Topology Preserving Dictionary Learning for Pattern Classification

Jun Guo[†], Yanqing Guo[*†], Bo Wang[†], Xiangwei Kong[†], and Ran He[*‡§]
[†]School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[‡]The Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[§]CAS Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Beijing 100190, China
Email: guoyq@dlut.edu.cn, rhe@nlpr.ia.ac.cn

***Abstract*—In recent years, dictionary learning (DL) has shown significant potential in various classification tasks. However, most of previous works aim to learn a synthesis dictionary. The other major category of DL-analysis dictionary learning has not been fully exploited yet. This paper proposes a novel DL method, named Topology Preserving Dictionary Learning (TPDL). First, we propose a triplet-constraint-based topology preserving loss function to capture the underlying local topological structures of data in a supervised manner. Second, a sparse-label-matrix-based function is integrated into the basic analysis model to improve discriminative ability. Third, Huber M-estimator is employed as a robust metric to handle the errors (*e.g.*, outliers and noise) that possibly exist in data. Then, an alternating optimization algorithm is developed based on half-quadratic minimization and alternate search strategy. Closed-form solutions in each alternating optimization stage speed up the minimization process. Experiments on four commonly used datasets show that our proposed TPDL achieves competitive performance in contrast to state-of-the-art DL methods.**

I. Introduction

Sparse representation (SR) has shown significant potential in computer vision [1]. The success of SR pushes forward the research of dictionary learning (DL). In sparse representation, a dictionary learned from data is often better than a set of predefined bases (*e.g.*, wavelets) for many pattern classification tasks. In DL, one popular line of research aims to learn a synthesis dictionary with some specific regularizations, such as transform-invariance of dictionary [2], structured incoherence of dictionary [3], joint dictionary learning and subspace clustering [4], Fisher discrimination on both dictionary and codes [5], discriminative locality of codes [6], label consistency of sparse codes [7]. Synthesis dictionary learning is widely used but it is very time-consuming. Hence, as its dual model, analysis dictionary learning has drawn much attention [8].

Analysis model aims to learn a proper transformation instead of employing off-the-shelf transformations like DWT, DCT, *etc*. In this model, the presentation of signal is also sparse [9]. In recent years, some analysis dictionary learning (ADL) methods have been developed. Ravishankar and Bresler [10] proposed well-conditioned square transforms and applied this concept to image denoising. Shekhar *et al.* [9] enhanced the method in [10] by imposing a full-rank constraint. Rubinstein and Elad [11] utilized hard thresholding for the analysis codes' sparsity, meanwhile, they employed synthesis dictionary for sparse construction. Gu *et al.* [12] learned analysis and synthesis class-specific dictionary pairs for pattern classification tasks.

These previous works in analysis dictionary learning benefit from the simpler optimization in training stage and the higher speed in testing stage. However, analysis dictionary learning has not been fully exploited yet for pattern classification tasks. Inspired by the previous significant works in synthesis dictionary learning, we further integrate discriminative and structure preserving characters into analysis dictionary learning scheme, which is effective for classification.

Main contributions are in the following aspects:

- To capture the underlying discriminative topological structures of data, we employ triplet constraints rather than conventional pairwise/doublet constraints, resulting in a novel topology preserving loss function. This topological loss function can preserve not only the similarity relationship, but also the neighborhood ranking information in a supervised manner.
- To learn discriminative sparse codes instead of merely reconstructing data, we explicitly introduce discrimination into analysis dictionary learning by utilizing a sparse label matrix. Then, the learned analysis dictionary could exploit the discriminability of data.
- We utilize Huber M-estimator as a robust metric to handle the errors (*e.g.*, outliers and noise) that possibly exist in data. Consequently, we develop an alternating optimization algorithm based on half-quadratic minimization and alternate search strategy. Experiments on four databases show that our proposed method outperforms state-of-the-art DL methods.

The rest of this paper is organized as follows. Section II provides a compact review of synthesis dictionary learning and analysis dictionary learning. In Section III, the proposed Topology Preserving Dictionary Learning (TPDL) method is introduced. Section IV is the optimization and classification procedure. Extensive experiments are described in Section V. In Section VI, we conclude this paper as well as discuss future works.

978-1-5090-0620-5/16/$31.00

II. Basics of Dictionary Learning

Let $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{R}^{r_1 \times N}$ be the original data set. The core idea of dictionary learning is to obtain an optimized dictionary that provides an effective representation for each sample $\mathbf{y}_i \in \mathbb{R}^{r_1}$. $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{r_2 \times N}$ denotes $\mathbf{Y}$'s corresponding coding coefficient matrix over the learned dictionary.

Synthesis dictionary learning: Most existing dictionary learning methods aim to learn a synthesis dictionary $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_{r_2}] \in \mathbb{R}^{r_1 \times r_2}$ by solving

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \|\mathbf{X}\|_p + \lambda_2 \rho(\mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{H}) \quad (1)$$

where λ_1 and λ_2 are regularization parameters, $\mathbf{H}$ is the label matrix of $\mathbf{Y}$. The term $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ is the reconstruction error. $\|\mathbf{X}\|_p$ is the l_p-norm regularizer on $\mathbf{X}$. $\rho(\mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{H})$ stands for some specific functions, which can propagate the geometrical structures of $\mathbf{Y}$ to $\mathbf{X}$ or promote the discriminability of $\mathbf{D}$ and $\mathbf{X}$.

Analysis dictionary learning: As a dual analysis viewpoint of the commonly used synthesis dictionary learning, analysis dictionary learning gives an intuitive explanation like feature transformation (*e.g.*, FFT). It aims to learn an analysis dictionary $\mathbf{\Omega} \in \mathbb{R}^{r_2 \times r_1}$ by solving

$$\begin{aligned} \min_{\mathbf{\Omega},\mathbf{X}} \quad & \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \\ s.t. \quad & \|\mathbf{X}\|_0 \leq T_0, \\ & \|\omega_i\|_2 = 1,\ i = 1, \cdots, r_2 \end{aligned} \quad (2)$$

where T_0 is a sparsity constraint factor and ω_i is the ith row of $\mathbf{\Omega}$. Here, the row-wise norm of the analysis dictionary is constrained to be unity for a stable solution.

III. The Proposed Method

A. Motivation

The kNN classifier is a classical classification method, which can perform exceptionally well without training effort. However, it critically relies on the topology property [13] as well as the discriminability of data. In many cases, high-dimensional data points from the same category often exhibit degenerate structures [14] and lie on low-dimensional subspaces or manifolds. Hence, it is significant to simultaneously exploit the underlying local topological structures and discriminability of data when applying kNN classifier to dictionary-learning-based classification tasks. Consequently, we propose a Topology Preserving Dictionary Learning (TPDL) method, which combines both local topological structures and discriminative information of the original data. The learned analysis dictionary can map the original samples into a new space where the same-category neighbors are orderly preserved and neighbors with different labels are repelled.

B. Topology Preserving Loss Function

Topology property is a description for the local topological structures of data. It emphasizes not only the neighborhood relationship between data points but also the ranking information of each sample's neighbors. The relative neighborhood proximities play an important role in kNN classification [13].

The effectiveness of triplet constraints [15] has been verified in hashing [16], metric learning [17], and deep learning based computer vision (CV) applications [18], [19]. Inspired by these recent advances, we propose a novel topology preserving loss function, which is different from similarity preserving loss function via pairwise/doublet constraints.

Definition 1. Given a triplet $(\mathbf{y}_i, \mathbf{y}_u, \mathbf{y}_v)$ comprised of sample $\mathbf{y}_i$ and its neighbors $\mathbf{y}_u$ and $\mathbf{y}_v$, their corresponding coding vectors also form a triplet $(\mathbf{x}_i, \mathbf{x}_u, \mathbf{x}_v)$. Let g_{i*} and h_{i*} denote the pairwise distance of $(\mathbf{y}_i, \mathbf{y}_*)$ and $(\mathbf{x}_i, \mathbf{x}_*)$, respectively. Subscript $*$ presents a placeholder for u or v. Then, the coding process is called *topology preserving* when the condition holds: if $g_{iu} \leq g_{iv}$, then $h_{iu} \leq h_{iv}$. □

Based on Definition 1 and Rearrangement Inequality[1], determining appropriate coding vectors is identical to optimize the following topology preserving loss function

$$\max_{\mathbf{X}} \sum_{i=1}^{N} \sum_{u=1}^{N} \sum_{v=1}^{N} (g_{iu} - g_{iv})(h_{iu} - h_{iv}). \quad (3)$$

Let $\mathbf{G}_i$ be an antisymmetric matrix with the $(u,v)^{th}$ element equals $(g_{iu} - g_{iv})$. The commonly used Euclidean distance is employed to compute each pairwise distance. Then (3) takes the following form:

$$\max_{\mathbf{X}} \sum_{i=1}^{N} \sum_{u=1}^{N} \sum_{v=1}^{N} \mathbf{G}_i(u,v) \left(\|\mathbf{x}_i - \mathbf{x}_u\|_2^2 - \|\mathbf{x}_i - \mathbf{x}_v\|_2^2 \right). \quad (4)$$

However, (4) is an unsupervised type. In consideration of each sample's label, we further propose a supervised-type loss function by replacing $\mathbf{G}_i$ with $\mathbf{G'}_i$ in (4). The $(u,v)^{th}$ element of $\mathbf{G'}_i$ is defined as

$$\mathbf{G'}_i(u,v) \triangleq \begin{cases} -\mathbf{G}_i(u,v)\, sgn[\mathbf{G}_i(u,v)] & , (i,u,v) \in \mathcal{O} \\ \mathbf{G}_i(u,v)\, sgn[\mathbf{G}_i(u,v)] & , (i,v,u) \in \mathcal{O} \\ \mathbf{G}_i(u,v) & , otherwise \end{cases} \quad (5)$$

where $sgn(z) = \begin{cases} -1 & , z < 0 \\ 0 & , z = 0 \\ +1 & , z > 0 \end{cases}$ is the sign function and $\mathcal{O} = \{(a,b,c)\}$ is a set of descriptions. If $(a,b,c) \in \mathcal{O}$, it means that the two samples $\mathbf{y}_a$ and $\mathbf{y}_b$ share the same label which is quite different from sample $\mathbf{y}_c$'s. It is obvious that $\mathbf{G'}_i$ is also an antisymmetric matrix. We obtain the supervised topology preserving loss function

$$\max_{\mathbf{X}} \sum_{i=1}^{N} \sum_{u=1}^{N} \sum_{v=1}^{N} \mathbf{G'}_i(u,v) \left(\|\mathbf{x}_i - \mathbf{x}_u\|_2^2 - \|\mathbf{x}_i - \mathbf{x}_v\|_2^2 \right). \quad (6)$$

Proposition 1. *Let* $\mathbf{W} \in \mathbb{R}^{N \times N}$ *be a weighting matrix with the* $(i,j)^{th}$ *element equals* $\sum_{u=1}^{N} \mathbf{G'}_i(u,j)$. *Objective function (6) is equivalent to* $\min_{\mathbf{X}} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{W}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$.

[1]Online: http://en.wikipedia.org/wiki/Rearrangement_inequality.

Proof. Recall that $\mathbf{G'}_i$ is an antisymmetric matrix, so $\mathbf{G'}_i(u,v) = -\mathbf{G'}_i(v,u)$. Then (6) is equivalent to

$$\max_{\mathbf{X}} \left\{ \begin{array}{l} -\sum_{i=1}^{N}\sum_{u=1}^{N}\sum_{v=1}^{N} \mathbf{G'}_i(v,u) \|\mathbf{x}_i - \mathbf{x}_u\|_2^2 \\ -\sum_{i=1}^{N}\sum_{v=1}^{N}\sum_{u=1}^{N} \mathbf{G'}_i(u,v) \|\mathbf{x}_i - \mathbf{x}_v\|_2^2 \end{array} \right\}, \quad (7)$$

which can be reformulated as

$$\max_{\mathbf{X}} \left\{ \begin{array}{l} -\sum_{i=1}^{N}\sum_{u=1}^{N} \mathbf{W}_{iu} \|\mathbf{x}_i - \mathbf{x}_u\|_2^2 \\ -\sum_{i=1}^{N}\sum_{v=1}^{N} \mathbf{W}_{iv} \|\mathbf{x}_i - \mathbf{x}_v\|_2^2 \end{array} \right\}. \quad (8)$$

The ultimate form $\min_{\mathbf{X}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{W}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ can be easily derived from (8). □

Let $\mathbf{S}$ denote a diagonal matrix with the $(i,i)^{th}$ element equals $\sum_{j=1}^{N} \frac{\mathbf{W}_{ij}+\mathbf{W}_{ji}}{2}$. We can define the Laplacian matrix $\mathbf{L} \triangleq \mathbf{S} - \frac{\mathbf{W}+\mathbf{W}^T}{2}$. It can be derived that $\min_{\mathbf{X}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{W}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ is equivalent to $\min_{\mathbf{X}} Tr\left(\mathbf{X}\mathbf{L}\mathbf{X}^T\right)$. Considering the important role of the analysis dictionary $\mathbf{\Omega}$ in coding process, we substitute $\mathbf{X}$ in the topology preserving loss function $\min_{\mathbf{X}} Tr\left(\mathbf{X}\mathbf{L}\mathbf{X}^T\right)$ with $\mathbf{\Omega}\mathbf{Y}$ so that $\mathbf{\Omega}$ can be directly learned in an efficient way, which is demonstrated to be effective in our experiments.

To simultaneously preserve neighborhood ranking as well as similarity relationship with a supervised approach, we employ (9) to calculate $\mathbf{W}_{ij}$ for each $\mathbf{y}_i$

$$\mathbf{W}_{ij} = \left\{ \begin{array}{ll} \sum_{\mathbf{y}_u \in \mathcal{N}_i} \mathbf{G'}_i(u,j) & , \mathbf{y}_j \in \mathcal{N}_i \\ 0 & , otherwise \end{array} \right. \quad (9)$$

where $\mathcal{N}_i$ is a set consisting of the k nearest neighbors of $\mathbf{y}_i$. In practice, we normalize each non-zero $\mathbf{W}_{ij}$ to $[0,1]$ by row and utilize the normalized Laplacian matrix $\bar{\mathbf{L}} \triangleq \mathbf{I} - \mathbf{S}^{-1/2}\bar{\mathbf{W}}\mathbf{S}^{-1/2}$ instead of $\mathbf{L}$, where $\mathbf{I}$ is an identity matrix[2] and $\bar{\mathbf{W}} \triangleq \frac{\mathbf{W}+\mathbf{W}^T}{2}$.

C. The Proposed TPDL Method

The analysis dictionary learning framework in (2) focuses on well representing the original data, neglecting its application to pattern classification tasks. To introduce discrimination into analysis dictionary learning, we set r_2 to the total number of categories and employ a sufficiently sparse label matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_N] \in \mathbb{R}^{r_2 \times N}$ as target codes. Each column of the sparse matrix $\mathbf{H}$ is a label vector: $\mathbf{h}_i = [0,0,...,1,...,0,0]^T$, whose non-zero position indicates the category label of $\mathbf{y}_i$. For example, assuming $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_6]$, where $\mathbf{y}_1$ and $\mathbf{y}_2$ are from class 1, $\mathbf{y}_3$ and $\mathbf{y}_4$ are from class 2, $\mathbf{y}_5$ and $\mathbf{y}_6$ are from class 3, the label matrix $\mathbf{H}$ can be defined as

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

[2]In this paper, $\mathbf{I}$, $\mathbf{0}$ and $\mathbf{1}$ denote the identity, all zeros and all ones matrix with compatible sizes, respectively.

For analysis dictionary learning, each sparse code often has a larger length than the total number of classes. We skillfully employ the Kronecker product[3] to generate a high-dimensional $\mathbf{H}$. To put it simply, the Kronecker product of two matrices $\mathbf{A} \in \mathbb{R}^{a_1 \times a_2}$ and $\mathbf{B} \in \mathbb{R}^{b_1 \times b_2}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is the $a_1 b_1 \times a_2 b_2$ block matrix:

$$\mathbf{A} \otimes \mathbf{B} \triangleq \begin{bmatrix} \mathbf{A}_{11}\mathbf{B} & \mathbf{A}_{12}\mathbf{B} & \cdots & \mathbf{A}_{1a_2}\mathbf{B} \\ \mathbf{A}_{21}\mathbf{B} & \mathbf{A}_{22}\mathbf{B} & \cdots & \mathbf{A}_{2a_2}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{a_1 1}\mathbf{B} & \mathbf{A}_{a_1 2}\mathbf{B} & \cdots & \mathbf{A}_{a_1 a_2}\mathbf{B} \end{bmatrix}. \quad (10)$$

There are various methods to obtain high-dimensional $\mathbf{H}$. In this paper, we only employ the Kronecker product for simplicity. The original low-dimensional $\mathbf{H}$ is replaced by the Kronecker product of itself and an all ones vector with appropriate size. We further propose a novel discriminative term $\|\mathbf{X} - \mathbf{H}\|_F^2$ added to (2) to fully exploit the discrimination embedded in the training samples. According to [9], the unit norm condition in (2) can be replaced by adding a small penalty of $\|\mathbf{\Omega}\|_F^2$ to the cost term. The optimization problem of our method now takes the following form:

$$\begin{array}{ll} \min\limits_{\mathbf{\Omega},\mathbf{X}} & \left\{ \begin{array}{c} \|\mathbf{X} - \mathbf{\Omega}\mathbf{Y}\|_F^2 \\ +\alpha Tr\left(\mathbf{\Omega}\mathbf{Y}\mathbf{L}\mathbf{Y}^T\mathbf{\Omega}^T\right) \\ +\beta \|\mathbf{X} - \mathbf{H}\|_F^2 + \gamma \|\mathbf{\Omega}\|_F^2 \end{array} \right\} \\ s.t. & \|\mathbf{x}_i\|_0 \le T_0, \ \forall i \end{array} \quad (11)$$

where α, β and γ are scalar constants that control the relative importance of the corresponding terms.

D. Problem Reformulation for Robustness

M-estimators are widely used to improve robustness of learning algorithms in computer vision [20]. Some popular M-estimators include Welsch function, l_p function, l_1-l_2 function and Huber function. Welsch function has a close relationship with Correntropy Induced Metric (CIM) [21]. The robustness of CIM has been verified in principal component analysis [22], non-negative matrix factorization [23], synthesis dictionary learning [24], low-rank representation [25], subspace clustering [26], [27] and other applications [28], [29], [30]. Besides, l_p function and l_1-l_2 function have been employed for various algorithms, *e.g.*, information theoretic subspace clustering [31] and robust sparse representation [32]. In this paper, we utilize Huber loss function $\mathcal{L}_H(z)$ to handle the errors (*e.g.*, outliers and noise) that possibly exist in data.

$$\mathcal{L}_H(z) = \left\{ \begin{array}{ll} \frac{z^2}{2} & , |z| \le \tau \\ \tau|z| - \frac{\tau^2}{2} & , |z| > \tau \end{array} \right. \quad (12)$$

Then, we obtain the final formulation of TPDL:

$$\begin{array}{ll} \min\limits_{\mathbf{\Omega},\mathbf{X}} & F = F_0 + \alpha F_1 + \beta F_2 + \gamma \|\mathbf{\Omega}\|_F^2 \\ s.t. & \|\mathbf{x}_i\|_0 \le T_0, \ \forall i \end{array} \quad (13)$$

[3]Online: https://en.wikipedia.org/wiki/Kronecker_product.

where

$$\begin{cases} F_0 = \sum_{i=1}^{N} \mathcal{L}_H\left(\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2\right) \\ F_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathcal{L}_H\left(\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\|_2\right) \\ F_2 = \sum_{i=1}^{N} \mathcal{L}_H\left(\|\mathbf{x}_i - \mathbf{h}_i\|_2\right) \end{cases}. \quad (14)$$

IV. Optimization and Classification

A. Half-quadratic Technique

Since problem (13) is not convex, it is difficult to directly optimize. Fortunately, we can employ the half-quadratic (HQ) technique to optimize it by alternately minimizing its augmented function $\hat{F}$. According to the HQ theory and the conjugate function theory [33], we have

Lemma 1. *Suppose that $\mathcal{F}(z)$ is a function which satisfies the conditions listed in [33], there exists a dual potential function $\varphi(\cdot)$, such that*

$$\mathcal{F}(z) = \inf_{r\in\mathbb{R}}\left\{rz^2 + \varphi(r)\right\} \quad (15)$$

where r is an auxiliary variable that is completely determined by the minimizer function $\delta(z)$ w.r.t. $\mathcal{F}(z)$.

As indicated in [32], $\delta(z) = \begin{cases} 1 & ,|z| \le \tau \\ \frac{\tau}{|z|} & ,|z| > \tau \end{cases}$ when $\mathcal{F}(z) = \mathcal{L}_H(z) = \begin{cases} \frac{z^2}{2} & ,|z| \le \tau \\ \tau|z| - \frac{\tau^2}{2} & ,|z| > \tau \end{cases}$. That is to say, the infimum of $\mathcal{F}(z)$ for a fixed z could be reached at $r = \delta(z)$.

According to Lemma 1, the augmented function $\hat{F}$ of (13) takes the following form

$$\min_{\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}} \hat{F} = \hat{F}_0 + \alpha\hat{F}_1 + \beta\hat{F}_2 + \gamma\|\mathbf{\Omega}\|_F^2 \quad s.t. \quad \|\mathbf{x}_i\|_0 \le T_0,\ \forall i \quad (16)$$

where

$$\begin{cases} \hat{F}_0 = \sum_{i=1}^{N}\left\{\mathbf{R}_{ii}\|\mathbf{x}_i - \mathbf{\Omega}\mathbf{y}_i\|_2^2 + \varphi_i(\mathbf{R}_{ii})\right\} \\ \hat{F}_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\mathbf{W}_{ij}\mathbf{P}_{ij}\|\mathbf{\Omega}\mathbf{y}_i - \mathbf{\Omega}\mathbf{y}_j\|_2^2 + \mathbf{W}_{ij}\phi_{ij}(\mathbf{P}_{ij})\right\} \\ \hat{F}_2 = \sum_{i=1}^{N}\left\{\mathbf{Q}_{ii}\|\mathbf{x}_i - \mathbf{h}_i\|_2^2 + \psi_i(\mathbf{Q}_{ii})\right\} \end{cases}. \quad (17)$$

The $N \times N$ matrices $\mathbf{R}$, $\mathbf{P}$ and $\mathbf{Q}$ store the auxiliary variables. Note that $\mathbf{R}$ and $\mathbf{Q}$ are diagonal. $\{\varphi_i\}_{i=1}^N$, $\{\psi_i\}_{i=1}^N$ and $\{\phi_{ij}\}_{i,j=1}^N$ are conjugate functions.

B. Optimization Procedure

Based on the HQ optimization theory, $\hat{F}(\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q})$ can be alternately minimized as follows:

1) Update the analysis dictionary and sparse codes.

$$\left(\mathbf{\Omega}^{t+1},\mathbf{X}^{t+1}\right) = \arg\min_{\mathbf{\Omega},\mathbf{X}} J = J_0 + \alpha J_1 + \beta J_2 + \gamma\|\mathbf{\Omega}\|_F^2 \quad s.t. \quad \|\mathbf{x}_i\|_0 \le T_0,\ \forall i \quad (18)$$

where

$$\begin{cases} J_0 = Tr\left((\mathbf{X} - \mathbf{\Omega}\mathbf{Y})\,\mathbf{R}^t(\mathbf{X} - \mathbf{\Omega}\mathbf{Y})^T\right) \\ J_1 = Tr\left(\mathbf{\Omega}\mathbf{Y}\mathbf{L}^{t+1}\mathbf{Y}^T\mathbf{\Omega}^T\right) \\ J_2 = Tr\left((\mathbf{X} - \mathbf{H})\,\mathbf{Q}^t(\mathbf{X} - \mathbf{H})^T\right) \end{cases}. \quad (19)$$

The weighting matrix is updated by $\mathbf{W}^{t+1} = \mathbf{P}^t \odot \mathbf{W}^t$ in the $(t+1)^{th}$ iteration. $\mathbf{L}^{t+1}$ is its corresponding Laplacian matrix. Alternate search strategy is employed to alternatively minimize (18) with respect to one while fixing the other one variable. To update $\mathbf{\Omega} \in \mathbb{R}^{r_2 \times r_1}$ with fixed $\mathbf{X}$, we solve

$$\min_{\mathbf{\Omega}} Tr\left\{\begin{array}{l} -2\mathbf{X}\mathbf{R}^t\mathbf{Y}^T\mathbf{\Omega}^T + \gamma\mathbf{\Omega}\mathbf{\Omega}^T \\ +\mathbf{\Omega}\mathbf{Y}\left(\mathbf{R}^t + \alpha\mathbf{L}^{t+1}\right)\mathbf{Y}^T\mathbf{\Omega}^T \end{array}\right\}, \quad (20)$$

The analytical solution is obtained by setting its first derivative to $\mathbf{0}$: $\mathbf{\Omega} = \mathbf{X}\mathbf{R}^t\mathbf{Y}^T\left[\mathbf{Y}\left(\mathbf{R}^t + \alpha\mathbf{L}^{t+1}\right)\mathbf{Y}^T + \gamma\mathbf{I}\right]^{-1}$. To update $\mathbf{X} \in \mathbb{R}^{r_2 \times N}$ with fixed $\mathbf{\Omega}$, we solve (21) for each column.

$$\min_{\mathbf{x}_i} \left\|\mathbf{x}_i - \frac{\mathbf{R}_{ii}^t\mathbf{\Omega}\mathbf{y}_i + \beta\mathbf{Q}_{ii}^t\mathbf{h}_i}{\mathbf{R}_{ii}^t + \beta\mathbf{Q}_{ii}^t}\right\|_2^2 \quad s.t. \quad \|\mathbf{x}_i\|_0 \le T_0 \quad (21)$$

The analytical solution is computed by hard thresholding: setting the smallest $(r_2 - T_0)$ elements (in magnitude) of $\frac{\mathbf{R}_{ii}^t\mathbf{\Omega}\mathbf{y}_i + \beta\mathbf{Q}_{ii}^t\mathbf{h}_i}{\mathbf{R}_{ii}^t + \beta\mathbf{Q}_{ii}^t}$ to 0. The update for $\mathbf{X}$ could be efficiently implemented in parallel.

2) Update auxiliary variables.

$$\mathbf{R}_{ii}^{t+1} = \delta\left(\left\|\mathbf{x}_i^{t+1} - \mathbf{\Omega}^{t+1}\mathbf{y}_i\right\|_2\right) \quad (22)$$

$$\mathbf{P}_{ij}^{t+1} = \delta\left(\left\|\mathbf{\Omega}^{t+1}\mathbf{y}_i - \mathbf{\Omega}^{t+1}\mathbf{y}_j\right\|_2\right) \quad (23)$$

$$\mathbf{Q}_{ii}^{t+1} = \delta\left(\left\|\mathbf{x}_i^{t+1} - \mathbf{h}_i\right\|_2\right) \quad (24)$$

As summarized in Algorithm 1, the two steps are alternatively minimized until convergence. It is empirically found that our method converges rapidly (please refer to Section IV-D for theoretical analysis). In most of our experiments, TPDL will converge in less than 15 iterations.

Algorithm 1 Topology Preserving Dictionary Learning (T-PDL)

Input:
Training data $\mathbf{Y}$ and the corresponding matrix $\mathbf{H}$;
Number of each data point's nearest neighbors k;
Regularization parameters α, β and γ;
The threshold parameter for Huber M-estimator τ.

Output:
The analysis dictionary $\mathbf{\Omega}$.

Set $\mathbf{X}^{(0)} = \mathbf{H}$, $\mathbf{R}^{(0)} = \mathbf{I}$, $\mathbf{P}^{(0)} = \mathbf{1}$, $\mathbf{Q}^{(0)} = \mathbf{I}$ for initialization. Compute the weighting matrix $\mathbf{W}^{(0)}$, $t = 0$;
while not convergence **do**
  $t \Leftarrow t + 1$;
  Update the weighting matrix $\mathbf{W}^{(t)}$ and compute its Laplacian matrix $\mathbf{L}^{(t)}$;
  Update $\mathbf{\Omega}^{(t)}$ by solving (20);
  Update $\mathbf{X}^{(t)}$ by solving (21) in parallel;
  Update $\mathbf{R}^{(t)}$, $\mathbf{P}^{(t)}$ and $\mathbf{Q}^{(t)}$ via (22), (23) and (24);
end while

C. Classification

Given training data and the corresponding matrix $\mathbf{H}$, an optimized dictionary could be learned by Algorithm 1. Then, we could code both training and testing samples by employing hard thresholding to solve

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & \|\mathbf{x}-\mathbf{\Omega}\mathbf{y}\|_2^2 \\ s.t. & \|\mathbf{x}\|_0 \leq T_0. \end{array} \tag{25}$$

Finally, these coding vectors are treated as new features and we employ kNN classifier to perform classification.

D. Discussion

Sparsity constraint: Actually, the role of sparsity in pattern classification is still an open problem. Some investigators argued that sparsity may not be relevant for classification tasks [12], [34], and our paper are supportive to this argument. We have observed in the experiment that our TPDL has very competitive classification performance when the sparsity constraint is omitted. This kind of model has less complexity than other sparsity-based DL methods.

Convergence analysis: According to Lemma 1, when fixing $\mathbf{\Omega}$ and $\mathbf{X}$, the following equation holds:

$$F\left(\mathbf{\Omega},\mathbf{X}\right) = \inf_{\mathbf{R},\mathbf{P},\mathbf{Q}} \hat{F}\left(\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}\right). \tag{26}$$

It follows that

$$\min_{\mathbf{\Omega},\mathbf{X}} F\left(\mathbf{\Omega},\mathbf{X}\right) = \min_{\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}} \hat{F}\left(\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}\right). \tag{27}$$

Therefore, minimizing $F\left(\mathbf{\Omega},\mathbf{X}\right)$ is equivalent to minimizing its augmented function $\hat{F}\left(\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}\right)$ on the enlarged domain. According to the properties of half-quadratic minimization [33] and the commonly used alternate search strategy, we have

$$\begin{array}{ll} & \hat{F}\left(\mathbf{\Omega}^{t+1},\mathbf{X}^{t+1},\mathbf{R}^{t+1},\mathbf{P}^{t+1},\mathbf{Q}^{t+1}\right) \\ \leq & \hat{F}\left(\mathbf{\Omega}^{t+1},\mathbf{X}^{t+1},\mathbf{R}^{t},\mathbf{P}^{t},\mathbf{Q}^{t}\right) \\ \leq & \hat{F}\left(\mathbf{\Omega}^{t},\mathbf{X}^{t},\mathbf{R}^{t},\mathbf{P}^{t},\mathbf{Q}^{t}\right) \end{array}. \tag{28}$$

The objective function is non-increasing at each alternative minimization step.

What's more, according to [20], [35], the objective function $F\left(\mathbf{\Omega},\mathbf{X}\right)$ in (13) is bounded below, and thus by (27) we obtain that $\hat{F}\left(\mathbf{\Omega},\mathbf{X},\mathbf{R},\mathbf{P},\mathbf{Q}\right)$ is also bounded. Consequently, we can conclude that $\hat{F}\left(\mathbf{\Omega}^t,\mathbf{X}^t,\mathbf{R}^t,\mathbf{P}^t,\mathbf{Q}^t\right)$ decreases step by step until Algorithm 1 converges.

V. Experiments

We demonstrate the performance of our TPDL on four typical visual classification databases: Two face datasets (AR [36] and Extended YaleB [37]), and one object categorization dataset (Caltech 101 [38]), and one action recognition dataset (UCF 50 action [39]). These databases are widely used in previous works to evaluate dictionary learning based classification methods.

We compare our TPDL with the following methods: the baseline Support Vector Machine (SVM) and Nearest Subspace Classifier (NSC), Sparse-Representation-based Classifier (SRC) [1] and Collaborative-Representation-based Classifier (CRC) [34], and three state-of-the-art DL methods: Dictionary Learning with Structured Incoherence (DLSI) [3], Fisher Discrimination Dictionary Learning (FDDL) [5], Label Consistent K-SVD (LC-KSVD) [7].

Fig. 1. Some sample objects from the AR database [36].

Fig. 2. Some sample objects from the Extended YaleB database [37].

We follow the experimental settings in [7] for all the above methods for fair comparison. We use the features of these datasets provided by Jiang[4] and Corso[5]. We empirically set $\beta = 10$ in this paper to force the discriminative codes $\mathbf{X}$ to approximate the sparse $\mathbf{H}$. The experimental results are insensitive to $\beta \in [10, 15]$. The Huber function's threshold parameter is estimated according to [32], in which τ is regarded as a function of median. The sparsity level T_0 depends on the length of all ones vector employed in the Kronecker product. For each database, we set T_0 to a proper positive integer so that $\mathbf{h}_i$'s length is a little bigger than $\mathbf{y}_i$'s. The other parameters of our TPDL are listed in Table I, which have been tuned on each database by cross validation.

The AR face database [36] includes more than 4,000 color face images from 126 persons. Each subject has 26 face images taken during two sessions. The characteristic of the AR face database is that it contains frontal views of faces with different occlusion conditions (sunglasses and scarves), facial expressions and lighting conditions. All the images are cropped and scaled to 165×120. Then, they are projected onto 540-dimensional vectors with a randomly generated matrix to obtain random-face features. A subset of 2,600 images from 50 females and 50 males is obtained. We randomly select 20 images each person for training and the other 6 images for testing. We report the average of 15 such random splits. As shown in Table II, our TPDL method gives a better result.

[4]Online: http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html.
[5]Online: http://www.cse.buffalo.edu/~jcorso/r/actionbank.

Fig. 3. Some sample objects from the Caltech 101 database [38].

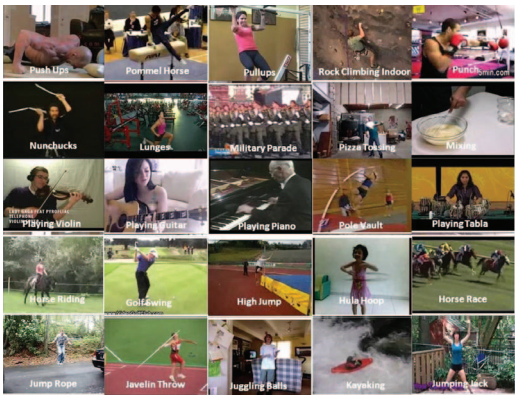

Fig. 4. Some sample objects from the UCF 50 database [39].

The Extended YaleB database [37] contains 2,414 face images of 38 people under 64 illumination conditions. This database is challenging for plentiful expressions and varying illumination conditions. All the original images are cropped to 192×168 pixels. Similar to AR's features, the features used here are 504-dimensional random-face features. We randomly select 32 images each human subject for training and the remaining images for testing. We repeated 15 times such a sampling process and calculated their average as the final classification accuracy. It can be seen from Table II that our proposed TPDL method obtains an improvement over other algorithms.

The Caltech 101 database [38] is comprised of 9,144 images from a background class and 101 common object classes. The number of samples in each class ranges from 31 to 800. We use the standard Bag-of-Visual-Words (BoVW) + Spatial Pyramid Matching (SPM) framework [40] for feature extraction. First, we extract dense SIFT descriptors from patches of size 16×16 which are sampled using a grid with a 6-pixel step size. Then, we compute the SPM features with 1×1, 2×2, and 4×4 subregions. To build fair comparison, the Vector Quantization (VQ) based coding method is employed to extract the mid-level features and the high-dimensional pooled features are obtained by utilizing the standard max pooling approach. Finally, the 21,504 dimensional features are reduced to 3,000 dimensions via Principal Component Analysis (PCA). For each category, 30 images are randomly selected for training and the rest for testing. Experimental results are summarized in Table II, which is the average of 15 random splits. Our TPDL approach achieves obviously higher accuracy than its competitors.

TABLE I
MAJOR PARAMETERS, DETERMINED BY CROSS-VALIDATION.

	AR	Extended Yale B	Caltech 101	UCF 50
k	5	5	9	7
α	0.001	0.001	0.010	0.010
γ	0.1	0.1	3.0	0.1

TABLE II
CLASSIFICATION ACCURACIES (%) ON FOUR DATASETS.

	AR	Extended Yale B	Caltech 101	UCF 50
SVM	96.5	95.6	64.6	72.3
NSC	92.0	94.7	70.1	67.5
SRC	97.5	96.5	70.7	75.0
CRC	98.0	97.0	68.2	75.6
DLSI	97.5	97.0	73.1	75.4
FDDL	97.5	96.7	73.2	76.5
LC-KSVD	97.8	96.7	73.6	70.1
TPDL	**98.1**	**97.2**	**73.9**	**76.8**

The UCF 50 [39] is an action recognition dataset with 50 action categories, containing 6,680 realistic human action videos from YouTube. For all the categories, the videos are divided into 25 groups and each group contains over four action clips. The action clips within the same group share some common characteristics, such as similar background, similar viewpoint, the same person, and so on. We utilize the action bank features [41] and five-fold data splitting for evaluation, where four folds are selected for training and the remaining one fold for testing. We also utilize PCA to reduce the high-dimensional features to 5,000 dimensions. The experiments are repeated 15 times with different random splits of training and testing images. Reliable results of different methods are reported in Table II. As expected, our proposed TPDL method achieves the best performance.

VI. CONCLUSION

In this paper, we proposed a novel analysis dictionary learning method by fully exploiting the topology property and discriminability of data. Based on triplet constraints, a novel topology preserving loss function has been developed, which simultaneously preserves neighborhood ranking as well as similarity relationship in a supervised manner. To make ADL suitable for pattern classification tasks, a sparse label matrix has been introduced for discrimination and Huber M-estimator has been employed as a robust metric. Based on half-quadratic minimization and alternate search strategy, we developed an iterative algorithm to speed up the analysis dictionary learning process. Experimental results on four pattern classification databases show the effectiveness of our proposed method. Future researches will focus on exploiting new topological criterions of similarity to measure the topology property of data.

ACKNOWLEDGMENT

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB02000000), the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the National Natural Science Foundation of China (Grant No. 61175003, 61172109, 61402079, 61473289, 61502076), the Youth Innovation Promotion Association CAS (Grant No. 2015190), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

REFERENCES

[1] 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. 1, 5

[2] S. Zhang, M. Zhang, R. He, and Z. Sun, "Transform-invariant dictionary learning for face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 348–352. 1

[3] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, San Francisco, CA, Jun. 2010, pp. 3501–3508. 1, 5

[4] G. Zhang, R. He, and L. S. Davis, "Jointly learning dictionaries and subspace structure for video-based face recognition," in *Proc. Asian Conf. Comput. Vision (ACCV)*, Singapore, Nov. 2014, pp. 348–352. 1

[5] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 543–550. 1, 5

[6] J. Guo, Y. Guo, Y. Li, B. Wang, and M. Li, "Locality sensitive discriminative dictionary learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, Canada, Sep. 2015, pp. 1558–1562. 1

[7] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013. 1, 5

[8] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010. 1

[9] S. Shekhar, V. M. Patel, and R. Chellappa, "Analysis sparse coding models for image-based classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 5207–5211. 1, 3

[10] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1072–1086, Mar. 2013. 1

[11] R. Rubinstein and M. Elad, "Dictionary learning for analysis-synthesis thresholding," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5962–5972, Nov. 2014. 1

[12] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *Proc. Conf. Advances Neural Inform. Process. Syst. (NIPS)*, Montreal, Canada, Dec. 2014, pp. 793–801. 1, 5

[13] D. Luo, C. Ding, F. Nie, and H. Huang, "Cauchy graph embedding," in *Proc. Int. Conf. Mach. Learning (ICML)*, Bellevue, Washington, Jun. 2011, pp. 553–560. 2

[14] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010. 2

[15] A. Frome, Y. Singer, F. Sha, and J. Malik, "Learning globally-consistent local distance functions for shape-based image retrieval and classification," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2007, pp. 1–8. 2

[16] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2015, pp. 3270–3278. 2

[17] F. Wang, W. Zuo, L. Zhang, D. Meng, and D. Zhang, "A kernel classification framework for metric learning," *IEEE Trans. Neural Networks Learning Syst.*, vol. 26, no. 9, pp. 1950–1962, Sep. 2015. 2

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2015, pp. 815–823. 2

[19] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2015, pp. 1556–1564. 2

[20] R. He, T. Tan, and L. Wang, "Recovery of corrupted low-rank matrix by implicit regularizers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 770–783, Apr. 2014. 3, 5

[21] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007. 3

[22] R. He, B. Hu, W. Zheng, and X. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1485–1494, Jun. 2011. 3

[23] L. Du, X. Li, and Y. Shen, "Robust nonnegative matrix factorization via half-quadratic minimization," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Brussels, Belgium, Dec. 2012, pp. 201–210. 3

[24] P. Hao and S. Kamata, "Maximum correntropy criterion for discriminative dictionary learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, Australia, Sep. 2013, pp. 4325–4329. 3

[25] Y. Zhang, Z. Sun, R. He, and T. Tan, "Robust low-rank representation via correntropy," in *Proc. Asian Conf. Pattern Recognition (ACPR)*, Okinawa, Japan, Nov. 2013, pp. 461–465. 3

[26] ——, "Robust subspace clustering via half-quadratic minimization," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Sydney, Australia, Dec. 2013, pp. 3096–3103. 3

[27] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, "Correntropy induced l_2 graph for robust subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Sydney, Australia, Dec. 2013, pp. 1801–1808. 3

[28] B. Chen and J. C. Principe, "Maximum correntropy estimation is a smoothed map estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 491–494, Aug. 2012. 3

[29] B. Chen, L. Xing, J. Liang, N. Zheng, and J. C. Principe, "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion," *IEEE Signal Process. Lett.*, vol. 21, no. 7, pp. 880–884, Jul. 2014. 3

[30] B. Chen, J. Wang, H. Zhao, N. Zheng, and J. C. Principe, "Convergence of a fixed-point algorithm under maximum correntropy criterion," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1723–1727, Oct. 2015. 3

[31] R. He, L. Wang, Z. Sun, Y. Zhang, and B. Li, "Information theoretic subspace clustering," *IEEE Trans. Neural Networks Learning Syst.*, vol. PP, no. 99, pp. 1–13, Dec. 2015. 3

[32] R. He, W. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 261–275, Feb. 2014. 3, 4, 5

[33] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Scientific Computing*, vol. 27, no. 3, pp. 937–966, Dec. 2005. 4, 5

[34] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 471–478. 5

[35] X. Yuan and B. Hu, "Robust feature extraction via information theoretic learning," in *Proc. Int. Conf. Mach. Learning (ICML)*, Montreal, Canada, Jun. 2009, pp. 1193–1200. 5

[36] A. Martinez and R. Benavente, "The AR face database," *CVC Tech. Rep.*, vol. 24, 1998. 5

[37] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001. 5, 6

[38] F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vision Image Understanding*, vol. 106, no. 1, pp. 59–70, Apr. 2007. 5, 6

[39] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Mach. Vision Applicat.*, vol. 24, no. 5, pp. 971–981, Jul. 2013. 5, 6

[40] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, vol. 2, New York, NY, Jun. 2006, pp. 2169–2178. 6

[41] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, Providence, RI, Jun. 2012, pp. 1234–1241. 6