IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2017, accepted July 12, 2017, date of publication July 21, 2017, date of current version August 8, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2730281

Discriminant Manifold Learning via Sparse Coding for Robust Feature Extraction

MENG PANG[1], BINGHUI WANG[2], (Student Member, IEEE), YIU-MING CHEUNG[1], (Senior Member, IEEE), AND CHUANG LIN[3], (Member, IEEE)

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Corresponding author: Chuang Lin (chuang.lin@siat.ac.cn)

This work was supported in part by the Faculty Research Grant of Hong Kong Baptist University under Grant FRG2/16-17/051, in part by the National Natural Science Foundation of China under Grant 61272366 and Grant 61672444, in part by the SZSTI under Grant JCYJ20160531194006833, in part by the National Key Basic Research Development Program of China under Grant 2013CB329505, in part by the Shenzhen High-level Overseas Talent Program, Shenzhen Peacock Plan, under Grant KQCX2015033117354152, and in part by the Shenzhen Governmental Basic Research under Grant JCYJ20170413152804728.

ABSTRACT Most off-the-shelf subspace learning methods directly calculate the statistical characteristics of the original input images, while ignoring different contributions of different image components. In fact, to extract efficient features for image analysis, the noise or trivial structure in images should have little contribution and the intrinsic structure should be uncovered. Motivated by this observation, we propose a new subspace learning method, namely, discriminant manifold learning via sparse coding (DML_SC) for robust feature extraction. Specifically, we first decompose each input image into several components via dictionary learning, and then regroup the components into a more important part (MIP) and a less important part (LIP). The MIP can be considered as the clean portion of the image residing on a low-dimensional submanifold, while the LIP as noise or trivial structure within the image. Finally, the MIP and LIP are incorporated into manifold learning to learn a desired discriminative subspace. The proposed method is general for both cases with and without class labels, hence generating supervised DML_SC (SDML_SC) and unsupervised DML_SC (UDML_SC). Experimental results on four benchmark data sets demonstrate the efficacy of the proposed DML_SCs on both image recognition and clustering tasks.

INDEX TERMS Subspace learning, manifold learning, dictionary learning, feature extraction, image decomposition.

I. INTRODUCTION

In the areas of image processing and pattern recognition, the input image is always of quite high dimensionality, which makes it difficult to apply statistical techniques to conduct image analysis. Hence, it is of great importance to seek an efficient subspace representation (i.e., feature) with lower-dimensionality to represent the original image data. To this end, subspace learning methods, which lower down the dimensionality of the input images, have attracted considerable attentions in recent years [1]–[5].

One plausible assumption is that naturally occurring high-dimensional data probably lie on or close to a lower dimensional submanifold of the ambient space [6]. Benefiting from above assumption, manifold learning has been studied as one type of the most successful subspace learning techniques in past decade [7]–[14]. Typical manifold learning methods include locality preserving projection (LPP) [7], marginal fisher analysis (MFA) [10], locality sensitive discriminant analysis (LSDA) [11] and neighborhood sensitive preserving embedding (NSPE) [14] etc. Furthermore, to reserve useful structural information embedded in the original images, a series of tensor based manifold learning methods [15]–[18], e.g., two dimensional locality preserving projection (2DLPP) [15], tensor subspace analysis (TSA) [16] and two dimensional neighborhood preserving projection (2DNPP) [17], have been developed to work on

natural high-order of the input image data. Overall speaking, these manifold learning methods can achieve good performances for image recognition or clustering task. However, almost all above manifold learning methods directly compute the statistical characteristics of the original input images without eliminating the noise or trivial structure in advance, which would hinder the discovery of the intrinsic structure of images. Therefore, it is significant to find an efficient technique to conduct image decomposition and consider the characteristics of different image components.

Previous study has found that natural images can be represented by a small number of bases chosen from an over-complete code set [19], which provides the theoretical feasibility for image decomposition. Recently, with the rapid progress of norm minimization techniques [20]–[22], a variety of sparse coding (SC) and dictionary learning (DL) methods [23]–[27], e.g., sparse representation classifier (SRC) [23], Laplacian sparse coding (LapSC) [24], metaface learning (MFL) [25] and Fisher discrimination dictionary learning (FDDL) [26] have been developed to achieve sparse representations of original images under some learnt or pre-given bases. Furthermore, Zhang *et al*. [28] introduced dictionary learning techniques to decompose each image into several image components, and employed Fisher criterion to regroup these components into new representations for subspace learning. Tang *et al*. [29] presented a new face recognition framework by utilizing low-rank matrix recovery [30] to decompose each image into a low-rank part and a sparse error part, then combining them to encode a query face image with sparsity constraint for recognition. Although the ideas in [28] and [29] are encouraging and inspiring, the underlying structures of the regrouped representations in [28] and the decomposed parts in [29] have not been further studied. Besides, the regrouped representations in [28] cannot provide explicit intuitive explanations.

To exploit the contributions of different image components and meanwhile uncover the intrinsic structure in images, we propose a new subspace learning method called discriminant manifold learning via sparse coding (DML_SC) for robust feature extraction. DML_SC is a two-step method, i.e., 1) dictionary learning and feature regrouping and 2) graph embedding. In the first step, we leverage an efficient dictionary learning algorithm and utilize a regrouping criterion to decompose each input image into a more important part (MIP) and a less important part (LIP). One should note that the learnt MIP and LIP have explicit meaning. That is, the MIP can be viewed as the clean portion of the original image residing on a low-dimensional submanifold, and the LIP as trivial structure or noise. In the second step, we embed the MIP and LIP into a spectral graph [10] to learn a desired discriminative subspace, where the MIP is preserved while the LIP is suppressed. Also of note, since the regrouping criterion can be designed in both a supervised and an unsupervised way to tackle image recognition and clustering problems, we denote our DML_SC as supervised DML_SC (SDML_SC) and unsupervised DML_SC (UDML_SC), respectively. The contributions of our work can be summarized as follows:

- We propose an effective subspace learning method called DML_SC by integrating image decomposition and manifold learning into a unified model for robust feature extraction. In DML_SC model, the noise or trivial structure (i.e., LIP) and the clean portion (i.e., MIP) of images can be successfully separated, and their different contributions are both considered to learn a desired discriminative subspace.
- We develop two graph embedding algorithms in correspondence with the supervised regrouping and unsupervised regrouping criteria, and implement DML_SC in both the supervised form (SDML_SC) and unsupervised form (UDML_SC).
- We evaluate DML_SC on four benchmark image and face datasets. Experimental results demonstrate that our method achieves better performances on image recognition and clustering tasks than classical subspace learning methods and state-of-the-art sparse coding and dictionary learning methods.

Preliminary results of the method have been published in [31]. Compared with the conference version, this paper has made three major extensions. 1) The detailed description of dictionary learning procedure and the corresponding convergence curve are presented. 2) The parameter sensitivity of the proposed DML_SC is studied. 3) More extensive experiments are conducted to evaluate the recognition and clustering performances of the proposed method and compare with other state-of-the-art methods.

The rest of this paper is organized as follows: In Section 2, we will introduce the proposed DML_SC algorithm in details. The experimental results are presented in Section 3. Finally, we conclude this paper in Section 4.

II. DISCRIMINANT MANIFOLD LEARNING VIA SPARSE CODING (DML_SC)

A. MOTIVATION AND FLOWCHART

Given a matrix with N images $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \Re^{M \times N}$, each column of the matrix is a M-dimensional sample vector corresponding to an image. For each image $\mathbf{x}_i$, we expect to decompose it into two different parts, i.e., a more important part (MIP) $\mathbf{x}_i^m$ and a less important part (LIP) $\mathbf{x}_i^l$. Specifically, the MIP reserves main features of the images and contains vital discriminant information. In contrast, the LIP indicates noise or trivial structure containing interference information. As a result, the training image set $\mathbf{X}$ can also be split into two parts, i.e., $\mathbf{X}^m$ and $\mathbf{X}^l$, and rewritten as $\mathbf{X} \approx \mathbf{X}^m + \mathbf{X}^l$. Moreover, we aim to seek a desired discriminative subspace to enhance the MIP and suppress the LIP, so that the recognition and clustering performances can be improved. To this end, we attempt to learn a projection $\mathbf{P}$ to project $\mathbf{X}$ into a lower dimensional subspace, where the intrinsic structure

TABLE 1. The related symbols and the corresponding definitions.

Symbol	Definition
$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \Re^{M \times N}$	Original high-dimensional image data
$\mathbf{\Omega} \in \Re^{M \times d} (d \ll M),\ \mathbf{\Omega}^T \mathbf{\Omega} = \mathbf{I}$	Eigenfaces with orthonormal columns
$\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_N] \in \Re^{d \times N}$	Low-dimensional image data after dimensionality reduction
$\widehat{\mathbf{\Phi}} = [\widehat{\mathbf{d}}_1, \widehat{\mathbf{d}}_2, \cdots, \widehat{\mathbf{d}}_k] \in \Re^{d \times k},\ \widehat{\mathbf{d}}_i^T \widehat{\mathbf{d}}_i = 1, \forall i$	k-atom over-complete dictionary with normalized columns for low-dimensional image data $\widehat{\mathbf{X}}$
$\mathbf{\Lambda} = [\alpha_1, \alpha_2, \cdots, \alpha_N] \in \Re^{k \times N}$	Sparse coefficients matrix of $\widehat{\mathbf{X}}$ over dictionary $\widehat{\mathbf{\Phi}}$
$\mathbf{\Phi} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_k] \in \Re^{M \times k}$	k-atom sample-based dictionary for original high-dimensional image data $\mathbf{X}$

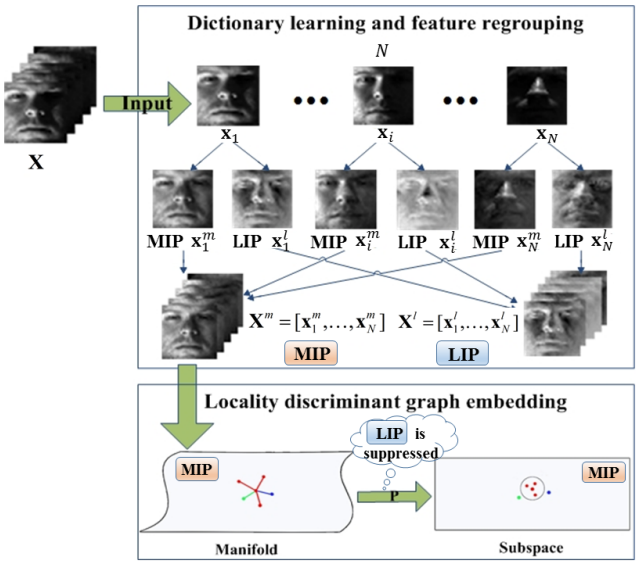

Fig. 1. The flowchart of the supervised DML_SC (SDML_SC). The points with same color indicate the MIP of images from the same class.

and energy of $\mathbf{X}^m$ on the manifold is preserved while the energy of $\mathbf{X}^l$ is suppressed. In what follows, the DML_SC is presented by posing it as a two-step scheme: 1) dictionary learning and feature regrouping and 2) graph embedding. The flowchart of the supervised DML_SC (SDML_SC) is shown in Fig. 1.

B. DICTIONARY LEARNING AND FEATURE REGROUPING

The detailed structure of dictionary learning and feature regrouping process can be stated formally as follows.

1) DICTIONARY LEARNING

In this phase, the key issue is how to decompose the training images $\mathbf{X}$ into different representation components. Motivated by the success of sparse coding and dictionary learning in image processing [23], [25], we expect to learn an adaptive over-complete dictionary $\mathbf{\Phi}$ to represent $\mathbf{X}$ under sparse coding constraints. Note that, since the dimensionality of the input image $\mathbf{x}_i$ is often very high, it is difficult to directly learn an over-complete dictionary for $\mathbf{X}$. To address the above issue, Zhang *et al.* [28] adopted the strategy by partitioning each image into several overlapped patches and thus learning a patch-based dictionary. In our work, instead of following the strategy described in [28], we prefer to generate Eigenfaces [1] to perform dimensionality reduction and also learn a sample-based dictionary. The reasons are twofold:

- First, since the atom number of patch-based dictionary is far larger than that of sample-based dictionary, learning a sample-based dictionary can be more beneficial to computational tractability.
- Second, the primary purpose in dictionary learning step is to decompose each image into several image components, while patch-based dictionary learning tends to find the representation components of each partitioned patch. Although we can use the partitioned patches to reconstruct the image, the global structure of the whole image cannot be well captured. Worse still, when handling the image that is corrupted or contaminated, some patches may be meaningless.

Hence, as stated above, we attempt to learn a k-atom over-complete sample-based dictionary with normalized columns, i.e., $\widehat{\mathbf{\Phi}} = [\widehat{\mathbf{d}}_1, \widehat{\mathbf{d}}_2, \cdots, \widehat{\mathbf{d}}_k] \in \Re^{d \times k}$ where $\widehat{\mathbf{d}}_i^T \widehat{\mathbf{d}}_i = 1$, for the low-dimensional image data, i.e., $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_N] \in \Re^{d \times N}$. According to [1], $\widehat{\mathbf{X}}$ is obtained as follows:

$$\mathbf{X} = \mathbf{\Omega}\widehat{\mathbf{X}} \rightarrow \mathbf{\Omega}^T \mathbf{X} = \mathbf{\Omega}^T \mathbf{\Omega} \widehat{\mathbf{X}} \xrightarrow{\mathbf{\Omega}^T \mathbf{\Omega} = \mathbf{I}} \widehat{\mathbf{X}} = \mathbf{\Omega}^T \mathbf{X}, \quad (6)$$

where $\mathbf{X} \in \Re^{M \times N}$ represent the original high-dimensional image data, $\mathbf{\Omega} \in \Re^{M \times d} (d \ll M)$ indicate Eigenfaces with orthonormal columns ($\mathbf{\Omega}^T \mathbf{\Omega} = \mathbf{I}$). Therefore, the optimization problem is defined as:

$$J(\widehat{\mathbf{\Phi}}, \mathbf{\Lambda}) = \underset{\widehat{\mathbf{\Phi}}, \mathbf{\Lambda}}{\operatorname{argmin}} (||\widehat{\mathbf{X}} - \widehat{\mathbf{\Phi}}\mathbf{\Lambda}||_F^2 + \gamma ||\mathbf{\Lambda}||_1), \quad (7)$$

where $\mathbf{\Lambda} = [\alpha_1, \alpha_2, \cdots, \alpha_N] \in \Re^{k \times N}$ is the sparse coefficients matrix of $\widehat{\mathbf{X}}$ over dictionary $\widehat{\mathbf{\Phi}}$, and γ is the regularization parameter. The related symbols and the corresponding definitions are listed in Table 1. It is worth noting that, the optimization problem in Eq. (2) is not jointly convex to $\widehat{\mathbf{\Phi}}$ and $\mathbf{\Lambda}$, but it is convex to $\widehat{\mathbf{\Phi}}$ or $\mathbf{\Lambda}$ when the other is fixed. Thus, we prefer to solve Eq. (2) by optimizing $\widehat{\mathbf{\Phi}}$ and $\mathbf{\Lambda}$, respectively. In this paper, we employ metaface learning (MFL) [25] to work it out. The detailed procedure is presented in Algorithm 1.

After obtaining $\widehat{\mathbf{\Phi}}$ and $\mathbf{\Lambda}$, the dictionary $\mathbf{\Phi} \in \Re^{M \times k}$ for the original image data $\mathbf{X} \in \Re^{M \times N}$ can be estimated by

Algorithm 1 Metaface Learning to Solve Problem (2)

Input: $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_N] \in \Re^{d\times N}$: image data after dimensionality reduction

Output: $\widehat{\mathbf{\Phi}} = [\widehat{\mathbf{d}}_1, \widehat{\mathbf{d}}_2, \cdots, \widehat{\mathbf{d}}_k] \in \Re^{d\times k}$: k-atom dictionary; $\mathbf{\Lambda} = [\alpha_1, \alpha_2, \cdots, \alpha_N] \in \Re^{k\times N}$: sparse coefficient of $\widehat{\mathbf{X}}$ over $\widehat{\mathbf{\Phi}}$

1: Initialize $\widehat{\mathbf{\Phi}}$: each column of $\widehat{\mathbf{\Phi}}$ is initialized as a random unit vector under l_2-norm constraint;

2: **repeat**

3: Fix $\widehat{\mathbf{\Phi}}$, solve $\mathbf{\Lambda}$: by fixing $\widehat{\mathbf{\Phi}}$, problem (2) becomes

$$J(\mathbf{\Lambda}) = \underset{\mathbf{\Lambda}}{\text{argmin}}(||\widehat{\mathbf{X}} - \widehat{\mathbf{\Phi}}\mathbf{\Lambda}||_F^2 + \gamma||\mathbf{\Lambda}||_1), \tag{1}$$

where $||.||_F$ is the Frobenius norm. The solution of (3) can be easily obtained by fast l_1-ls algorithm [20].

4: Fix $\mathbf{\Lambda}$, update $\widehat{\mathbf{\Phi}}$: by fixing $\mathbf{\Lambda}$, problem (2) becomes

$$J(\widehat{\mathbf{\Phi}}) = \underset{\widehat{\mathbf{\Phi}}}{\text{argmin}}\, ||\widehat{\mathbf{X}} - \widehat{\mathbf{\Phi}}\mathbf{\Lambda}||_F^2 \quad s.t.\ \widehat{\mathbf{d}}_i^T\widehat{\mathbf{d}}_i = 1, \quad \forall i. \tag{2}$$

Here, $\widehat{\mathbf{\Phi}}$ is updated column by column. Let $\mathbf{\Lambda} = [\boldsymbol{\beta}_1; \boldsymbol{\beta}_2; \cdots; \boldsymbol{\beta}_k] \in \Re^{k\times N}$, where $\boldsymbol{\beta}_i \in \Re^{1\times N}$ is a row vector. When updating $\widehat{\mathbf{d}}_i$, we fix all the other $\widehat{\mathbf{d}}_j (j \neq i)$. Thus, (4) is converted to

$$J(\widehat{\mathbf{d}}_i) = \underset{\widehat{\mathbf{d}}_i}{\text{argmin}}\, ||\widehat{\mathbf{X}} - \Sigma_{j\neq i}\widehat{\mathbf{d}}_j\boldsymbol{\beta}_j - \widehat{\mathbf{d}}_i\boldsymbol{\beta}_i||_F^2 \quad s.t.\ \widehat{\mathbf{d}}_i^T\widehat{\mathbf{d}}_i = 1, \quad \forall i. \tag{3}$$

Let $\mathbf{S} = \widehat{\mathbf{X}} - \Sigma_{j\neq i}\widehat{\mathbf{d}}_j\boldsymbol{\beta}_j$, problem (5) can be rewritten as the following lagrange multiplier function:

$$J(\widehat{\mathbf{d}}_i, \theta) = \underset{\widehat{\mathbf{d}}_i, \theta}{\text{argmin}}\, ||\mathbf{S} - \widehat{\mathbf{d}}_i\boldsymbol{\beta}_i||_F^2 - \theta(\widehat{\mathbf{d}}_i^T\widehat{\mathbf{d}}_i - 1). \tag{4}$$

By calculating the partial derivative of J with respect to $\widehat{\mathbf{d}}_i$ and setting it to 0, we obtain $\frac{\mathrm{d}J}{\mathrm{d}\widehat{\mathbf{d}}_i} = 0 \Longrightarrow \widehat{\mathbf{d}}_i = \mathbf{S}\boldsymbol{\beta}_i^T(\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T - \theta)^{-1}$. Note that $\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T - \theta$ is a constant, and $\widehat{\mathbf{d}}_i$ should be a unit vector, so we have the following normalization step:

$$\widehat{\mathbf{d}}_i = \mathbf{S}\boldsymbol{\beta}_i^T / ||\mathbf{S}\boldsymbol{\beta}_i^T||_2. \tag{5}$$

Repeating the above procedures for different i, $i = 1, 2\cdots, k$, then we can update all the dictionary basis $\widehat{\mathbf{d}}_i$, $i = 1, 2\cdots, k$, and hence the whole set $\widehat{\mathbf{\Phi}}$ is updated.

5: **until** the difference of $J(\widehat{\mathbf{\Phi}}, \mathbf{\Lambda})$ in adjacent iterations is smaller than a predefined threshold, or the maximum number of iterations is reached

$\mathbf{\Phi} \approx \mathbf{\Omega}\widehat{\mathbf{\Phi}}$, which is derived as follows:

$$\widehat{\mathbf{X}} \approx \widehat{\mathbf{\Phi}}\mathbf{\Lambda} \rightarrow \mathbf{\Omega}\widehat{\mathbf{X}} \approx \mathbf{\Omega}\widehat{\mathbf{\Phi}}\mathbf{\Lambda} \xrightarrow{\mathbf{X}=\mathbf{\Omega}\widehat{\mathbf{X}}} X \approx (\mathbf{\Omega}\widehat{\mathbf{\Phi}})\mathbf{\Lambda}. \tag{8}$$

Moreover, each column of $\mathbf{\Phi}$, i.e., $\mathbf{d}_i$, also satisfies $\mathbf{d}_i^T\mathbf{d}_i = \widehat{\mathbf{d}}_i^T\mathbf{\Omega}^T\mathbf{\Omega}\widehat{\mathbf{d}}_i = 1$. Hence, for each original image $\{\mathbf{x}_i\}_{i=1}^N$, we have $\mathbf{x}_i \approx \mathbf{\Phi}\alpha_i = \sum_{z=1}^k \mathbf{d}_z\alpha_{i,z}$, and $\alpha_{i,z}$ is the zth element of α_i. Let $\mathbf{x}_{i,z} = \mathbf{d}_z\alpha_{i,z}$, then $\mathbf{x}_i \approx \sum_{z=1}^k \mathbf{x}_{i,z}$. Thus each $\mathbf{x}_i$ can be approximately rewritten as the summation of k parts and each part is denoted by $\mathbf{x}_{i,z}$.

2) FEATURE REGROUPING

Considering whether the category information of training images is available or not, we leverage two reasonable criteria, i.e., *unsupervised* representation regrouping and *supervised* representation regrouping, to group each image into a MIP and LIP for effective graph embedding.

In the scheme of *unsupervised learning*, the representations are regrouped by their variances following the idea of principal component analysis (PCA) [1]. In general, if one representation part has a relative larger variance in image, then this part is more likely to cover the intrinsic features in image and can be more informative; whereas, the representation part with smaller variance is more likely to be noise. Accordingly, the variance of the zth representation $\mathbf{x}_{i,z}$ is defined as:

$$v_z = \sum_{i=1}^{N} (\mathbf{x}_{i,z} - \mu_z)^T(\mathbf{x}_{i,z} - \mu_z), \tag{9}$$

where μ_z denotes the mean of zth representation of all images, i.e., $\mu_z = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_{i,z}$.

To implement unsupervised representation regrouping, we expect to put the decomposed representations with larger v_z into the MIP $\mathbf{x}_i^m$, and the remaining into the LIP $\mathbf{x}_i^l$. To this end, we first reorder the decomposed k representations of $\mathbf{x}_i$ according to their variances v_z in a descending order. Then the first $\lfloor \tau \cdot k \rfloor$ representations, e.g., $\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,\lfloor\tau\cdot k\rfloor}\}$, are regrouped into the MIP $\mathbf{x}_i^m$, and the remaining representations, i.e., $\{\mathbf{x}_{i,\lfloor\tau\cdot k\rfloor+1}, \mathbf{x}_{i,\lfloor\tau\cdot k\rfloor+2}, \cdots, \mathbf{x}_{i,k}\}$, are regrouped into the LIP $\mathbf{x}_i^l$. τ is a scalar (e.g., $\tau = 0.8$) which decides the truncated threshold to generate the MIP and LIP of original image. Thus, the MIP $\mathbf{x}_i^m$ and LIP $\mathbf{x}_i^l$ can be calculated as follows:

$$\mathbf{x}_i^m = \mathbf{x}_{i,1} + \mathbf{x}_{i,2} + \cdots + \mathbf{x}_{i,\lfloor\tau\cdot k\rfloor} \tag{10}$$

$$\mathbf{x}_i^l = \mathbf{x}_{i,\lfloor\tau\cdot k\rfloor+1} + \mathbf{x}_{i,\lfloor\tau\cdot k\rfloor+2} + \cdots + \mathbf{x}_{i,k}. \tag{11}$$

For *supervised* learning, the category information can be leveraged to measure the discriminant capability of each representation part. In this scenario, the supervised feature regrouping criterion is designed to extract more discriminative features of image by ranking the discriminant capability of all representation parts. Considering that the maximum margin criterion (MMC) [32] can be employed to evaluate the discriminant capability of labeled samples, then the discriminative factor df_z of $\mathbf{x}_{i,z}$ is defined as:

$$df_z = \sum_{c=1}^{C}(\mu_z^c - \mu_z)^2 - \sum_{c=1}^{C}\frac{1}{N^c}\sum_{\mathbf{x}_i\in\mathbf{X}^c}(\mathbf{x}_{i,z} - \mu_z^c)^2, \quad (12)$$

where $\mathbf{X}^c$ denotes image set of the cth class, μ_z is the mean vector of $\mathbf{x}_{i,z}$, μ_z^c is the mean vector of $\mathbf{x}_{i,z}$ belongs to the cth class, and N^c indicates the number of the cth class.

Similarly for unsupervised representation regrouping, the representations possessing larger df_z are superposed to construct the MIP $\mathbf{x}_i^m$ and the remaining to form the LIP $\mathbf{x}_i^l$ of each face image $\mathbf{x}_i$. Finally, for whole image set, we obtain the MIP $\mathbf{X}^m = [\mathbf{x}_1^m, \cdots, \mathbf{x}_N^m]$, and the LIP $\mathbf{X}^l = [\mathbf{x}_1^l, \cdots, \mathbf{x}_N^l]$.

C. GRAPH EMBEDDING

After representation regrouping, the important issue is how to utilize the MIP and LIP of the input images to learn a desired subspace for recognition or clustering. In general, the MIP $\mathbf{X}^m$ plays a more important role than the LIP $\mathbf{X}^l$ on image recognition. However, the contribution of the LIP $\mathbf{X}^l$ should not be ignored, because the LIP is also useful to determine the projection directions. To summarize, it is desired to learn a discriminative subspace where the intrinsic structure and energy of the MIP $\mathbf{X}^m$ are both preserved, while the energy of the LIP $\mathbf{X}^l$ is suppressed simultaneously.

Inspired from LPP [7] and LSDA [11] methods, we respectively develop locality preserving graph embedding (LPGE) and locality discriminant graph embedding (LDGE), to keep consistency with above *unsupervised* representation regrouping and *supervised* representation regrouping. The illustrations of LPGE and LDGE are presented in Fig. 2.

1) LPGE

It is an unsupervised graph embedding, which aims to preserve the geometrical structure as well as the energy of the MIP, while suppressing the energy of the LIP. Similar to the strategy in LPP [7], we first construct two graphs G^m and G^l to depict the geometrical structure of the MIP $\mathbf{X}^m$ and LIP $\mathbf{X}^l$, respectively. Accordingly, the affinity weight matrices $\mathbf{W}^m$ and $\mathbf{W}^l$ are computed in a "simple-minded" way [8]; namely, $\mathbf{W}_{ij}^m = 1$ if $\mathbf{x}_i^m$ and $\mathbf{x}_j^m$ are connected in graph G^m, and $\mathbf{W}_{ij}^m = 0$ otherwise. The diagonal matrices $\mathbf{D}^m$, $\mathbf{D}^l$ and the Laplacian matrices $\mathbf{L}^m$, $\mathbf{L}^l$ of the MIP $\mathbf{X}^m$ and LIP $\mathbf{X}^l$ are also defined by $\mathbf{D}_{ii}^m = \Sigma_j \mathbf{W}_{ij}^m$, $\mathbf{D}_{ii}^l = \Sigma_j \mathbf{W}_{ij}^l$, $\mathbf{L}^m = \mathbf{D}^m - \mathbf{W}^m$ and $\mathbf{L}^l = \mathbf{D}^l - \mathbf{W}^l$.

Suppose the linear projection vector is $\mathbf{p}$, to preserve the geometrical structure of the MIP, we consider to map the connected MIP points in graph G^m into a subspace, where the similarity of these points are maintained. Therefore, we need to minimize the following objective function:

$$\sum_{i,j}(\mathbf{p}^T\mathbf{x}_i^m - \mathbf{p}^T\mathbf{x}_j^m)^2\mathbf{W}^m{}_{ij} = \mathbf{p}^T\mathbf{X}^m\mathbf{L}^m(\mathbf{X}^m)^T\mathbf{p}. \quad (13)$$

Moreover, LPGE also aims to preserve the global variance (i.e., energy) of the MIP $\mathbf{X}^m$ on the manifold, while suppressing the variance of the LIP $\mathbf{X}^l$. Suppose the MIP $\mathbf{X}^m$ and LIP $\mathbf{X}^l$ both have zero means, then the variance of $\mathbf{X}^m$ and $\mathbf{X}^l$ can be estimated as follows:

$$\sum_i ||\mathbf{p}^T\mathbf{x}_i^m||^2\mathbf{D}_{ii}^m = \mathbf{p}^T\mathbf{X}^m\mathbf{D}^m(\mathbf{X}^m)^T\mathbf{p} \quad (14)$$

$$\sum_i ||\mathbf{p}^T\mathbf{x}_i^l||^2\mathbf{D}_{ii}^l = \mathbf{p}^T\mathbf{X}^l\mathbf{D}^l(\mathbf{X}^l)^T\mathbf{p}. \quad (15)$$

In a nutshell, LPGE seeks for a projection $\mathbf{p}$ to minimize the variance of $\mathbf{X}^l$, while maximizing the variance of $\mathbf{X}^m$, and at the same time preserving the geometrical structure of $\mathbf{X}^m$ (refer to Table 2). Finally, we obtain the following objective function:

$$\min_{\mathbf{p}} \frac{\mathbf{p}^T\mathbf{X}^m\mathbf{L}^m(\mathbf{X}^m)^T\mathbf{p} + \rho\mathbf{p}^T\mathbf{X}^l\mathbf{D}^l(\mathbf{X}^l)^T\mathbf{p}}{(1-\rho)\mathbf{p}^T\mathbf{X}^m\mathbf{D}^m(\mathbf{X}^m)^T\mathbf{p}}, \quad (16)$$

where ρ is a balance parameter controlling the variance of the LIP. The projection vector $\mathbf{p}$ that minimizes the objective function in (16) is given by the minimum eigenvalue solution to the generalized eigen-problem [10]:

$$\Psi\mathbf{p} = \lambda(1-\rho)\mathbf{X}^m\mathbf{D}^m(\mathbf{X}^m)^T\mathbf{p}, \quad (17)$$

where $\Psi = \{\mathbf{X}^m\mathbf{L}^m(\mathbf{X}^m)^T + \rho\mathbf{X}^l\mathbf{D}^l(\mathbf{X}^l)^T\}$. Obviously, when $\rho = 0$, LPGE degenerates to LPP.

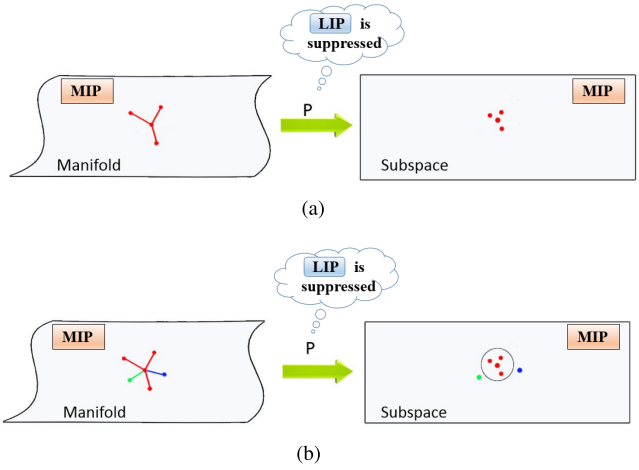

Fig. 2. (a) and (b) are illustrations of locality preserving graph embedding (LPGE) and locality discriminant graph embedding (LDGE), respectively. The points with same color belong to the same class.

TABLE 2. The motivations and the corresponding terms in LPGE.

Motivation	Term
Minimize the variance of $\mathbf{X}^l$	$\min\ \mathbf{p}^T\mathbf{X}^l\mathbf{D}^l(\mathbf{X}^l)^T\mathbf{p}$
Maximize the variance of $\mathbf{X}^m$	$\max\ \mathbf{p}^T\mathbf{X}^m\mathbf{D}^m(\mathbf{X}^m)^T\mathbf{p}$
Preserve the geometrical structure of $\mathbf{X}^m$	$\min\ \mathbf{p}^T\mathbf{X}^m\mathbf{L}^m(\mathbf{X}^m)^T\mathbf{p}$

2) LDGE

It is a supervised graph embedding, which aims to capture both the geometrical structure and the hidden discriminant information of the MIP, while suppressing the energy of the LIP. Similar to LSDA [11], we first construct two graphs, i.e., within-class graph G_w^m and between-class graph G_b^m to represent the MIP $\mathbf{X}^m$ of all training images. Meanwhile, we also construct a graph G^l to represent the LIP $\mathbf{X}^l$ of all training images. Let $l(\mathbf{x}_i^m)$ be the class label for $\mathbf{x}_i^m$ of the MIP. For each $\mathbf{x}_i^m$, its k nearest neighborhood set $N(\mathbf{x}_i^m) = \{\mathbf{x}_{i,1}^m, \mathbf{x}_{i,2}^m, \cdots, \mathbf{x}_{i,k}^m\}$ are split into $N_w(\mathbf{x}_i^m)$ and $N_b(\mathbf{x}_i^m)$. $N_w(\mathbf{x}_i^m)$ contain the neighbors sharing the same class label with $\mathbf{x}_i^m$, while $N_b(\mathbf{x}_i^m)$ contain the neighbors with different class labels of $\mathbf{x}_i^m$.

Let $\mathbf{W}_w^m$ and $\mathbf{W}_b^m$ be the weight matrices of G_w^m and G_b^m, respectively. $\mathbf{W}_{w\,ij}^m = 1$ if $\mathbf{x}_i^m \in N_w(\mathbf{x}_j^m)$ or $\mathbf{x}_j^m \in N_w(\mathbf{x}_i^m)$ and $\mathbf{W}_{b\,ij}^m = 1$ if $\mathbf{x}_i^m \in N_b(\mathbf{x}_j^m)$ or $\mathbf{x}_j^m \in N_b(\mathbf{x}_i^m)$. Let $\mathbf{W}^l$ be the weight matrix of G^l. Then, the Laplacian matrices $\mathbf{L}_w^m$, $\mathbf{L}_b^m$ and $\mathbf{L}^l$ are defined by $\mathbf{L}_w^m = \mathbf{D}_w^m - \mathbf{W}_w^m$, $\mathbf{L}_b^m = \mathbf{D}_b^m - \mathbf{W}_b^m$ and $\mathbf{L}^l = \mathbf{D}^l - \mathbf{W}^l$.

Now, we consider to preserve the geometrical structure and extract discriminant information of the MIP by mapping the MIP sample points in G_w^m and G_b^m into a low-dimensional subspace, where the connected points in G_w^m (i.e., within-class points) stay as close as possible while the connected points in G_b^m (i.e., between-class points) stay as distant as possible. Hence, for the MIP, we have the following two objective functions:

$$\min \sum_{i,j} (\mathbf{p}^T \mathbf{x}_i^m - \mathbf{p}^T \mathbf{x}_j^m)^2 \mathbf{W}_{w\,ij}^m \tag{18}$$

$$\max \sum_{i,j} (\mathbf{p}^T \mathbf{x}_i^m - \mathbf{p}^T \mathbf{x}_j^m)^2 \mathbf{W}_{b\,ij}^m, \tag{19}$$

with the constraint $\mathbf{p}^T \mathbf{X}^m \mathbf{D}_w^m (\mathbf{X}^m)^T \mathbf{p} = 1$. The objective function in Eq. (18) can be derived as follows:

$$\begin{aligned}
&\sum_{i,j} (\mathbf{p}^T \mathbf{x}_i^m - \mathbf{p}^T \mathbf{x}_j^m)^2 \mathbf{W}_{w\,ij}^m \\
&= \sum_{i,j} \mathbf{p}^T (\mathbf{x}_i^m - \mathbf{x}_j^m)(\mathbf{x}_i^m - \mathbf{x}_j^m)^T \mathbf{p} \mathbf{W}_{w\,ij}^m \\
&= 2tr\{\mathbf{p}^T \mathbf{X}^m \mathbf{D}_w^m (\mathbf{X}^m)^T \mathbf{p} - \mathbf{p}^T \mathbf{X}^m \mathbf{W}_w^m (\mathbf{X}^m)^T \mathbf{p}\} \\
&= 2tr\{\mathbf{p}^T \mathbf{X}^m (\mathbf{D}_w^m - \mathbf{W}_w^m)(\mathbf{X}^m)^T \mathbf{p}\} \\
&= 2tr\{\mathbf{p}^T \mathbf{X}^m \mathbf{L}_w^m (\mathbf{X}^m)^T \mathbf{p}\}.
\end{aligned}$$

In a similar way, the objective function in Eq. (19) can also be rewritten as $2tr\{\mathbf{p}^T \mathbf{X}^m \mathbf{L}_b^m (\mathbf{X}^m)^T \mathbf{p}\}$. Hence, the minimization problem in Eq. (18) and the maximization problem in Eq. (19) can be converted to $\max \mathbf{p}^T \mathbf{X}^m \mathbf{W}_w^m (\mathbf{X}^m)^T \mathbf{p}$ and $\max \mathbf{p}^T \mathbf{X}^m \mathbf{L}_b^m (\mathbf{X}^m)^T \mathbf{p}$, respectively.

Furthermore, we also aim to minimize the variance of the LIP $\mathbf{X}^l$, which is expressed as $\min \mathbf{p}^T \mathbf{X}^l \mathbf{D}^l (\mathbf{X}^l)^T \mathbf{p}$. Intuitively, the motivations and the corresponding terms of LDGE are summarized in Table 3. Consequently, the final optimization problem becomes:

TABLE 3. The motivations and the corresponding terms in LDGE.

Motivation	Term
Minimize the variance of $\mathbf{X}^l$	$\min\ \mathbf{p}^T \mathbf{X}^l \mathbf{D}^l (\mathbf{X}^l)^T \mathbf{p}$
Preserve the within-class points of $\mathbf{X}^m$	$\max\ \mathbf{p}^T \mathbf{X}^m \mathbf{W}_w^m (\mathbf{X}^m)^T \mathbf{p}$
Suppress the between-class points of $\mathbf{X}^m$	$\max\ \mathbf{p}^T \mathbf{X}^m \mathbf{L}_b^m (\mathbf{X}^m)^T \mathbf{p}$

$$\begin{aligned}
&\max_{\mathbf{p}}\ \mathbf{p}^T \{\mathbf{X}^m (\eta \mathbf{W}_w^m + \beta \mathbf{L}_b^m)(\mathbf{X}^m)^T - \epsilon \mathbf{X}^l \mathbf{D}^l (\mathbf{X}^l)^T\} \mathbf{p} \\
&s.t.\ \mathbf{p}^T \mathbf{X}^m \mathbf{D}_w^m (\mathbf{X}^m)^T \mathbf{p} = 1,
\end{aligned} \tag{20}$$

where η, β, ϵ are trade-off parameters with $\eta + \beta + \epsilon = 1$. The optimization problem in Eq. (20) leads to solving the following generalized eigenvalue problem:

$$\boldsymbol{\Upsilon} \mathbf{p} = \lambda \mathbf{X}^m \mathbf{D}_w^m (\mathbf{X}^m)^T \mathbf{p}, \tag{21}$$

where $\boldsymbol{\Upsilon} = \mathbf{X}^m (\eta \mathbf{W}_w^m + \beta \mathbf{L}_b^m)(\mathbf{X}^m)^T - \epsilon \mathbf{X}^l \mathbf{D}^l (\mathbf{X}^l)^T$. When $\epsilon = 0$, LDGE reduces to LSDA. Also of note, in the following recognition experiments, the optimal hyper-parameters, i.e., α, β, and ϵ, are chosen via grid search based cross validation.

III. EXPERIMENT RESULTS

In this section, several experiments are conducted to show the efficacy of the proposed method. For simplicity, we denote the proposed supervised DML_SC as SDML_SC, and unsupervised DML_SC as UDML_SC. We conduct face recognition experiments on three benchmark face datasets, i.e., CMU PIE, Extended YaleB and FERET, to evaluate the classification performance of SDML_SC. Furthermore, we also conduct image clustering experiments on COIL20 image library and Extended YaleB face database to test the clustering ability of UDML_SC. All experiments are conducted on a PC (CPU: Intel Core i7-4790K 4.00GHz, RAM: 16GB). The data preparation and representation are described below.

A. DATA PREPARATION AND REPRESENTATION

The CMU PIE dataset [33] consists of 68 subjects with 41,368 face images, and each subject involves 43 different illumination conditions, 13 different poses, and 4 different expressions. Following the experimental settings in [12] and [14], we also selected a subset of 1700 images of 10 subjects that contain five near frontal poses (C05, C07, C09, C27, C29) and all the images involve different expressions and illuminations. In the experiment, we randomly selected 85 images per subject include 850 images in total for training and the other images for testing.

The Extended YaleB (E-YaleB) dataset [34] consists of 38 subjects with 2,414 frontal face images under 9 poses and 64 illumination conditions, which is more challenging than CMU PIE dataset. In recognition experiment, we randomly took 32 images of 20 subjects thus in total 640 images as the training set and the remaining face images

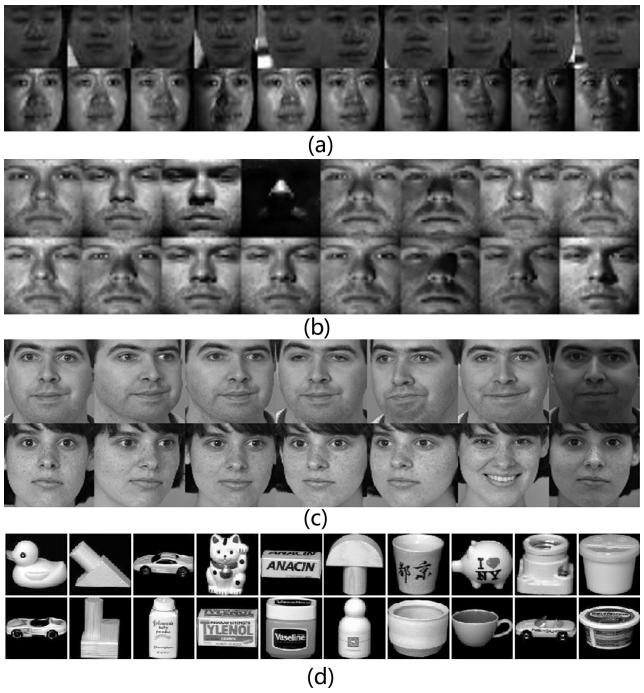

Fig. 3. Some sample images of four benchmark datasets: (a) CMU PIE; (b) E-YaleB; (c) FERET; (d) COIL20.

(nearly 32 images per subject) as the testing set. All face images were cropped and normalized to 32 × 32 pixels.

The FERET dataset [35] contains 14,126 images from 1199 subjects. Similar to the previous work in [36] and [37], we selected a subset of 700 face images of 100 subjects who have the same set of poses from the FERET dataset. For each subject, there are seven different images whose names are marked with two character strings: ''ba'', ''bj'', ''bk'', ''be'', ''bf'', ''bd'', and ''bg'' to denote different facial poses, illuminations and expressions. Each image in this subset was manually cropped and resized to 40 × 40 pixels. For recognition experiment, we randomly took 4 images per subject to form the training set, i.e., 4 × 100. The rest 300 images were taken to form the testing set.

The COIL 20 image library [38] contains gray scale images of 20 objects viewed from varying angles and each object has 72 images. The COIL20 is one of the most common datasets to test clustering performance of algorithms. Fig. 3 presents some sample images in the CMU PIE, E-YaleB, FERET and COIL20 datasets, respectively.

For recognition experiments, to learn an over-complete dictionary for SDML_SC, Eigenface [1] was first applied to reduce the dimensionality of face images to 300. In each face dataset, in order to prove the convergence of the dictionary learning (DL) procedure in Algorithm 1, we make a drawing where the objective function $J(\widehat{\mathbf{\Phi}}, \mathbf{\Lambda})$ versus the iteration number is illustrated. It can be seen from Fig. 4 that the DL technique in SDML_SC achieves fast convergence speed over three tested face datasets. Moreover, to provide an intuition of the MIP and LIP, we also conduct a face representation experiment using the supervised representation regrouping criterion to show the MIP and LIP of a given face image. The MIP and LIP of given face images on three face datasets using supervised representation regrouping criterion ($\tau = 0.8$) are shown in Fig. 5. We observe that, the dictionary learning and feature regrouping procedure in the proposed DML_SC method can successfully separate the noise or trivial structure (e.g., illumination or shadow) and the relatively clean portion of the original image. As a result, the learnt MIP and LIP both have explicit intuitive explanations.

B. FACE RECOGNITION RESULTS

In this subsection, three benchmark face datasets: CMU PIE, E-YaleB and FERET are adopted to evaluate the classification performance of the proposed SDML_SC method. For each dataset, the recognition experiments are repeated 5 times with different training-testing partitions.

1) COMPARING ALGORITHMS

To demonstrate how the face recognition performance can be improved by the proposed SDML_SC method, ten popular subspace learning methods including PCA [1], graph-based LDA [7], supervised LPP (sLPP) [7], supervised neighborhood preserving embedding (sNPE) [9], sparsity preserving projection (SPP) [39], LSDA [11], NSPE [14], discriminative sparsity preserving projections (DSPP) [40] and two tensor based manifold learning methods, i.e., 2DLPP [15] and 2DNPP [17] are used for comparison. Moreover, we also compare SDML_SC with recent dictionary learning methods, i.e., SRC [23], MFL [25] and the state-of-the-art FDDL [26].

2) PARAMETER SETTING

For PCA and SPP, the only parameter is the subspace dimension. The model parameters for sLPP, sNPE, LSDA and NSPE are empirically configured according to [31]. For DSPP, we set the value of trade-off parameter ρ as 0.0005 to obtain optimal recognition results. The values of regularization parameter λ for SRC and MFL are fixed to 0.01. For FDDL, the parameters are chosen via cross-validation as depicted in [26]. Moreover, to conduct a fair comparison with dictionary learning methods (i.e., SRC, MFL and FDDL), the above dictionary learning methods and SDML_SC all generate Eigenfaces to perform DR, the reduced dimension of each feature is set as 300. The atom numbers of the over-complete dictionaries for SDML_SC on CMU PIE, E-YaleB and FERET datasets are set as 680, 500 and 300, respectively.

3) RECOGNITION RESULTS

Table 4 lists the top average recognition rates of SDML_SC and the comparing subspace learning methods, while Table 5 reports the performances of SDML_SC and the comparing dictionary learning methods. We highlight the best and comparable results in bold font and underline the second best ones. From Table 4, we can observe that the top average recognition rates of SDML_SC are higher than other comparing subspace learning methods. Particularly on

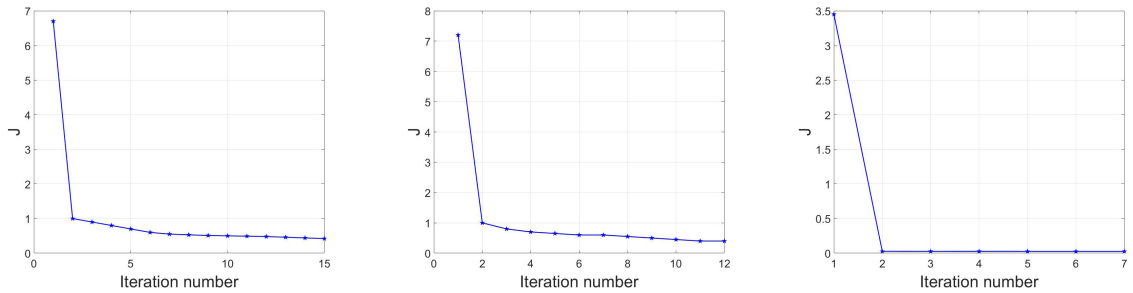

Fig. 4. Convergence of the dictionary learning process on (a) CMU PIE; (b) E-YaleB; (d) FERET.

Fig. 5. (a) to (c) are the results on the CMU PIE, E-YaleB and FERET datasets. From left to right are: a face sample; the MIP and the LIP after supervised feature regrouping.

TABLE 4. Top average recognition rates (%) of different subspace learning methods (Best; Second Best).

Methods	CMU PIE	E-YaleB	FERET	Year
PCA	83.9	53.0	70.6	1991
LDA	91.6	75.2	57.0	1997
sLPP	90.3	76.5	57.3	2005
sNPE	92.1	83.2	58.3	2007
2DLPP	79.2	75.9	58.0	2007
2DNPP	86.9	80.3	70.3	2012
SPP	89.2	77.2	66.7	2010
LSDA	91.4	78.0	60.6	2007
NSPE	92.1	76.2	75.6	2014
DSPP	93.8	83.4	67.7	2015
SDML_SC	**94.9**	**88.0**	**84.0**	

E-YaleB and FERET datasets, SDML_SC boosts over 4% recognition rates compared to DSPP and NSPE methods, and even improves the recognition rates of sLPP and sNPE by a margin as large as 4.8%-25.7%. Regarding the tensor based manifold learning methods, 2DLPP and 2DNPP perform close to their 1D versions (i.e., sLPP and sNPE) on CMU PIE and E-YaleB datasets, and obtain better recognition results on FERET dataset. Furthermore, we are interested to find that, although SPP and PCA are unsupervised methods, PCA performs modestly well on FERET dataset, and SPP even obtains comparable recognition results compared to supervised methods (i.e., LDA and sLPP) on CMU PIE and E-YaleB datasets.

As shown in Table 5, the proposed SDML_SC still obtains promising recognition results among the comparing sparse coding and dictionary learning methods. In addition, FDDL ranks the first and the second on CMU PIE and E-YaleB datasets, respectively, while performing worse than SDML_SC and other two dictionary learning methods on FERET dataset. MFL performs stably across all the

TABLE 5. Top average recognition rates (%) and corresponding number of dictionary atoms of different dictionary learning methods (Best; Second Best).

Methods	CMU PIE	E-YaleB	FERET	Year
SRC	92.5(850)	70.1(640)	78.3(400)	2009
MFL	93.3(590)	82.9(440)	79.0(400)	2010
FDDL	**97.8(850)**	**87.9(640)**	67.0(400)	2014
SDML_SC	94.9(680)	**88.0(500)**	**84.0(300)**	

datasets and gets goodish results next to DML_SC. Although SRC achieves similar recognition accuracy to SDML_SC and MFL on CMU PIE dataset, it is not competitive with SDML_SC and MFL on E-YaleB dataset.

Fig.6 (a)-(c) present the top average recognition rates of the involved subspace learning methods (except tensor based methods) versus the variation of feature dimensions on CMU PIE, E-YaleB and FERET datasets, respectively. We observe that SDML_SC outperforms other comparing methods almost across all the dimensions, which confirms the effectiveness and rationality to leverage the contributions of different image components for image recognition.

4) EXPLANATION OF ABOVE EXPERIMENTAL RESULTS

- PCA is a typical unsupervised subspace learning method, which usually performs much worse than other discriminative subspace learning methods. Hence, PCA is usually applied as a baseline method in different face recognition tasks. However, when few samples of each person are available (e.g, FERET), most supervised subspace learning methods, such as LDA, sLPP, sNPE and LSDA, suffer serious performance drop and reduce to baseline PCA or even worse [41].
- SPP is also an unsupervised subspace learning method, which aims to preserve the sparse reconstructive

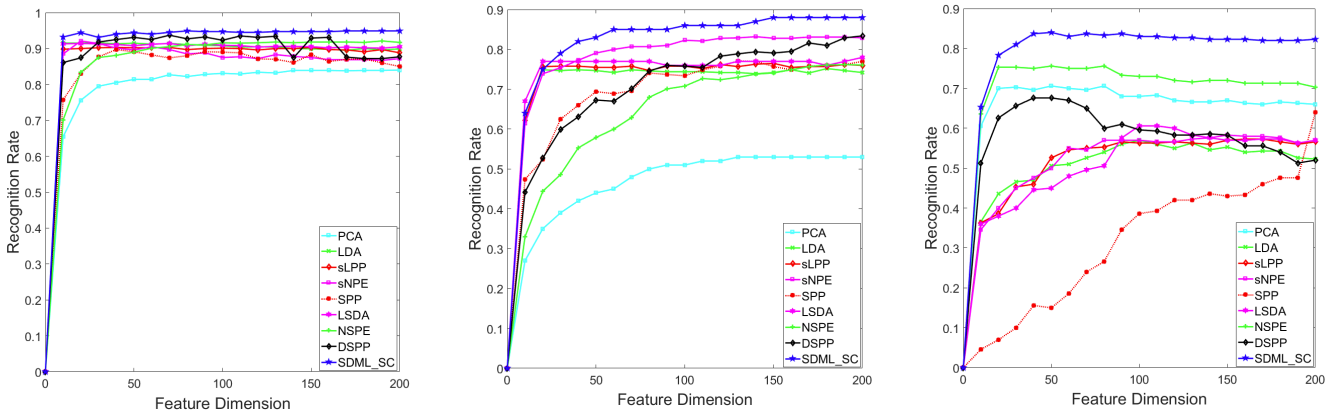

Fig. 6. (a) to (c) are the top average recognition rates of all comparing subspace learning methods versus dimensions on the CMU PIE, E-YaleB and FERET datasets, respectively.

relationship of the image samples. Although category information is not needed, SPP tends to find the discriminative mapping since the sparse representation (SR) has natural discriminating power. Consequently, SPP obtains comparable or even better recognition results compared to supervised methods (e.g., LDA and sLPP) over three tested datasets.

- DSPP, LSDA, NSPE, sNPE and sLPP are supervised graph-based manifold learning methods. Specifically, DSPP, LSDA and NSPE are all two-graphs based supervised methods, which construct within-class adjacency graph and between-class penalty graph, to find a projection maximizing the margin between data points from different classes. Notably, DSPP even applies the core idea of SPP [39] by minimizing a l_1-regularization related objective function to preserve the sparse reconstruction relationships of within-class samples. By contrast, sLPP and sNPE only construct within-class adjacency graph, while neglecting the relationship between classes. Consequently, DSPP, LSDA and NSPE generally perform better than sLPP and sNPE on three tested datasets.
- 2DLPP and 2DNPP are developed to directly work on input image matrix rather than concatenated vectors, which generally require extra coefficients for representing an image as compare to other vector-based methods [42]. Moreover, by exploiting the spatial structure information embedded in limited training data, 2DLPP and 2DNPP outperform their 1D versions (i.e., sLPP and sNPE) on FERET dataset.
- SRC, MFL and FDDL are three recent dictionary learning methods. SRC straightforwardly uses the original training images to construct dictionary, while MFL tries to learn an enriched but more representative dictionary based on training images. Compared to the two former dictionary learning methods, FDDL pays attention to training a class-specific dictionary with discriminative representation coefficients, by reducing the within-class variation and expanding the between-class differences correspondingly [26]. Hence, FDDL can be more robust against facial variations in training images than SRC and MFL. One should note that, since FDDL aims to distinguish the reconstructive ability of within-class image samples and between-class image samples in its objective function, the scarcity of training samples per subject (refer to FERET dataset) would still be a key factor to undermine the performance of FDDL.
- DML_SC is designed as a two-step subspace learning method, i.e., 1) dictionary learning and feature regrouping and 2) graph embedding. In the first step, DML_SC attempts to separate the noise or trivial structure (i.e., LIP) from original data, and obtain relatively clean portion of image (i.e., MIP). Subsequently, SDML_SC tires to learn a desired discriminative subspace where the intrinsic structure and the hidden discriminant information of the MIP are both captured, while the energy of the LIP is suppressed simultaneously. These procedures enable SDML_SC be robust against severe variation of illuminations and image noises existed in raw image data. Therefore, SDML_SC consistently outperforms the comparing subspace learning methods with NN classifier, and even obtains better recognition results compared to recent dictionary learning methods over three face datasets.

5) STUDY OF PARAMETER SELECTION

In this subsection, we probe further the effects of two key parameters of the proposed SDML_SC: the balance parameter ϵ in locality discriminant graph embedding (LDGE) and the control parameter τ in feature regrouping phase, over CMU PIE, E-YaleB and FERET three benchmark datasets.

In SDML_SC model, we observe that the value of parameter ϵ is controlled by the constraint $\epsilon = 1 - (\eta + \beta)$. Inspired from the strategy in LSDA [11], we consider the

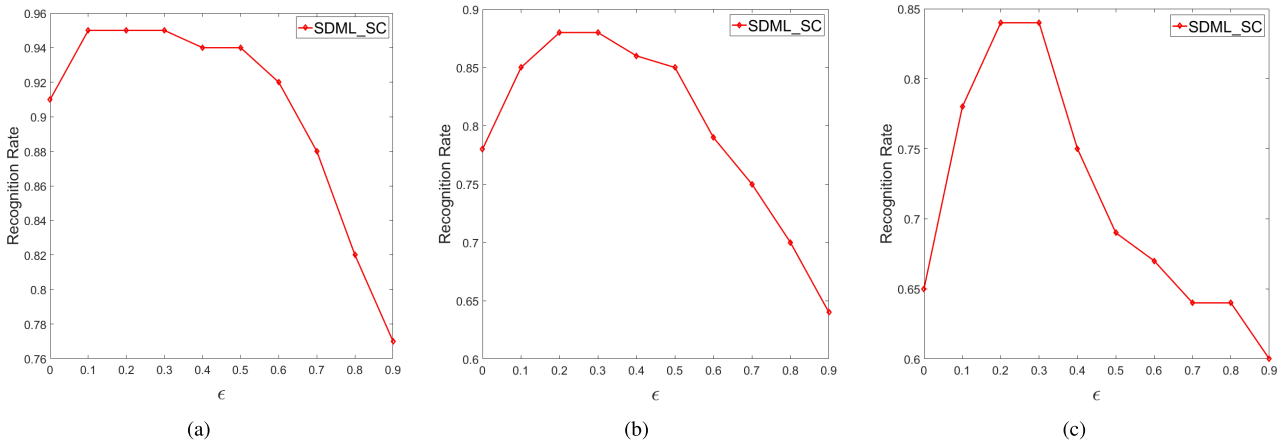

Fig. 7. (a) to (c) are the recognition rates versus the parameter ϵ (from 0.0 to 0.9) of SDML_SC on the CMU PIE, E-YaleB and FERET datasets.

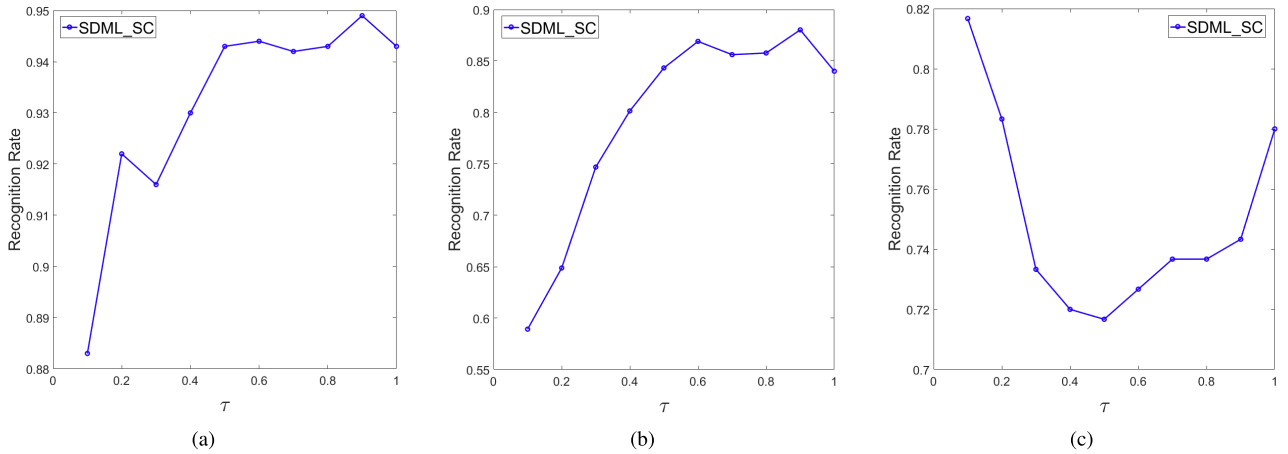

Fig. 8. (a) to (c) are the recognition rates versus the parameter τ (from 0.1 to 1.0) of SDML_SC on the CMU PIE, E-YaleB and FERET datasets.

combination of η and β as a single parameter, and regard it as a function of ϵ. Hence, we choose to evaluate the performance of SDML_SC on the effect of ϵ separately.[1] Furthermore, the control parameter τ decides the truncated threshold to generate the MIP and LIP of original image in feature regrouping phase, which can also be a key factor to affect the classification performance of SDML_SC.

Fig. 7 and Fig. 8 present the effects of two parameters ϵ and τ on the recognition rates of SDML_SC over CMU PIE, E-YaleB and FERET three datasets, respectively. As shown in Fig. 7, the balance parameter ϵ of SDML_SC that controls the energy of the LIP is playing roles for face recognition. Specifically, when $0 < \epsilon \leq 0.6$, the recognition results of SDML_SC are higher than that of LSDA ($\epsilon = 0$, SDML_SC degrades to LSDA). When the value of ϵ is extremely large, i.e., 0.8 or 0.9, the SDML_SC performs worse than LSDA. A reasonable explanation could be that ϵ controls the energy of the LIP in SDML_SC, the excessive weights on the LIP may cover up the influence of graph embedding by the MIP and bring about negative effect on learning desired discriminative subspace. It is worth mentioning that, SDML_SC consistently achieves goodish performance with the ϵ varying from 0.1 to 0.3 over three tested datasets. Regarding the parameter τ, as shown in Fig. 8, the recognition rates of SDML_SC present a rising tendency with the increase of the value of τ, and achieve best results ranging from 0.7-0.9 on CMU PIE and E-YaleB datasets. By contrast, SDML_SC obtains stable recognition results across different values of parameter τ, and performs especially well at both ends of values of parameter τ on FERET dataset.

6) SUMMARY

- SDML_SC consistently outperforms classical subspace learning methods, and obtains comparable or better recognition results compared to state-of-the-art dictionary learning methods.
- SDML_SC is, to some extent, sensitive to the variation of parameters ϵ and τ. However, in practical

[1]Through experiments, we observed that when the value of ϵ was fixed, we always obtained stable recognition results when tuning different combinations of η and β.

TABLE 6. Database description.

Dataset	Size	Dimensionality	Number of classes
COIL20	1440	1024	20
E-YaleB	2414	1024	38

applications, we can leverage model selection techniques, e.g., grid search-based cross validation, to choose the optimal ϵ and τ.

C. IMAGE CLUSTERING RESULTS

In this subsection, we conduct clustering experiment on COIL20 and more challenging E-YaleB face database (listed in Table 6). In order to randomize the experiments, we evaluate the clustering performance with different number of clusters ($k = 4, 8, 12, 16, 20$). For each given cluster number, 5 tests are conducted on different randomly chosen classes.

1) COMPARING ALGORITHMS

Nine popular clustering algorithms including K-means, nonnegative matrix factorization (NMF) [43], LapSC [24], graph regularized nonnegative matrix factorization (GNMF) [44], nonnegative local coordinate factorization (NLCF) [45], sparse concept coding (SCC) [46], graph regularized nonnegative matrix factorization with sparse coding (GRNMFSC) [47] and the state-of-the-art low rank representation (LRR) [48] and latent low rank representation (LatLRR) [49] are selected as the comparing algorithms.

2) PARAMETER SETTING

About the parameter setting, we report the matrix factorization based methods (except for Kmeans, LRR and LatLRR) with the number of basis vectors equal to the number of clusters. For LRR and LatLRR, the ranks of adjacency graph (i.e., the number of subspaces) are also set as the number of clusters. For GNMF, NLCF and GRNMFSC, the maximum iterations are both set as 500 and the sparse regularization parameter λ of GRNMFSC is set as 0.01. The parameters of SCC and LapSC are configured according to [46]. For UDML_SC, the dimensions of images are reduced to 200 in DR process on COIL20 and E-YaleB datasets. Consequently, to construct over-complete dictionary for each test (*atom number* ≥ 200), we select 80% number of samples to be the atom numbers of the over-complete dictionary in dictionary learning phase. Note also that, there are three parameters in UDML_SC algorithm: the number of nearest neighbors k^m in graph G^m, the balance parameter ρ and the threshold τ to divide the MIP and LIP. We empirically set $k^m = 5$, $\rho = 0.3$ and $\tau = 0.7$.

3) CLUSTERING EVALUATION METRICS

The clustering result is evaluated by comparing the obtained label of each sample with the label provided by the dataset. Two metrics, the accuracy (AC) and the normalized mutual information metric (NMI) are used to measure the clustering performance. According to [50], the clustering accuracy (AC) is defined as follows:

$$AC = \frac{\Sigma_{i=1}^{N} \delta(c_i, map(l_i))}{N}, \tag{22}$$

where N is the total number of samples, c_i stands for the provided label, $map(l_i)$ is a mapping function that maps the obtained cluster label l_i to the equivalent label from the data corpus. $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise.

Let C denote the set of clusters obtained from the ground truth and $\widehat{C}$ obtained from our algorithm. Their mutual information metric $MI(C, \widehat{C})$ is defined according to [44], [45]:

$$MI(C, \widehat{C}) = \sum_{c_i \in C, \widehat{c_j} \in \widehat{C}} p(c_i, \widehat{c_j}).log \frac{p(c_i, \widehat{c_j})}{p(c_i)p(\widehat{c_j})}, \tag{23}$$

where $p(c_i)$ and $p(\widehat{c_j})$ denote the probabilities that a sample arbitrarily selected from the data set belongs to the clusters c_i and $\widehat{c_j}$, respectively. $p(c_i, \widehat{c_j})$ is the joint probability that the arbitrarily selected sample belongs to the clusters c_i and $\widehat{c_j}$ at the same time. In our experiment, we also use normalized mutual information (NMI) to evaluate clustering performance:

$$NMI(C, \widehat{C}) = \frac{MI(C, \widehat{C})}{max(H(C), H(\widehat{C}))}, \tag{24}$$

where $H(C)$ and $H(\widehat{C})$ are the entropies of C and $\widehat{C}$, respectively. NMI metric reflects the similarity of the distribution of C and $\widehat{C}$, if the two sets of clusters are identical, $NMI = 1$, otherwise NMI falls in between 0 and 1. The worst case is that the two sets are independent, then $NMI = 0$.

4) CLUSTERING RESULTS

Table 7 and Table 8 show the clustering results on the COIL20 and E-YaleB, respectively, the average clustering performances of different number of clusters are reported in two tables. We also highlight the best and comparable results in bold font and underline the second best ones.

As shown in Table 7 and Table 8, the proposed UDML_SC always result in the best performance in all the cases. On COIL20 dataset, the average clustering accuracies obtained by Kmeans, NMF, GNMF, NLCF, LapSC, LRR, LatLRR, GRNMFSC, SCC, and UDML_SC are 67.6%, 66.3%, 81.9%, 72.9%, 74.0%, 76.4%, 72.3% 84.0%, 81.1%, and 88.6%, respectively. Comparing with the second best method, that is, GRNMFSC, UDML_SC delivers 4.6% accuracy improvement. For mutual information, it can be seen that UDML_SC also achieves 3.4% improvement over GRNMF_SC. On E-YaleB dataset, the average clustering accuracies obtained by Kmeans, NMF, GNMF, NLCF, LapSC, LRR, LatLRR, GRNMFSC, SCC, and UDML_SC are 18.4%, 24.9%, 25.7%, 25.8%, 17.8%, 57.0%, 59.5%, 49.2%, 43.0%, and 63.8%, respectively. On this dataset, it is clear to find that UDML_SC makes greater advantages with respect to clustering accuracy. Specifically, UDML_SC improves 4.3% and 6.8% clustering accuracy compared to

TABLE 7. Clustering performance on COIL20 (Best; Second Best).

	AC(%)						NMI(%)					
Method	4	8	12	16	20	Avg.	4	8	12	16	20	Avg.
Kmeans	83.3	66.4	63.8	62.8	61.5	67.6	74.6	68.0	68.7	70.0	73.9	71.0
NMF	81.0	65.4	62.7	62.2	60.4	66.3	74.1	72.8	72.1	70.3	70.3	71.9
GNMF	89.9	83.0	80.8	79.0	77.2	81.9	84.1	86.7	84.7	83.2	85.4	84.8
NLCF	84.9	75.6	75.1	65.9	63.2	72.9	74.9	78.2	79.5	73.5	72.8	75.8
LapSC	82.9	72.0	72.2	71.8	71.1	74.0	74.0	76.3	79.5	80.5	79.8	78.0
LRR	85.0	82.1	73.7	70.9	70.3	76.4	80.1	82.1	78.8	76.9	78.7	79.3
LatLRR	83.4	73.8	72.4	67.8	64.2	72.3	75.3	74.8	76.2	75.0	74.0	75.1
GRNMFSC	90.6	85.4	83.0	81.9	79.0	84.0	**90.3**	86.4	87.3	86.3	87.2	87.5
SCC	91.1	81.1	79.4	77.9	76.2	81.1	88.4	79.4	85.8	84.9	83.1	84.3
UDML_SC	**93.8**	**90.1**	**90.9**	**83.5**	**84.7**	**88.6**	**90.8**	**89.9**	**92.9**	**89.7**	**91.2**	**90.9**

TABLE 8. Clustering performance on E-YaleB (Best; Second Best).

	AC(%)						NMI(%)					
Method	4	8	12	16	20	Avg.	4	8	12	16	20	Avg.
Kmeans	27.9	22.7	16.0	13.2	12.1	18.4	5.5	8.7	8.4	8.4	9.8	8.2
NMF	28.5	26.6	23.4	23.7	22.6	24.9	10.0	16.4	20.6	25.7	27.3	20.0
GNMF	31.2	26.7	24.0	24.0	22.4	25.7	11.9	17.4	19.7	26.0	27.8	20.6
NLCF	30.6	25.9	25.7	23.0	23.6	25.8	12.3	15.1	22.8	24.0	27.6	20.4
LapSC	28.1	18.4	15.3	14.0	13.1	17.8	4.2	5.9	7.6	8.3	14.0	8.0
LRR	64.6	55.6	57.2	54.6	53.0	57.0	49.7	54.3	58.1	56.8	55.6	54.9
LatLRR	62.2	59.2	60.6	**59.4**	55.9	59.5	48.2	59.4	61.9	63.0	60.1	58.5
GRNMFSC	59.9	58.1	43.6	43.7	40.5	49.2	45.7	57.9	48.6	51.1	50.8	50.8
SCC	49.1	47.6	42.4	39.8	36.2	43.0	29.4	42.1	44.1	44.3	44.2	40.8
UDML_SC	**72.0**	**65.6**	**63.4**	**60.1**	**58.0**	**63.8**	**64.1**	**62.2**	**63.3**	**64.7**	**64.2**	**63.7**

LatLRR and LRR, respectively. Moreover, UDML_SC even boosts 14.6% clustering accuracy compared to GRNMFSC. The significant improvement could be explained by that UDML_SC aims to separate the noise or trivial structure from original image and seek an efficient subspace representation (i.e., feature), which considers the intrinsic structure as well as the importance of different image components; consequently, by simply using k-means on the low-dimensional feature, UDML_SC can achieve impressive clustering performance even when the target image data are contaminated by illumination or shadow. In addition, LatLRR and LRR are both multi-subspace learning methods with error correction, so they perform especially well on this face dataset. In addition, GRNMFSC obtains better clustering results than SCC and GNMF, while LapSC and K-means perform the worst amidst all comparing methods.

5) SUMMARY

- UDML_SC achieves the best clustering results in both datasets compared with the state-of-the-art LatLRR, LRR and other comparing clustering methods.
- UDML_SC shows high resistance to images that are contaminated with severe variations of illumination or shadow.

IV. CONCLUDING REMARKS

In this paper, we have proposed a new subspace learning method, namely discriminant manifold learning via sparse coding (DML_SC). DML_SC aims to decompose the original image into a more important part (MIP) and a less important part (LIP), and learn a desired discriminative subspace where the intrinsic structure and energy of the MIP are preserved, while the energy of the LIP is simultaneously suppressed. Hence, DML_SC can exploit different contributions of different image components for robust feature extraction. Experimental results on image recognition and clustering tasks have demonstrated that DML_SC performs better than classical subspace learning methods and state-of-the-art sparse coding and dictionary learning methods.

It is worth mentioning that, since the DML_SC algorithm is directly applied on the original pixel intensity, the accuracy of image recognition and clustering in our experiments has still a room to make an improvement towards the practical pattern classification and clustering applications. For example, one feasible way is to replace the input features by local features, such as Gabor, LBP, super-pixel [51] or even latest deep features [52]. We will leave it as our future work.

ACKNOWLEDGMENT

The authors would like to express our sincere gratitude to all the reviewers and Dr. Jiawei Li for their constructive and valuable comments. The preliminary work was done in Dalian University of Technology, China, and then further extended and completed in Hong Kong Baptist University. *Meng Pang and Binghui Wang contribute equally to this work.*

REFERENCES

[1] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1991, pp. 586–591.

[2] S. Nikitidis, A. Tefas, and I. Pitas, "Maximum margin projection subspace learning for visual data analysis," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4413–4425, Oct. 2014.
[3] W. Jin, R. Liu, Z. Su, C. Zhang, and S. Bai, "Robust visual tracking using latent subspace projection pursuit," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.
[4] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.
[5] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.
[6] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[7] 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.
[8] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jul. 2005, pp. 846–853.
[9] X. Zeng and S. Luo, "A supervised subspace learning algorithm: Supervised neighborhood preserving embedding," in *Proc. Int. Conf. Adv. Data Mining Appl.*, 2007, pp. 81–88.
[10] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[11] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2007, pp. 708–713.
[12] Y. Liu, Y. Liu, and K. C. Chan, "Supervised manifold learning for image and video classification," in *Proc. 18th ACM Int. Conf. Multimedia (ACM MM)*, 2010, pp. 859–862.
[13] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao, "Maximal linear embedding for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1776–1792, Sep. 2011.
[14] B.-H. Wang, C. Lin, X.-F. Zhao, and Z.-M. Lu, "Neighbourhood sensitive preserving embedding for pattern classification," *IET Image Process.*, vol. 8, no. 8, pp. 489–497, Aug. 2014.
[15] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition," *Pattern Recognit.*, vol. 40, no. 1, pp. 339–342, 2007.
[16] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 499–506.
[17] H. Zhang, Q. M. J. Wu, T. W. S. Chow, and M. Zhao, "A two-dimensional neighborhood preserving projection for appearance-based face recognition," *Pattern Recognit.*, vol. 45, no. 5, pp. 1866–1876, 2012.
[18] Y. Guo *et al.*, "Tensor manifold discriminant projections for acceleration-based human activity recognition," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 1977–1987, Oct. 2016.
[19] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.
[20] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale $\ell 1$-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.
[21] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[22] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, no. 1, pp. 490–530, May 2015.
[23] 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, Feb. 2009.
[24] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.
[25] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1601–1604.
[26] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.
[27] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1864–1870, Sep. 2012.
[28] L. Zhang, P. Zhu, Q. Hu, and D. Zhang, "A linear subspace learning approach via sparse coding," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 755–761.
[29] X. Tang, G.-C. Feng, X.-X. Li, and J.-X. Cai, "Learning low-rank class-specific dictionary and sparse intra-class variant dictionary for face recognition," *PloS one*, vol. 10, no. 11, p. e0142403, 2015.
[30] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, May 2011.
[31] M. Pang, B. Wang, X. Fan, and C. Lin, "Discriminant manifold learning via sparse coding for image analysis," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, 2016, pp. 244–255.
[32] H. R. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Feb. 2006.
[33] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2002, pp. 46–51.
[34] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[35] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[36] J. Lu and Y. P. Tan, "Regularized locality preserving projections and its extensions for face recognition," *IEEE Trans. Syst., Man, B, Cybern.*, vol. 40, no. 3, pp. 958–963, Jun. 2010.
[37] Y. Xu, X. Li, J. Yang, Z. Lai, and D. Zhang, "Integrating conventional and inverse representation for face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1738–1746, Oct. 2014.
[38] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Tech. Rep. CUCS-005-96, 1996.
[39] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.
[40] Q. Gao, Y. Huang, H. Zhang, X. Hong, K. Li, and Y. Wang, "Discriminative sparsity preserving projections for image recognition," *Pattern Recognit.*, vol. 48, no. 8, pp. 2543–2553, Aug. 2015.
[41] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.
[42] J. H. Shah, M. Sharif, M. Raza, and A. Azeem, "A survey: Linear and non-linear PCA based face recognition techniques," *Int. Arab J. Inf. Technol.*, vol. 10, no. 6, pp. 536–545, 2013.
[43] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[44] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[45] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 969–979, Mar. 2013.
[46] D. Cai, H. Bao, and X. He, "Sparse concept coding for visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2011, pp. 2905–2910.
[47] C. Lin and M. Pang, "Graph regularized nonnegative matrix factorization with sparse coding," *Math. Problems Eng.*, vol. 2015, Feb. 2015, Art. no. 239589.
[48] 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, Jan. 2013.
[49] G. Liu and S. Yan, "Latent low-rank representation," in *Low-Rank and Sparse Modeling for Visual Analysis*. Springer, 2014, pp. 23–38.
[50] H. Jia, Y.-M. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1065–1079, May 2016.
[51] H. E. Tasli, R. Sicre, and T. Gevers, "Superpixel based mid-level image description for image recognition," *J. Vis. Commun. Image Represent.*, vol. 33, pp. 301–308, Nov. 2015.

[52] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2015, pp. 689–692.

MENG PANG received the B.Sc. degree in embedded engineering and the M.Sc. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include image processing, pattern recognition, and data mining.

BINGHUI WANG (S'16) received the B.Sc. degree in network engineering and the M.Sc. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Iowa State University, Ames, IA, USA. His research interests include machine learning, big data mining, data-driven security and privacy, and adversarial machine learning.

YIU-MING CHEUNG (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, in 2000. He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, information security, image and video processing, and pattern recognition. He is also a Senior Member of the Association for Computing Machinery. He is the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section and the Vice Chair of Technical Committee on Intelligent Informatics of the IEEE Computer Society.

CHUANG LIN (M'14) received the M.Sc. and Ph.D. degrees in signal processing from the Harbin Institute of Technology, Harbin, China, in 2004 and 2008, respectively. He is currently an Associate Professor with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Research Center for Neural Engineering, Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences. His research interests include biomedical signal processing, pattern recognition, and machine learning.