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[28] M. Jian and C. Jung, “Semi-supervised bi-dictionary leaning for image classification with smooth representation-based lablel propagation,” *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 458–473, Mar. 2016.

[29] C. Luo, B. Ni, and S. Yan, “Image classification by selective regularized subspace learning,” *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, Jan. 2016.

[30] Y. Lu *et al.*, “Nuclear norm-based 2DLPP for image classfication,” *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2391–2403, Nov. 2017.

[31] J. Liu, Y. Chen, J. Zhang, and Z. Xu, “Enhancing low-rank subspace clustering by manifold regularization,” *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4022–4030, Sep. 2014.

[32] G. Liu, Q. Liu, and P. Li, “Blessing of dimensionality: Recovering mixture data via dictionary pursuit,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 47–60, Jan. 2017.

[33] C. Lang, J. Feng, S. Feng, J. Wang, and S. Yan, “Dual low-rank pursuit: Learning salient features for saliency detection,” *IEEE Trans. Neural Netw. Learn. Syst.* vol. 27, no. 6, pp. 1190–1200, Jun. 2016.

[34] L. Zhuang *et al.*, “Constructing a nonnegative low-rank and sparse graph with data-adaptive features,” *IEEE Trans. Image Process.* vol. 24, no. 11, pp. 3717–3728, Nov. 2015.

[35] Z. Ding and Y. Fu, “Robust multi-view subspace learning through dual low-rank decompositions,” *Proc. 13th AAAI Conf. Artif. Intell. Amer. Assoc. Artif. Intell.*, 2016, pp. 1181–1187.

[36] X. Niyogi, “Locality preserving projections,” *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160.

[37] M. Fazel, *Matrix Rank Minimization With Applications*, Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2002.

[38] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[39] Z. Lin, M. Chen, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” UIUC Tech. Rep. UILU-ENG-09-2215, Nov. 2009.

[40] Z. Wen, D. Goldfarb, and W. Yin, “Alternating direction augmented lagrangian methods for semidefinite programming,” *Math. Program. Comput.* vol. 2, no. 3/4, 2010, pp. 203–230.

[41] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

[42] R. Liu, Z. Lin, and Z. Su, “Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning.,” in *Proc. Asian Conf. Mach. Learn.*, 2013, pp. 116–132.

[43] D. Cai, X. He, and J. Han, “Spectral regression: A unified approach for sparse subspace learning,” in *Proc. Int. Conf. Data Mining*, 2007, pp. 73–82.

[44] X. Lu, Y. Wang, and Y. Yuan, “Graph-regularized low-rank representation for destriping of hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.* vol. 51, no. 7, pp. 4009–4018, Jul. 2013.

[45] M. Yin, J. Gao, and Z. Lin, “Laplacian regularized low-rank representation and its applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.

[46] F. R. Chung, *Spectral Graph Theory*, vol. 92. Providence, RI, USA: Amer. Math. Soc., 1997.

[47] M. Shao, D. Kit, and Y. Fu, “Generalized transfer subspace learning through low-rank constraint,” *Int. J. Comput. Vis.*, vol. 109, no. 1/2, pp. 74–93, 2014.

[48] H. Hu, H. Gao, Z. Lin, J. Feng, and J. Zhou, “Smooth representation clustering,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3834–3841.

[49] A. Frank, “On Kuhn’s hungarian method—A tribute from hungary,” *Naval Res. Logistics*, vol. 52, no. 1, pp. 2–5, 2005.

[50] X. He, D. Cai, S. Yan, and H. J. Zhang, “Neighborhood preserving embedding,” *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1208–1213.

[51] 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.

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

Shuangyan Yi received the M.S. degree in mathematics from Harbin Institute of Technology Shenzhen Graduate School, China. She is currently working toward the Ph.D. degree in computer science and technology at Harbin Institute of Technology Shenzhen Graduate School, China. Her current research interests include object tracking, pattern recognition, and machine learning.

Yingyi Liang received the B.S. degree from Wuhan University of Science and Technology, Wuhan, China, in 2007, and the M.S. degree from Guangdong University of Technology, Guangzhou, China, in 2010. He is currently working toward the Ph.D. degree at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include pattern recognition, computer vision, and machine learning.

Zhenyu He (SM’12) received the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2007. From 2007 to 2009, he worked as a Postdoctoral Researcher with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. He is currently a Full Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research interests include machine learning, computer vision, image processing, and pattern recognition.

Yi Li received the M.S. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, where he is currently working toward the Ph.D. degree. His research interests include visual tracking, image processing, and machine learning.

Yiu-Ming Cheung (F'17) is a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. His current research interests focus on artificial intelligence, visual computing, and optimization. Prof. Cheung is the Founding and the Past Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. Also, he is now serving as an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, KNOWLEDGE AND INFORMATION SYSTEMS, and the *International Journal of Pattern Recognition and Artificial Intelligence*, among others.