A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification

Xiaofeng Liu*
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
liuxiaofeng@cmu.edu

Zhaofeng Li*
CIOMP, CAS
University of Chinese Academy of Sciences
Beijing, China
lizhaofeng14@mails.ucas.ac.cn

Lingsheng Kong, Zhihui Diao, Junliang Yan
CIOMP
Chinese Academy of Sciences
Changchun, China
konglingsheng, diaozh@ciomp.ac.cn; jlyciomp@163.com

Chao Yang
Viterbi school of engineering
University of Southern California
Los Angeles, USA
harryyang.hk@gmail.com

***Abstract*—Various representation-based methods have been developed and shown great potential for pattern classification. To further improve their discriminability, we propose a Bi-level optimization framework in terms of both low-dimensional projection and collaborative representation. Specifically, during the projection phase, we try to minimize the intra-class similarity and inter-class dissimilarity, while in the representation phase, our goal is to achieve the lowest correlation of the representation results. Solving this joint optimization mutually reinforces both aspects of feature projection and representation. Experiments on face recognition, object categorization and scene classification dataset demonstrate remarkable performance improvements led by the proposed framework.**

Keywords-classification; joint optimization; low-dimensional projection; sparse representation

I. INTRODUCTION

Classification plays a prominent role in many real-world applications [1]. Conventional classification methods, such as the SVM [2] and neural network [3], train a specific classifier on a training set for a certain task, which are sensitive to the variance of the testing sets. Recently, the sparse representation classification (SRC) [4] and collaborative representation classification (CRC) [5] have aroused broad research interest in the pattern recognition and computer vision areas, especially for the adaptation to real-world applications.

The basic idea of SRC aims to find the minimum residuals between a given test sample and its sparse linear combination of all the training samples in different classes. The weighted non-zero representation coefficients are supposed to indicate the contribution of each class and are referred to be the sparse solution obtained by l_1 regularization. Considering the iterative numerical optimization of l_1 regularization, it always leads to heavy computation burden. The CRC was proposed for this issue, in which the l_2 regularization is employed, and it is computationally efficient due to its closed-form solution. Despite we cannot guarantee the sparse property by applying the native l_2 regularization directly on a classification model, a recent study indicates that it is the collaborative representation instead of the l_1-norm sparsity that makes SRC powerful for classification [6]. It usually uses a linear combination of all the *n* training samples $\mathrm{X} = [x_1, x_2, \dots, x_n,]$ to represent a test sample y (*i.e.*, $y = \sum_{i=1}^{n} a_i\, x_i$). Thus, this can be expressed to solve the optimization problem:

$$\min_{a} \{\|y - \mathrm{X}a\|_2^2 + \sigma\|a\|_2^2\} \quad (1)$$

where the $a = [a_1, a_2, \dots, a_n,]$ is the representation coefficients, and the σ as the regularization parameter.

Despite its wide use, it still suffers from the insufficient discriminability and sensitivity to noise. As classification or recognition by machines always involves a measurement of similarity, a key challenge is to develop a more discriminative and efficient representation to balance the complex distribution of intra- and inter- class variations.

For the small sample problems (*e.g.* face recognition), where the training sample size is smaller than the dimension of original images, the feature extraction (dimensionality reduction) becomes necessary before we implement CRC or SRC. Moreover, with relatively insufficient training data, the high dimensionality normally leads to a "curse of dimensionality" problem, which degrades the classification accuracy [7]. Therefore, it is necessary to find a low-dimensional representation that can retain sufficient discriminative information from the high-dimension data. How should we perform dimensionality reduction, so that CRC

*contribute equally

Hong Kong Government General Research Fund (152202/14E)
Youth Innovation Promotion Association, CAS (2017264)
Innovative Foundation of CIOMP, CAS (Y586320150)

achieves the optimum performance in the reduced-dimensional space still remains as a challenging task. The conventional solution usually involves the global Euclidean structure, and neglect the local manifold structure of the images, such as principle component analysis (PCA) [8] and linear discriminant analysis (LDA) [8]. To make up their shortages, locality preserving projection (LPP) [9] is proposed to optimally preserve the neighborhood structure of the data set. Nonetheless, these methods fail to utilize the high-level discriminative information [10]. Lately, several subspace learning methods are proposed for extracting more discriminative features for image understanding [11]. Li et al. [12] proposed a robust subspace learning method for feature extraction, and the fisher criterion was embedded into the SRC for dimensionality reduction, which achieves a good performance in image classification task [13]. However, even if some improvements have been made, these methods may still encounter the problems of ineffective representation and insufficient robustness *etc.* In [14], a dimensionality reduction projection is steered by the decision rule of SRC. However, their classification steps are neither two-fold discriminative nor jointly optimized.

Our work focuses on the low-dimensional projection space and then we develop the loss function based on CRC. A task driven Bi-level optimization framework is then built to incorporate these two phases together. The projection phase is considered as the lower-level constraints, where the minimum intra-class similarity and inter-class dissimilarity are enforced. The cost function of lowest correlation collaborative representation is formulated as the upper-level constraint. Then, the parameters of both phases are updated via the Stochastic gradient decent algorithm.

The three major contributions in this paper are: 1) We propose a novel framework to integrate minimum similarity projection and lowest correlation representation. 2）Stochastic gradient decent algorithms are developed to solve the Bi-level model. 3) The low-dimensional projection and the closed form solution of both phases can make the algorithm more efficient.

II. MODEL FORMULATION

A. Low-dimensional feature projection

We aim to generate a projection matrix $\mathbf{P}$ by minimizing the intra-class similarity and inter-class dissimilarity in order to map the original samples $\mathrm{X} = \{x_i | x_i \in \mathbb{R}^m, i = 1, \dots, n\}$ to a d-dimensional space ($d < m$). The projection phase not only reduces the computation burdens, but also tries to offer a more effective and discriminative representation for the classification by exploring the label information with distance measurements.

The triplets $\{x, x^-, x^+\}$ of samples are constructed, where the x is a randomly selected sample, $x^-$and x^+ are the samples with different and same class labels respectively. The negative pairs $\{x, x^-\}$ constructed the negative set X^-, while the positive pairs $\{x, x^+\}$ constructed the positive set X^+. By taking the advantages of both the with-in class similarity and between-class dissimilarity constraints, our methods to some extent consider more discrimination than the original LDA Hash [15]. Based on this, the optimization problem can be expressed as the following loss function:

$$\mathcal{L} = \mathbb{E}\{\|\mathbf{P^T}x_i - \mathbf{P^T}x_m^+\|^2 | \mathrm{X}^+\} - \mu\mathbb{E}\{\|\mathbf{P^T}x_i - \mathbf{P^T}x_n^-\|^2 | \mathrm{X}^-\} \quad (2)$$

where the

$$\mathbb{E}\{\|\mathbf{P^T}x_i - \mathbf{P^T}x_m^+\|^2 | \mathrm{X}^+\} = Tr\{\mathbf{P}\textstyle\sum_{\mathrm{X}^+}\mathbf{P^T}\} \quad (3)$$

$$\mathbb{E}\{\|\mathbf{P^T}x_i - \mathbf{P^T}x_n^-\|^2 | \mathrm{X}^-\} = Tr\{\mathbf{P}\textstyle\sum_{\mathrm{X}^-}\mathbf{P^T}\} \quad (4)$$

where the $\sum_{\mathrm{X}^+} = \mathbb{E}\{(x_i - x_m^+)(x_i - x_m^+)^T | \mathrm{X}^+\}$ and $\sum_{\mathrm{X}^-} = \mathbb{E}\{(x_i - x_n^-)(x_i - x_n^-)^T | \mathrm{X}^-\}$ are the covariance matrices of positive pairs and negative pairs respectively. Thus, Eq. (1) can be formulated as:

$$\mathcal{L} = Tr\{\mathbf{P}\textstyle\sum_{\mathrm{X}^+}\mathbf{P^T}\} - \mu Tr\{\mathbf{P}\sum_{\mathrm{X}^-}\mathbf{P^T}\} \quad (5)$$

Finally, we have:

$$\mathcal{L} \propto Tr\{\mathbf{P}\textstyle\sum_{\mathrm{X}^+}\sum_{\mathrm{X}^-}^{-1}\mathbf{P^T}\} = Tr\{\mathbf{P}\sum_R\mathbf{P^T}\} \quad (6)$$

where $\sum_{\mathrm{X}^+}\sum_{\mathrm{X}^-}^{-1}$ is a ratio matrix between positive and negative covariance matrices. It is obvious that $\sum_R$ can be decomposed by singular value decomposition (SVD) for an orthogonal matrix. Thus, this orthogonal matrix $\mathbf{P}$ can map the samples into a space spanned by d smallest eigenvectors, which has the minimum similarity of different classes. By doing this, the dimension of the sample turns to be much smaller, but yet quite effective for classification.

B. Bi-level Optimization Formulation

We formulate our objective cost function to be the following bi-level optimization:

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{w}} \quad C(\mathbf{P}, \mathbf{w}) \\ &s.t. \quad \mathbf{P} = arg\ min_{\mathbf{P}} Tr\{\mathbf{P}\textstyle\sum_{\mathrm{X}^+}\mathbf{P}^T\} - \mu Tr\{\mathbf{P}\sum_{\mathrm{X}^-}\mathbf{P^T}\} \end{aligned} \quad (7)$$

where $C(\mathbf{P}, \mathbf{w})$ is the cost function that evaluates the loss of sparse coding, with $\mathbf{P}$ as its input and $\mathbf{w}$ as the parameters need to be learned. Based on different graph regularization methods, it can be formulated as various objectives of sparse representations. Limited by the length, we only discus and solve the l_2 regularization based representation (*i.e.*, CSC) in subsection C.

Both the theories and applications of the Bi-level optimization have been explored in recent years. A general Bi-level sparse coding model was proposed to learn dictionaries across the coupled signal spaces [16]. The similar formulations have also been extended to clustering [17] and regression tasks [18], which are related to and have inspired our proposed architecture, despite they are limited in a pure sparse coding framework.

C. Collaborative Feature Representation

In this phase, we represent the feature vectors on the low-dimensional space for the classification. As a typical l_2 regularization based representation, the CRC is more computationally efficient and robust than the regression-based classification [19] and two-phase test sample representations [20]. A recent study demonstrates that collaborative representation plays a more important role in classification than the sparse representation.

As we project the original samples to the d-dimensional space, we represent these feature vectors in it using the $v = \mathbf{P^T}y$ and $\mathrm{F} = \mathbf{P^T}\mathrm{X}$. Following the previous notations, the loss function of the CRC can be formulated as:

$$C(\mathbf{P},\rho) = \overset{min}{\rho}\{\|v - \mathrm{F}\rho\|_2^2 + \sigma\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2\} \quad (8)$$

where the **P** is the input and the $\rho = [\rho_1, \rho_2, \ldots, \rho_z,]$ is the to be learned representation coefficients.

Except for the label information, we also consider the specific class samples on the regularization term, which may be also useful to generate the decorrelation effect for representation and classification. Considering the convexity and differentiability of the loss function $C(\mathbf{P},\rho)$, we can obtain the optimal solution by setting objective function to 0 and taking the derivative with respect to the parameter ρ.

The derivative with respect to ρ of the first term can be calculated as:

$$\frac{d}{d\,\rho}\|v - \mathrm{F}\rho\|_2^2 = -2\mathrm{F}^{\mathrm{T}}(v - \mathrm{F}\rho) \quad (9)$$

As the second term dose not explicitly contain the ρ, the n partial derivates are computed:

$$\frac{\partial\,\sigma\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2}{\partial\,\rho_i} = \sigma\sum_{i=1}^{n}\frac{\partial\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2}{\partial\,\rho_i}$$
$$= \sigma K\frac{\partial}{\partial\,\rho_i}\|\mathrm{F}_i\rho_i\|_2^2 = 2\sigma K\mathrm{F}_i^{\mathrm{T}}(\mathrm{F}_i\rho_i) \quad (10)$$

Then we get their sum to form the derivative of the second term:

$$\frac{d\,\sigma\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2}{d\,\rho} = \begin{pmatrix}\frac{\partial\,\sigma\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2}{\partial\,\rho_1}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\partial\,\sigma\sum_{i=1}^{n}\|\mathrm{F}_i\rho_i\|_2^2}{\partial\,\rho_n}\end{pmatrix} = \begin{pmatrix}2\sigma K\mathrm{F}_1^{\mathrm{T}}(\mathrm{F}_1\rho_1)\\ \cdot\\ \cdot\\ \cdot\\ 2\sigma K\mathrm{F}_n^{\mathrm{T}}(\mathrm{F}_n\rho_n)\end{pmatrix}$$

$$= 2\sigma K\begin{pmatrix}\mathrm{F}_1^{\mathrm{T}}\mathrm{F}_1 & & O\\ \vdots & \ddots & \vdots\\ O & & \mathrm{F}_n^{\mathrm{T}}\mathrm{F}_n\end{pmatrix} \quad (11)$$

Let the $\mathbf{Q}=\begin{pmatrix}\mathrm{F}_1^{\mathrm{T}}\mathrm{F}_1 & & O\\ \vdots & \ddots & \vdots\\ O & & \mathrm{F}_n^{\mathrm{T}}\mathrm{F}_n\end{pmatrix}$ and combing the derivative of the first and second term, we get the derivative of our loss function $\frac{d\,C(\mathbf{P},\rho)}{d\rho} = -2\mathrm{F}^{\mathrm{T}}(v - \mathrm{F}\rho) + 2\sigma n\mathbf{Q}$. By setting it to be 0, we can derivate the closed-form solution of the optimal ρ with:

$$\rho = (\mathrm{F}^{\mathrm{T}}\mathrm{F} + \sigma n\mathbf{Q})^{-1}\mathrm{F}^{\mathrm{T}}v \quad (12)$$

D. Classification in test phase

The proposed method classifies the samples on a discriminative *d*-dimensional space. Thus, training samples X and test samples y should be projected as F and *v* with the projection matrix **P**. Finally, we use Eq. (12) to compute the representation coefficient ρ of the projected test sample *v*. Then, a test sample *v* is classified to the c^{th} class according to the following optimization

$$c = \overset{\arg min}{i}\|v - \mathrm{F}_i\rho_i\|_2^2 \quad (13)$$

In summary, we present a detailed description of the proposed method in Algorithm 1.

Algorithm 1 Stochastic gradient descent algorithm for solving (7), with $C(\mathbf{P},\rho)$ as defined in (8)

Require: original training samples X; parameter σ; $\mathbf{P_0}, \rho_0$ (initial); ITER (number of iterations) ; t_0, μ(learning rate)

1. Generate triplets $\{x, x^-, x^+\}$ from the X.
2. Construct positive set $\mathrm{X}^+$and negative set X^- using the $\{x, x^+\}$ and $\{x, x^-\}$ respectively.
3. For t=1 to ITER DO
4. $\sum_{\mathrm{X}^+} = \mathbb{E}\{(x_i - x_m^+)(x_i - x_m^+)^T|\mathrm{X}^+\}$ and $\sum_{\mathrm{X}^-} = \mathbb{E}\{(x_i - x_n^-)(x_i - x_n^-)^T|\mathrm{X}^-\}$
5. Compute the projections: $v = \mathbf{P}^{\mathbf{T}}y$ and $\mathrm{F} = \mathbf{P}^{\mathbf{T}}\mathrm{X}$
6. Compute $\rho^* = (\mathrm{F}^{\mathrm{T}}\mathrm{F} + \sigma n\mathbf{Q})^{-1}\mathrm{F}^{\mathrm{T}}v$
7. Choose the learning rate$\mu_t = min\left(\mu, \mu\frac{t_0}{t}\right)$
8. Update **P** and ρ by a projected gradient step:
   $\rho = \prod_{\rho}\left[\rho - \mu_t\left(\nabla_\rho C(\mathbf{P},\rho)\right)\right]$
   $\mathbf{P} = \prod_{\mathbf{P}}[\mathbf{P} - \mu_t(\nabla_{\mathbf{P}}C(\mathbf{P},\rho))]$
   where $\prod\rho$ and $\prod\mathbf{P}$ are respectively orthogonal projections on the embedding spaces of ρ and **P**.
9. END FOR

Ensure: **P**, ρ

E. Convergence and Complexity Analysis

The Stochastic gradient decent algorithms converge to the stationary points under a few stricter assumptions than ones satisfied in this paper. A non-convex convergence proof assumes three times differentiable cost functions [21]. As a typical case in machine learning, we use SGD in a setting where it is not guaranteed to converge in theory, but it behaves well in several practices. Assuming *n* samples and the dimension of the representation coefficient z, the time complexity of triplet generation is $O(n^2)$. Specifically, in each iteration of Algorithm 1, step 8 takes $O(n)$ time. Step 4 is solved by the feature-sign algorithm [22], which is reduced to a series of quadratic programming (QP) problems. The computational bottleneck lies in solving the inverse of the matrix in step 6, where applying the Gauss-Jordan elimination method takes $O(z^3)$ time per sample. Thus, Algorithm 1 takes $O(nz^3)$ time per iteration, and $O(n^2+Cnz^3)$ in total, where C is a constant absorbing epoch numbers, *etc*.).

III. EXPERIMENTS

We conduct our classification experiments on three popular real datasets, which involves the face recognition, screen classification and object categorization.

A. Face Recognition

The Extended Yale B face dataset [23] consists of 2414 facial images from 38 identities with frontal pose under different illumination conditions, as shown in Figure 1. Each image has a size of 192×168, and we cropped and resized them to 84×96 in the preprocessing stage. We split the database to 5 subsets in a strict subject independent manner, and a 5-fold cross-validation is employed. Three of them are used for training and the other two are used for validation and testing respectively. We randomly

Figure 1. Some examples of from 2 subjects (rows) in YALE B.

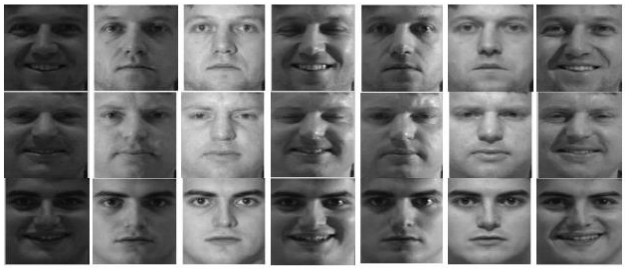

Figure 2. Some examples of from 3 subjects (rows) in CMU Multi-Pie.

selected 10, 20, 30, and 40 images of each individual as original training samples. The hyperparameter σ was set to 0.001. The classification results of nine methods are shown in Table I. Compared with other methods, the proposed method increases almost 10% on the recognition rates under different conditions.

Moreover, the superior recognition accuracy was obtained under low-dimension space (d=1000), while the other methods under original dimension (m=8064).

TABLE I. AVERAGE CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON THE EXTENDED YALE B DATASET

Methods	**Number of training samples per subject**			
	10	***20***	***30***	***40***
CRC[5]	67.54	96.63	98.86	99.46
L1LS[28]	73.25	74.52	84.13	84.98
INNC[29]	42.35	40.79	43.42	41.34
FCM[30]	43.86	46.41	49.38	40.79
L1+L2[32]	59.45	66.21	74.77	73.14
LRE[33]	72.47	77.21	82.74	82.35
(S+C)RC[34]	65.20	72.19	79.88	81.58
(S+C)RC+[35]	65.51	71.98	86.74	91.67
Proposed	77.73	85.35	94.58	96.49

To evaluate the robustness of illuminations and expressions, the CMU Multi-PIE face dataset [24] was used to test our method. The CMU Multi-PIE face dataset contains face images of 337 identities captured from various poses, expressions and illuminations. Figure 2. shows some examples of this dataset. In our experiment, we chose a subset composed of 249 identities with smiling expression under 7 different illumination conditions and with a frontal pose under 20 different illumination conditions. We cropped and resized all images to 40×30 manually. Similar to the Extended YALE B dataset, the same 5-fold setting was used. The first 3, 5, 7, 9 illumination images from the 20 illumination images and only one image from 7 smiling images were used as training samples and the remaining images. Parameter σ was set to 0.00001 and the dimension was set to 1000.

Figure 3. Some examples of of a class in Coil 100 dataset.

Figure 4. Some examples of of a class in Coil 100 dataset.

Table II reports the recognition rates of nine methods on four test sets. The results demonstrate that the proposed method obtains higher recognition rate than other eight methods, which means that our method owns more robustness to variations of illuminations and expressions.

TABLE II. AVERAGE CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON THE CMU MULTI-PIE FACE DATASET

Methods	**Number of training samples per subject**			
	4	***6***	***8***	***10***
CRC[5]	90.13	96.63	98.86	99.46
L1LS[28]	92.79	99.08	99.66	99.81
INNC[29]	46.74	50.01	57.26	71.16
FCM[30]	55.49	56.55	63.03	71.25
L1+L2[32]	68.05	67.62	72.27	84.67
LRE[33]	82.05	95.62	98.82	99.50
(S+C)RC[34]	90.34	95.47	97.36	98.32
(S+C)RC+[35]	89.80	95.09	98.01	99.31
Proposed	95.48	99.04	99.89	99.86

B. Scene classification

The scene 15 dataset [25] contains 15 categories of natural scenes, and image number of each category varies from 200 to 400. The average image size is about 250×300 pixels. Some of the scenes, such as the forest, bedroom, country scenes and office are shown in Figure 3. In this experiment, we use spatial pyramid feature provided by [26] to replace source image for the classification experiment, which have 3000 dimensions. The dataset was randomly separated to 5 parts equally to make the 5-fold cross-validation. We randomly extract the 50, 75 or 100 images from each scene in these 3 training subsets for the model training. The hyperparameter σ was set to 0.001. Table III lists the results of nine state-of-the-art methods. We tested all methods with original dimension. The results demonstrated that the proposed method obtains the best classification accuracy among those methods, and it is capable of improving discriminability.

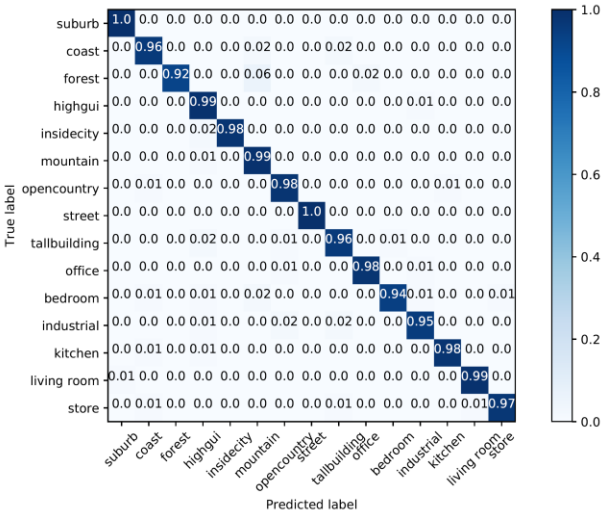

Figure 5. Confusion matrices under 100 training samples per class.

We also presented the details of classification results using the confusion matrices. The confusion matrices under 100 training samples per class are shown in Figure 5.

TABLE III. AVERAGE CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON THE SCENE15 DATASET

Methods	**Number of training samples per subject**		
	50	***75***	***100***
CRC[5]	95.20	96.55	96.55
L1LS[28]	95.23	96.55	96.65
INNC[29]	93.47	95.54	96.05
FCM[30]	89.39	91.57	92.19
(M+S)RC[31]	95.26	96.76	96.85
(S+C)RC[34]	88.73	90.36	90.75
(S+C)RC+[35]	95.20	96.55	96.65
MOLP[36]	88.67	90.92	96.18
Proposed	95.37	97.23	97.42

TABLE IV. AVERAGE CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON THE COIL-100 DATASET

Methods	**Number of training samples per subject**				
	10	***20***	***30***	***40***	***50***
CRC[5]	43.57	49.64	57.24	60.54	99.46
L1LS[28]	48.89	56.92	68.57	74.34	99.81
INNC[29]	49.32	53.59	65.93	70.81	71.16
FCM[30]	51.34	60.44	73.90	74.63	71.25
(M+S)RC[31]	54.76	63.60	77.55	79.13	84.67
(S+C)RC[34]	48.73	55.02	66.57	69.25	99.50
(S+C)RC+[35]	47.11	52.85	60.26	64.03	98.32
MOLP[36]	45.48	54.40	67.29	75.47	99.31
Proposed	55.42	65.65	78.55	81.03	82.41

C. *Object categorization*

The COIL100 dataset [27] contains 7200 images that from 100 classes. Each class has 72 images from different views. The resolution is 128 × 128 pixels and be converted into gray images in our preprocessing stage. Figure 4. shows some image examples of a class in the COIL100 dataset. We split the COIL100 database to 5 subsets in a strict subject independent manner, and a 5-fold cross-validation is employed. Data from 3 subsets are used for training and the others are used for validation and testing respectively. Then, we randomly extract the 10, 20, 30, 40, or 50 images from each subject in these 3 training subsets for the model training. The hyperparameter σ was set to 1. Table IV shows the classification accuracies of different methods on the Coil-100 dataset. Not surprisingly, it also beats the baseline methods obviously benefit from the combination of two-fold discriminative ability, which means that the extracted low-dimension features are efficient for improving the classification accuracy.

D. *Comparison of the computation time*

We presented the average running time of each test sample taken by nine methods on our experiment platform (a PC with 3.2 GHz i7 4770 CUP and 12GB RAM using the MATLAB 2017a). Table V lists the details of the running time on the COIL-100 dataset. It is obvious that the proposed method runs faster than L1LS, INNC, (S+C)RC+ and MOLP. The running speed of our method is close to those of FCM, (M+C)RC and CRC and lower than that of (S+C)RC. However, our method can extract more discriminative feature and obtains higher classification accuracy than those of comparison methods. Therefore, although our method did not obtain the fastest record in efficiency, its speed is still feasible to real-time face recognition applications with competitive its classification performance.

TABLE V. AVERAGE RUNNING TIME OF EACH TEST SAMPLES FROM THE COIL-100 DATASET

Methods	**Number of training samples per subject**				
	10	***20***	***30***	***40***	***50***
CRC[5]	0.18s	0.41s	0.95s	1.25s	1.44s
L1LS[28]	20.33s	33.71s	41.20s	49.15s	57.20s
INNC[29]	0.51s	0.83s	1.22s	2.45s	4.13s
FCM[30]	0.15s	0.49s	0.81s	1.20s	1.52s
(M+S)RC[31]	0.17s	0.58s	1.09s	1.26s	1.41s
(S+C)RC[34]	0.06s	0.14s	0.15s	0.17s	0.17s
(S+C)RC+[35]	1.15s	1.71s	2.39s	3.52s	5.11s
MOLP[36]	4.50s	5.44s	7.38s	9.19s	11.06s
Proposed	0.32s	0.67s	1.21s	2.30s	3.93s

Based on the experimental results, our method is propitious to the image classification task, and obtains competitive results in face recognition, object categorization, and scene classification. Especially, compared with other representation-based methods, the proposed method obtains relatively large improvement on some classification tasks, such as face recognition on the Extended Yale B dataset and object classification on the Coil-100 dataset. It demonstrates to be more robust to shape variations, as it extracts more discriminative information for resisting the influence of the size and view variations for object images or the illumination, pose and expression changes for face images. The best classification performance of our method can be achieved under low-dimension, which means that the proposed method can capture the low-dimension discriminative information and lead to an effective classification result. Even though the proposed method is performed on the data with original dimension, the discriminative property of the representation result remains retained.

IV. CONCLUSION

In this paper, we propose a joint optimization framework to combine the discriminative low-dimensional projection and lowest-correlation collaborative representation by design a specific regularization term and projection matrix. The projection minimizing the with-in class similar covariance and maximizing between-class dissimilar covariance. The task-driven Bi-level optimization mutually reinforces both the projection and representation phases via an end-to-end training. Explements on face recognition, object categorization and screen classification verify the remarkable performance improvements led by our joint optimization.

ACKNOWLEDGMENT

The careful work the anonymous reviewers are greatly appreciated. We wish to thank Professor Vijayakumar Bhagavatula for his constructive suggestions and Professor Ping Jia for his foresight supports.

REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, et al. Label-embedding for image classification[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(7): 1425-1438.

[2] C. Cortes, V. Vapnik. Support vector machine[J]. Machine learning, 1995, 20(3): 273-297.

[3] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[4] 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.

[5] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?" in Proc. Int. Conf. Comput. Vis. (ICCV), Nov. 2011, pp. 471-478.

[6] R. Baraniuk, "Compressive sensing," IEEE Signal Process. Magazine, vol. 24, no. 4, pp. 118-121, Jul. 2007.

[7] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in Proc. IEEE Conf. Comput. Vis. Pattern Recognition, Jun. 1991, pp. 586–591.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp: 711-720, 1997.

[9] 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.

[10] X. D. Jiang, B. Mandal and A. Kot, "Eigenfeature Regularization and Extraction in Face Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 3, pp. 383-394, March 2008.

[11] Q. Wang, L. Ma, Q. Gao, Y. Li, Y. Huang, Y. Liu, "Adaptive maximum margin analysis for image recognition," Pattern Recognit., vol. 61, pp: 339-347, 2017

[12] 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. 2015

[13] Y. Xu, Z. Zhong, J. Yang, J. You, D. Zhang, "A New Discriminative Sparse Representation Method for Robust Face Recognition via L2 Regularization," IEEE Trans. Neural Netw. Learn. syst., DOI:10.1109/TNNLS.2016.2580572, 2016.

[14] J. Yang, D. Chu, L. Zhang. "Sparse representation classifier steered discriminative projection with applications to face recognition". IEEE transactions on neural networks and learning systems, 24(7): 1023-1035, 2013.

[15] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua, "LDAHash: Improved matching with smaller descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 1, pp: 66-78, 2012.

[16] J. Yang, Z.Wang, Z. Lin, X. Shu, and T. Huang. Bilevel sparse coding. for coupled feature spaces. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2360–2367. IEEE, 2012.

[17] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(4):791–804, 2012.

[18] Z. Wang, Y. Yang, S. Chang, J. Li, S. Fong and T. S. Huang. A Joint Optimization Framework of Sparse Coding and Discriminative Clustering. In IJCAI, pages 3932-3938, 2015.

[19] I. Naseem, R. Togneri, and M. Bennamoun, "Robust regression for face recognition," Pattern Recognit., vol. 45, no. 1, pp. 104-118, Jau. 2012.

[20] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, "A two-phase test sample sparse representation method for use with face recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 9, pp. 1255-1262, Sep. 2011.

[21] H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. In Advances in neural information processing systems, pages. 801–808, 2006.

[22] Y. Yang, F. Liang, S. Yan, Z. Wang, and T. S. Huang. On a theory of nonparametric pairwise similarity for clustering: Connecting clustering to classification. In Advances in Neural Information Processing Systems, pages 145–153, 2014.

[23] 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.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and Baker S., "Multi-PIE," in Proc. IEEE Int. Conf. Automat. Face Gesture Recognit., Sep. 2008, pp. 1–8.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," In Computer Vision and Pattern Recognition, 2006 IEEE Conference on (Vol. 2, pp. 2169-2178). 2006

[26] Z. Jiang, Z. Li, and L.S. Davis. "Label consistent K-SVD: Learning a discriminative dictionary for recognition," IEEE Trans. Pattern Anal. Mach. Intell., 2013, 35(11): 2651-2664.

[27] S. A, Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)". Technical Report CUCS-005-96, 1996.

[28] V. Shia, A. Yang, S. Sastry, A. Wagner, and Y. Ma, "Fast l1-Minimization and Parallelization for Face Recognition," IEEE Trans. Image Process., vol. 22, no. 8, pp. 3234–3246, Aug. 2013.

[29] Y. Xu, Q. Zhu, Y. Chen, and J. S. Pan, "An improvement to the nearest neighbor classifier and face recognition experiments," Int. J. Innov. Comput., Inf. Control, vol. 8, no. 12, pp. 1349-4198,Feb. 2012.

[30] Z. Liu, J. Pu, T. Huang, and Y. Qiu, "A novel classification method for palmprint recognition based on reconstruction error and normalized distance," Applied Intell., vol. 39, no. 2, pp. 307-3148, Feb. 2013.

[31] Y. Xu, B. Zhang, Z. Zhong. Multiple representations and sparse representation for image classification[J]. Pattern Recognition Letters, 2015, 68: 9-14.

[32] S. Zeng, J. Gou, L. Deng. An antinoise sparse representation method for robust face recognition via joint l 1 and l 2 regularization[J]. Expert Systems with Applications, 2017, 82: 1-9.

[33] W. K. Wong, Z. Lai, J. Wen, et al. Low-Rank Embedding for Robust Image Feature Extraction[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2905-2917.

[34] C. Tian, G. Sun, Q. Zhang, et al. Integrating Sparse and Collaborative Representation Classifications for Image Classification[J]. International Journal of Image and Graphics, 2017, 17(02): 1750007.

[35] S. Zeng, J. Gou, X. Yang. Improving sparsity of coefficients for robust sparse and collaborative representation-based image classification[J]. Neural Computing and Applications, 2017: 1-14.

[36] Y. Yu, C. Ren, D. Dai, et al. Kernel Embedding Multiorientation Local Pattern for Image Representation[J]. IEEE Transactions on Cybernetics, 2017.