This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

Double Relaxed Regression for Image Classification

Na Han, Jigang Wu, *Member, IEEE,* Xiaozhao Fang, *Member, IEEE,* Wai Keung Wong, Yong Xu, *Member, IEEE,* Jian Yang, *Member, IEEE,* and Xuelong Li, *Fellow, IEEE,*

***Abstract*—This paper addresses two fundamental problems of 1) learning discriminative model parameters and 2) avoiding over-fitting which often occurs in regression based classification tasks. We formulate these two problems in terms of relaxing both the strict binary label matrix and graph regularization term into more flexible forms so that the margins between different classes are enlarged as much as possible and the problem of over-fitting is avoided to some extent. This task is accomplished by the proposed double relaxed regression (DRR) method. The convex problem of DRR is solved efficiently with an iterative procedure. Extensive experiments on synthetic and real world image data sets demonstrate the effectiveness of the proposed method in terms of both classification accuracy and running time.**

***Index Terms*—Regression, image classification, convex problem, optimization, computer vision.**

I. INTRODUCTION

Least squares regression (LSR) is a simple but efficient tool for data analysis. LSR has been widely applied in fields of machine learning and computer vision [1], [2], [3], [4]. Due to the mathematically tractable and computation efficient of LSR, many variants such as robust regression (RR) [5], weight LSR [6], partial LSR [7] and nonnegative least squares (NNLS) [8] have been proposed for classification and regression. In addition, many popular models usually have strong connections to conventional LSR. For example, ridge regression [9], LASSO problem [10], support vector machine (SVM) [11] and logistic regression (Log_R) [12], and so on. LSR has been also used for feature selection. For example, Xiang et al. [13] proposed a discriminative least squares regression (DLSR) framework for feature selection and multiclass classification. The core idea is to enlarge the margins by using the ϵ-dragging technique. The least squares model has been also used for semi-supervised learning [14], [15]. Recently, various algorithms have been proposed to extend LSR to the reproducing kernel Hilbert space [16], [17].

Linear regression (LR) is also a simple regression analysis method [18]. For a collection of n training samples represented as a matrix $X = [x_1, ..., x_n] \subset \Re^{m\times n}$, LR can be defined as

$$\min_{W} \|W^T X - Y\|_F^2 + \lambda \|W\|_F^2 \tag{1}$$

where $W \in \Re^{m\times c}$ (c denotes the number of classes) is the transformation matrix which transforms original inputs into their label space and $Y \in \Re^{c\times n}$ is the binary label matrix that is defined as follows: for each training sample x_i ($i = 1, ..., n$), $y_i \in \Re^c$ is its label vector. If x_i is from the kth class ($k = 1, ..., c$), then only the kth entry of y_i is one and all the other entries are zero. $\lambda \geq 0$ is a scalar that weights the second term in (1). As shown in [19], the LS loss between the regression results and the binary labels cannot closely reflect the classification ability of the regression model. In other words, in most of the above regression-based methods the training samples are exactly transformed into a linear or a strict binary label matrix, such as Y in (1) which is too rigid to learn a discriminative transformation in practice [13], [14].

One strategy is to use different loss functions for LSR. For example, the squared hinge loss and the hinge loss [11] are selected as the surrogate loss functions of the classification error [11], [20]. The negative log-exponent loss [12] and multicategory hinge loss [21] [22] are also used for multiclass classification. Another strategy is to preserve the least square loss but adopt other regression targets. For example, An et al. proposed a kernel ridge regression method for face recognition in which the regression targets are defined as the regular simplex vertices in $c-1$ space (c is the number of classes). Xiang et al. [13] proposed a so called ϵ-dragging technique to enlarge the distances between different classes during regression. Although these methods show the remarkable performance in classification, it is inevitable that they encounter the problem of over-fitting. For example, DLSR [13] uses the so called ϵ-dragging technique to force the regression targets of different classes moving along opposite directions which makes the transformation excessively fit the labels so as to obtain large margins. To this end, Fang *et al.* [23] proposed a regularized label relaxation (RLR) linear regression method in which the class compactness graph is used to avoid the problem of over-fitting. Graph embedding technique [24] is commonly used to address the problem of over-fitting [14], [23]. In these methods, they only use a single transformation matrix to transform samples into a subspace in which the local structure of data is preserved. However, such a single transformation may be too strict to provide more freedom for learning better margins. A natural idea is to use more

N. Han, J. Wu and X. Fang and are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: hannagdut@126.com, asjgwu@gmail.com, xzhfang168@126.com,).

W. K. Wong is with Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong, and The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China. (e-mail: calvin.wong@polyu.edu.hk).

Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, Guangdong, P. R. China. He is also with the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen 518055, Guangdong, P. R. China (e-mail: yongxu@ymail.com).

J. Yang is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: csjyang@mail.njust.edu.cn).

X. Li is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xian 710072, P. R. China. (email: xuelong_li@opt.ac.cn).

Corresponding author: Jigang Wu and Wai Keung Wong (asjgwu@gmail.com and calvin.wong@polyu.edu.hk).

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

transformations for providing more freedom. The goal of our paper is use more transformation matrices to well address the problem of over-fitting. Our previous work [23] uses a single transformation matrix to address the problem of over-fitting. Although the classification accuracy is improved, the subsequent experiments in this paper verify that the use of more transformation matrices can better address the problem of over-fitting. Since our paper focuses on image classification, we give a brief review of some recently proposed representative image classification methods in the following.

Recently, many excellent image classification methods have been proposed. For example, some methods were proposed to learn discriminative features for image classification [25], [26], [27]. Representation-based feature learning methods achieve outstanding performance in image classification [28]-[34]. The first representation-based method is sparse representation classification [29] which finds the smallest number of training samples to represent the test sample and uses the representation results to perform classification. In addition, dictionary learning, which is based on the sparse representation, is also used for image classification. For example, Mairal et al. [35] proposed a task-driven dictionary learning (TDDL) for recognition. Jiang et al. [36] proposed a label consistence K-SVD dictionary learning (LC-KSVD) method in which the discriminative dictionary and classifier parameter are jointly optimized in a framework. The second representation based methods are low-rank representation based classification [30], [31], [34], [37]. In these methods, the dimensionality of subspace corresponds to the rank of the corresponding representation matrix and thus the correction among the representation coefficient vectors is exploited to perform face and objective recognition and clustering. Other classification methods have also shown their power in image classification tasks, such as kernel methods [38], [40] and regression methods [32], [39], [41]. Deep learning based methods shown very impressive improvement on objective recognition and image classification [42][43]. Li et al proposed a deep collaborative embedding learning framework for image classification and image retrieval [44]. Trigeorgis et al proposed a deep matrix factorization method for learning image attribute representation [45]. Krizhevsky et al proposed a imagenet classification with deep conventional neural networks learning method and achieved better image classification results [46].

Inspired by DLSR [13] and RLR [23], in this paper a double relaxed regression (DRR) method is proposed for image classification. Specifically, DRR has the following remarkable advantages. Instead of using a single transformation matrix, DRR uses two transformation matrices to address the problem of over-fitting by introducing the class compactness graph. The reason for introducing two transformation matrices is twofolds: First, they enable DRR to obtain better margins. Second, it relaxes the constraint of using a single transformation matrix into two more flexible transformation matrices which enables one of these two matrices to have more freedom so that the problem of over-fitting can be addressed well. Furthermore, inspired by the observation that these two transformation matrices share similar structure of data, i.e., the local data structure, a constraint is introduced to ensure the structure-consistency among these two transformation matrices. To solve the proposed problem, we propose an effective and efficient iterative algorithm with fast convergence. Extensive experiments are conducted on synthetic and real world image data sets to verify the effectiveness of the proposed method.

The remainder of this paper is arranged as follows. In Section II, the related work is introduced. The proposed DRR method is described in Section III. This is followed by extensive experiments using five standard image data sets in Section IV. The paper concludes in Section V.

II. Related Work

In this section, we review the work of RLR [23] for the sake of completeness.

Let us introduce our notations. Let $X \in \Re^{m\times n}$ be the training samples, where m and n are the dimensionality and number of training samples, respectively. The definitions of binary label matrix Y and transformation A are the same as those in (1). We consider the Frobenius norm of matrix X: $\|X\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = Tr[X^T X]$, where X^T denotes the transposed matrix of X and Tr is the trace operator of matrix. In our paper, $\odot$ presents a Hadamard product operator of matrices and I denotes an identity matrix with a suitable size.

To enlarge margins between different classes, DLSR [13] and RLR relax the strict binary label constraint into a slack variable matrix by introducing a non-negative label relaxation matrix (In DLSR, the technique is called ϵ-draggings). RLR further introduces the class compactness graph to address the problem of over-fitting. The objective function of RLR is as follows

$$\min_{W,M} \|W^T X - (Y + B \odot M)\|_F^2 + \lambda Tr(W^T X L X^T W) \quad (2)$$
$$s.t. \;\; M \geq 0$$

where $M = [m_1, ..., m_n] \in \Re^{c\times n}$ is the non-negative label relaxation matrix and B is a luxury binary matrix which is defined as $B_{ij} = \begin{cases} +1 & \text{if } Y_{ij} = 1 \\ -1 & \text{if } Y_{ij} = 0 \end{cases}$. It can be seen that the binary label matrix Y in (1) is redefined as $Y^\circ = Y + B \odot M$. In (2), L is the graph Laplacian and defined as $L = D - Z$, where D is a diagonal matrix and its diagonal entries are defined as $D_{ij} = \sum_j Z_{ij}$. Here, Z is the weight of the class compactness graph and defined as

$$Z_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma}} & \text{if } x_i \text{ and } x_j \text{ share the same labels} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where σ is the heat kernel parameter. In practical, σ is away set to 1 and such setting can ensure a good classification result.

We take three samples for example to show that Y° is more discriminative than Y. Let x_1, x_2, x_3 be three training samples that are from the second, first and third class and their corresponding binary label matrix is defined as $Y = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \in \Re^{3\times 3}$.

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

It is easy to see that the distance between any two samples from different classes is $\sqrt{2}$. For example, the distance between the first and second samples is set to $\sqrt{(0-1)^2+(1-0)^2+(0-0)^2}=\sqrt{2}$. Such definition has a drawback that it cannot closely reflect the classification ability of the model. If we use Y° to substitute Y, then we have

$$Y^\circ = \begin{bmatrix} -m_{11} & 1+m_{12} & -m_{13} \\ 1+m_{21} & -m_{22} & -m_{23} \\ -m_{31} & -m_{32} & 1+m_{33} \end{bmatrix}, \quad s.t. \quad m_{ij} \geq 0.$$

The distance between the first and second samples becomes $\sqrt{(-m_{11}-1-m_{12})^2+(1+m_{12}+m_{22})^2+(-m_{31}+m_{32})^2} \geq \sqrt{2}$. This means that the margins between different classes are enlarged by introducing the non-negative label relaxation matrix M. However, as shown in RLR, such relaxation may make the transformation matrix W excessively fit the labels. As a result, the over-fitting may occur in such case. To solve the problem, Fang et al. [23] proposed the model of (2). By introducing the class compactness graph, the samples from the same class can be kept close together when they are transformed into their label space. In this way, the problem of over-fitting can be avoided to some extent. However, is the formulation in (2) really perfect to solve the over-fitting problem? In the next section, we observe that if we can give more freedom for the transformation matrix, then the over-fitting problem can be solved more perfectly. To this end, a double relaxed regression (DRR) is proposed in which we use double transformation matrices to solve the problem of over-fitting. As far as we know, our DRR is the first-of-its-kind of the idea and method of the graph embedding with two different transformation matrices to address the problem of over-fitting.

III. DRR

In this section, we introduce DRR in detail.

In (2), a single matrix W has less freedom to obtain better margins. In other words, the transformation matrix W faces a dilemma here: on the one hand, it should have the power to transform training samples into their label space and enlarge the margins between different classes as much as possible; on the other hand, it also should have power to guarantee that the samples sharing the same class labels should be kept close together when they are transformed. Facing with such dilemma, we propose to perform these two tasks by using two transformation matrices. Thus, we rewrite (2) as

$$\min_{W,M,A} \|W^T X-(Y+B\odot M)\|_F^2+\lambda_1 \sum_i^n \sum_j^n \|W^T x_i - A^T x_j\|^2 Z_{ij} \quad s.t. \quad M \geq 0 \tag{4}$$

where $A \in \Re^{m\times c}$ is another transformation matrix which is used to share part of the responsibility of W for learning better margins. It is easy to see that these two transformation matrices should share similar structure, i.e., the local structure of data. To capture such similar structure, these two matrices should resemble mutually. In other words, W and A have some similar structure and thus there should be some correspondence between them, i.e., $W = AS$, in which $S \in \Re^{c\times c}$ is a square matrix. In this paper, we call this resemblance as *structure-consistency*, modeled by the following term:

$$\|W - AS\|_F^2 \tag{5}$$

We then minimize this term by adding it to the formulation of (4) and reach the following objective function for DRR:

$$\min_{W,M,A,S} \|W^T X - (Y + B \odot M)\|_F^2 + \lambda_1 \sum_i^n \sum_j^n \|W^T x_i - A^T x_j\|^2 Z_{ij} + \lambda_2 \|W - AS\|_F^2 \quad s.t. \quad M \geq 0 \tag{6}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are two trade-off parameters. The definition of Z is the same as that in (3). From the above objective function, we can see that two strict constraints of rigid binary label matrix and the use of a single transformation matrix are relaxed into more flexible forms. Therefore, DRR is much more flexible and accurate than DLSR and RLR. In this way, we can obtain the optimal margins. By using the *structure-consistency* term, we can effectively capture the resemblance among these two transformation matrices. As a result, transformation matrix A shares part of responsibility of W to address the problem of over-fitting and guarantees that transformation matrix W classifies the data more accurately. From (6), it can be see that differences between RLR and DRR are aspects: (1) the graph embedding with a single transformation matrix is used in RLR, whereas the graph embedding with two different transformation matrix is used in DRR, which is an entirely new way to perform graph embedding and better experimental results are achieved in image classification. Thus, the idea can be extended to more widespread applications as long as they involve the graph embedding. (2) DRR adds the extra term of "structure consistency" to capture the similar data structure embedded in the two transformation matrices. The subsequent analysis indicate that this term is very useful to guarantee the "structure consistency" (see Fig. 6).

In (6), we do not force that S not be a identity matrix. When S is equal to a identity matrix, (6) degrades into (2). Thus, the performance of our DRR is, in theory, equal to that of RLR at least. However, in practical we find that the value of S is very large. The reason may be as follows: in real world applications, the problem of over-fitting is very serious owing to the uncontrolled environment of samples collection. Thus, we should give more freedom for transformation matrix W so that it can fit Y^0 and address the problem of over-fitting well as much as possible. In this case, the value of S should be large (or it should not be identity at least) since it should burden part of the responsibility of W so that the samples from the same classes can be kept close together. Under the circumstances, only the value of S is very large, the objective function can achieve the minimum. The results in Fig. 6 also verify that S is not a identity matrix.

A. Solving the optimization problem

In order to solve problem (6), we adopt an iterative optimization algorithm by iteratively updating W, A, S and M. Next, we will prove that problem (6) is convex.

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

Proposition 1. Problem (6) is convex with respect to W, A, S and M.

Proof. Based on the convex optimization theory [13], [47], it can be easily justified that the first and third terms in problem (6) are convex respect to W, S, A and M ($M \geq 0$).

For the second term: (1) the jth dimension in $\|W^T x_i - A^T x_k\|_2^2$ is $(W^T x_{ij} - A^T x_{kj})^2$, whose convexity can be proved by verifying its Hessian to be positive semi-definite. (2) since Z_{ik} is nonnegative, thus the second term is also convex as a nonnegative weighted sum of convex functions is convex.

Therefore, problem (6) is convex as it is the sum of three convex terms. □

To facilitate the optimization, we rewrite problem (6) as

$$\begin{aligned}\Phi =&\|W^T X - (Y + B \odot M)\|_F^2 \\ &+ \lambda_1\{Tr(W^T X D X^T W) + Tr(A^T X D X^T A) \\ &- 2Tr(W^T X Z X^T A)\} + \lambda_2 \|W - AS\|_F^2 \\ &s.t. \quad M \geq 0\end{aligned} \tag{7}$$

where $D_{ij} = \sum_j Z_{ij}$. In the following, we introduce the proposed update rules in brief.

Update S as given W, A and M.

By setting the derivative $\partial\Phi/\partial S = 0$, we obtain

$$\lambda_2(A^T A S - A^T W) = 0 \Rightarrow S = (A^T A + \tau I)^{-1}(A^T W) \tag{8}$$

where τ is a very small positive constant which is used to obtain numerically more stable solution.

Update A as given W, S and M.

By setting the derivative $\partial\Phi/\partial A = 0$, we obtain

$$\lambda_1 X D X^T A + \lambda_2 A S S^T - \lambda_1 X Z^T X^T W - \lambda_2 W S^T = 0 \tag{9}$$

A is essentially updated by solving a Sylvester equation.

Update W as given A, S and M.

By setting the derivative $\partial\Phi/\partial W = 0$, we obtain

$$\begin{aligned}&X X^T W - X K^T + \lambda_1 X D X^T W - \lambda_1 X Z X^T A + \\ &\qquad \lambda_2 W - \lambda_2 A S = 0 \\ &\Rightarrow W = (X X^T + \lambda_1 X D X^T + \lambda_2 I)^{-1}(X K^T \\ &\qquad + \lambda_1 X Z X^T A + \lambda_2 A S)\end{aligned} \tag{10}$$

where $K = Y + B \odot M$.

Update M as given A, S and W.

M can be solved from the following optimization problem [13], [23]:

$$\min_M \|P - B \odot M\|_F^2, \quad s.t. \quad M \geq 0 \tag{11}$$

where $P = W^T X - Y$.

It is known to all that the squared Frobenius norm of a matrix can be decoupled element by element. Thus, (11) can be decoupled equivalently into $c \times n$ subproblems. For the ith row and jth column element M_{ij}, we have

$$\min_{M_{ij}} (P_{ij} - B_{ij} M_{ij})^2, \quad s.t. \quad M_{ij} \geq 0 \tag{12}$$

where P_{ij} and B_{ij} are the ijth elements of P and B, respectively. Considering the face of $B_{ij}^2 = 1$, we have $(P_{ij} - B_{ij} M_{ij})^2 = (B_{ij} P_{ij} - M_{ij})^2$. Thus, M can be calculated as follows

$$M = \max(B \odot P, 0) \tag{13}$$

In summary, the process of solving problem (6) is summarized in Algorithm 1.

Algorithm 1 : DRR

Input: Training samples matrix X; Label matrix Y;
The luxury matrix B; Class compactness graph weight matrix Z;
Parameters λ_1 and λ_2;
Output: The transformation matrix W.
Initialization: $M = \mathbf{1}_{n \times c}$; $A = W = (X X^T)^{-1}(X Y^T)$
Set $t = 0$;
repeat
1. Update S by (8).
2. Update A by (9).
3. Update W by (10).
4. Update M by (13).
3. Update $t = t + 1$.

until *Convergence*

Next, we discuss the computation complexity of Algorithm 1.

The main time-consuming component of Algorithm 1 are:

a) Matrix multiplication and inverse in solving problems (8), (9) and (10).

b) Sylvester equation in solving problem (9).

The general matrix multiplication takes $\mathcal{O}(n^3)$, and since there are k multiplications, the total time complexity of these operations is $\mathcal{O}(kn^3)$. The inverse of a $n \times n$ matrix is $\mathcal{O}(n^3)$. For problem (10), the inverse operation of matrix $X X^T + \lambda_1 X D X^T + \lambda_2 I$ can be pre-calculated before before going to the loop. The complexity of classical solution for the Sylvester equation is $\mathcal{O}(m^3)$. The final time cost for Algorithm 1 is about $\mathcal{O}(n^3 + \mathcal{T}((k+1)n^3 + m^3))$, where $\mathcal{T}$ is the number of iterations.

B. Classification

When (6) is solved, we obtain regression parameter W. We can directly use the obtained classification parameter W for classification. Suppose X_t is the test sample set, their final output results are $W^T X_t$. Then, we use the nearest-neighbor (NN) to classify them. For other classifiers, the results may be improved but is more involved. We leave it as a future work.

IV. Experiments And Analysis

In this section, we first evaluate our DRR on three widely used face data sets: 1) Extended YaleB [48], 2) CMU PIE [49] and 3) AR [29], [36]. The difficulties of these three face data sets are not the same. As shown in Figure 1, the Extended YaleB is relatively simple. For each individual, it has about 64 near frontal images under different illuminations. The CMU PIE data set is taken under different poses, expressions, and illumination conditions. Compared with the Extended YaleB data set, CMU PIE data set is more difficult to identify. The

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

challenge of AR is that it contains different facial expressions, illumination conditions, and occlusions (sun glass and scarf). We also test our method on two more different types of databases: 1) Fifteen Scene categories data set [36] for scene classification; 2) Caltech 101 and Imagenet data sets for classification with deep learning feature; 3) COIL100 objective data set [50] for objective classification. The descriptions of these five data sets are summarized in Table I.

TABLE I
DESCRIPTIONS OF 5 BENCHMARK DATA SETS

Data set	Number of samples	Dimensionality	Classes
Extended Yale B	2414	1024	38
CMU PIE	11554	1024	68
AR	2600	540	100
Fifteen Scene categories	4485	3000	15
Caltech 101	9144	4096	102
Imagenet	71990	4096	52
COIL100	7200	1024	100

(a) Extended YaleB

(b) AR

(c) CMU PIE

Fig. 1. Some face images from (a) Extended YaleB, (b) AR and (c) CMU PIE data sets.

We compare our method with SRC [29], CRC [32], the locality constrained linear coding (LLC) method [33], LRC [41], low-rank matrix recovery with structural incoherence based classification (LRSIC) [30], low-rank representation for classification (LRRC) [31], structured LRRC (SLRRC) [31], TDDL [35], SVM [40], RLR [23], Robust PCA [37], DLSR [13], latent low-rank representation (LatLRR) [34], traditional low-rank linear regression (LRLR)[57], low-rank ridge regression (LRRR) [57], sparse low-rank regression (SLRR) [57], low-rank matrix recovery method by embedding the structure incoherence (LRSI) [58], class-wise block-diagonal structure (CBDS) dictionary learning method [59], extreme learning machine (ELM) [60], random forest (RF) [60]. For fairness, Robust PCA and LatLRR first extract features by corresponding methods. Then, they use model (1) to learn the transformation matrix. Finally, they use the NN classifier to classify them. For RLR and DLSR, we also use the NN classifier for the sake of fairness. For the other methods, we use the classification methods mentioned in their papers to perform the final classification. The platform is MATLAB 2010b under Windows 7 on PC equipped with a 3.30-GHZ CPU and 8-GB memory. The MATLAB code of DRR is publicly available at http://www.yongxu.org/lunwen.html.

TABLE II
CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON THE EXTENDED YALE B DATA SET

Alg.	10	15	20	25
SRC [29]	87.8±0.3	92.6±0.6	94.4±0.6	96.7±0.5
CRC [32]	86.1±0.5	90.7±0.3	93.0±0.2	94.1±0.3
LLC [33]	79.8±0.4	88.6±0.3	91.5±0.4	94.3±0.5
LRC [41]	83.3±0.4	89.4±0.5	92.4±0.2	93.6±0.3
LRSIC [30]	87.0±0.6	92.7±0.5	94.2±0.3	96.1±0.5
LRRC [31]	84.3±0.6	91.5±0.4	93.3±0.5	95.8±0.7
SLRRC [31]	85.5±0.4	91.4±0.6	94.0±0.5	95.6±0.7
TDDL [35]	84.3±0.6	88.9±0.3	92.5±0.4	95.0±0.6
Robust PCA [37]	86.1±0.2	90.5±0.4	93.5±0.6	95.4±0.3
LatLRR [34]	84.0±0.5	88.8±0.3	92.1±0.5	93.8±0.6
SVM [40]	81.5±1.4	89.2±0.9	92.6±0.7	94.5±0.6
ELM [60]	85.5±0.2	91.2±0.6	93.7±0.4	95.2±0.5
RF [60]	83.4±0.4	88.5±0.3	91.1±0.5	94.6±0.4
RLR [23]	88.4±0.3	92.8±0.4	96.1±0.3	97.5±0.2
DLSR [13]	86.2±0.9	92.3±0.7	94.7±0.7	95.8±0.4
LRLR [57]	78.2±1.7	82.0±0.9	83.8±1.5	85.0±1.0
LRRR [57]	78.6±1.7	82.3±1.2	83.6±0.7	85.4±0.9
SLRR [57]	78.0±1.7	82.3±1.0	84.2±0.7	85.1±1.1
LRSI [58]	87.1±0.6	92.7±0.5	94.2±0.3	96.1±0.5
CBDS [59]	85.8±1.8	93.1±1.3	95.8±1.0	96.3±0.8
DRR	**90.4±0.2**	**94.9±0.7**	**97.1±0.5**	**98.2±0.4**

A. Face Recognition

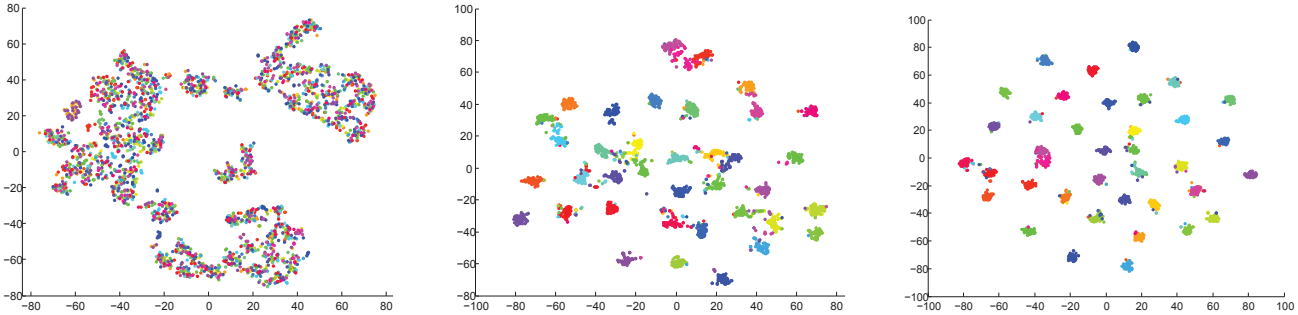
1) Extended YaleB: The Extended YaleB data set consists of 2414 cropped frontal face images of 38 peoples. There are between 59 and 64 images for each person. Every image has 32×32 pixels. We randomly select 10, 15, 20 and 25 training samples from each person for training and the remaining images for testing. Every experiment runs 30 times.

When we evaluate SRC, CRC, LRC, and LRSIC, all training samples are used as the dictionary. The number of neighbors of LLC is set to 5, which is the same as that in [33]. Following [31], the dictionary size for LRRC, SLRRC, and TDDL is all set to 140 (each person has five atoms). The experimental results are shown in Table II. Note that we report the mean classification accuracy and corresponding standard deviation (mean±std) of different methods, and the bold numbers suggest the best classification accuracies. We can see that with different numbers of training sample, DRR always achieves the best classification results.

To further prove that our method can address the problem of over-fitting by enlarging the margins between different classes, we visualize the experimental results of RLR and DRR and the original samples of the Extended Yale B data set by suing the t-SNE algorithm [56] in Figure 2 in which 5 samples per class are randomly selected as the training samples. From Figure 2, it is obvious that the transformed samples of DRR have much better separability than the original samples. Moreover, margins between different classes achieved by our method are obviously larger than those achieved by RLR, which proves that our method can compactly pull samples of the same class to their own subspace. This also proves that the use of two different transformation matrices is helpful to enlarge the margins and thus the over-fitting problem can be addressed to some extent.

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology



(a) Original samples (b) Transformed samples by RLR (c) Transformed samples by DRR

Fig. 2. t-SNE visualization of (a) original samples (b) Transformed samples by RLR [23] and (c) Transformed samples by DRR (our method). In this experiments, 5 samples per class are randomly selected as the training samples and the rest are used as testing samples. Please note that all samples (include training and testing samples) are simultaneously visualized in these figures.

TABLE III
CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON THE CMU PIE DATA SET

Alg.	10	15	20	25
SRC [29]	77.3±0.7	87.2±0.6	90.5±0.4	93.3±0.7
CRC [32]	83.3±0.9	88.1±0.6	90.4±0.7	93.1±0.8
LLC [33]	77.1±0.4	85.5±0.5	89.9±0.3	93.0±0.4
LRC [41]	79.1±0.5	84.7±0.6	88.3±0.4	93.4±0.6
LRSIC [30]	82.4±0.8	87.7±0.8	90.6±0.6	93.5±0.5
LRRC [31]	79.8±0.6	85.2±0.7	89.1±0.7	91.3±0.8
SLRRC [31]	80.9±0.5	86.0±0.5	89.9±0.4	91.8±0.3
TDDL [35]	78.4±0.6	84.4±0.7	87.9±0.6	91.0±0.8
Robust PCA [37]	80.3±0.3	84.1±0.5	87.8±0.4	90.7±0.3
LatLRR [34]	79.4±0.3	85.8±0.6	89.6±0.3	91.6±0.4
SVM [40]	77.9±1.1	86.8±0.7	90.7±0.4	93.6±0.7
ELM [60]	77.6±0.6	88.1±0.4	91.2±0.5	94.3±0.4
RF [60]	76.8±0.4	87.8±0.3	91.1±0.5	93.5±0.5
RLR [23]	89.0±0.4	92.1±0.3	93.2±0.6	94.9±0.4
DLSR [13]	86.4±0.5	90.7±0.5	92.5±0.4	94.5±0.2
LRLR [57]	79.8±1.2	83.8±0.5	86.8±0.4	87.0±0.4
LRRR [57]	79.9±1.1	83.9±0.6	87.6±0.5	88.9±0.6
SLRR [57]	78.1±0.8	84.7±0.6	87.8±0.6	89.4±0.7
LRSI [58]	83.1±0.5	87.9±0.7	90.8±0.6	93.8±0.6
CBDS [59]	82.9±1.2	89.5±0.7	91.6±0.4	93.6±0.6
DRR	**90.6±0.5**	**93.8±0.3**	**95.4±0.3**	**96.1±0.2**

2) CMU PIE: The CMU PIE data set contains 41368 images of 68 people, each with 13 different poses, 43 different illumination conditions, and 4 different expressions. In this experiment, we select a subset of PIE, which contains five near frontal poses (C05, C07, C09, C27, C29) and all the images are taken under different illuminations and expressions, to test different methods. Thus, there are 170 images for each persons. Since LLC encodes the Scale-Invariant Feature Transform (SIFT) features and we should keep a certain amount of SIFT features. Thus, in this experiment, the face images is normalized to a size of 64×64 pixels for LLC [33]. In all the other methods, the size of each image is only of 32×32. All the training samples are used as the dictionary for SRC, CRC, LRC, and LRSIC. The size of dictionary for LRRC, SLRRC and TDDL is 340. We also select different training samples per person for training and remaining for testing. The classification results are summarized in Table III in which the best results are denoted by bold numbers. Again, our method obtains the best classification results on all cases.

3) AR: The AR data set contains over 4000 color images with 126 persons (70 men and 56 women) and each provides 26 face images taken during two sessions. In each session, each person provides 13 images, in which three images with sunglasses, another three with scarfs, and the remaining seven with different facial expressions and illumination conditions. Following the standard evaluation procedure [29], we, in this experiment, use a subset consisting 2600 images from 50 male and 50 female. For each person, we randomly select 20 images for training and the other 6 for testing. Each image is projected onto a 540-dimension vector with a randomly generated matrix [36]. The experimental results are shown in Table IV. Note that some experimental results are directly cited from [36]. We set the parameters $\lambda_1 = 10^{-5}$ and $\lambda_2 = 0.2$ in our method. Our methods achieves the best experiment result.

B. Scene Classification

The Fifteen Scene Categories data set contains 15 natural scene categories that expands on the 13-category database related in [53]. This data set contains 4485 images falling into 15 categories such as bedrooms, kitchens, streets, and country scenes. Each category has 200 to 400 images. Figure 3 shows some images from this data set.

Fig. 3. Some images from the Fifteen Scene Categories data set.

We use the features of this data set provided by Jiang et al. in [36]. The features are obtained by the following steps: First, computing a spatial pyramid feature with a four-level spatial pyramid and a SIFT-descriptor codebook with size of 200. Then, PCA is applied to reduce the feature dimension to 3000 dimensions. Following the common experimental settings, we randomly select 100 images per category as training set and use remaining as testing set. The comparison results are shown

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

TABLE IV
CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON THE AR DATA SET

Alg.	Accuracy	Alg.	Accuracy
LLC (30 local bases)[33]	69.5	SRC (all train. samp.)[29]	97.5
LLC* (70 local bases)[33]	88.7	SRC* (5 per person)[29]	66.5
K-SVD (5 per person)[51]	86.5	CRC [32]	97.3
D-SVD (5 per person)[52]	88.8	LRC [41]	94.5
LC-KSVD1 (5 per person)[36]	92.5	SVM [40]	96.7
LC-KSVD2 (5 per person)[36]	93.7	RLR [23]	98.1
LC-KSVD2 (all train.samp)[36]	97.8	LatLRR [34]	97.6
ELM [60]	96.4	DLSR [13]	97.6
DRR	**98.85±0.44**		

in Table V in which SRC, CRC, LRC, DLSR, RLR, SVM, LRSIC, SLR, LRRC, SLRRC, LatLRR, Robust PCA and our method all use the spatial pyramid feature provided by in [36]. The dictionary size of SRC, CRC, LRC, LRSIC, LRRC SLRRC, and TDDL are all set to 450. The neighborhoods of LLC and LLC* are set to 30. We set the parameters $\lambda_1 = 0.001$ and $\lambda_2 = 0.005$ in our method. Our method also achieves the best result. Figure 4 shows the confusion matrix of our method obtained from the Fifteen Scene Categories database, where the classification accuracy for each class is along the diagonal. All classes are classified well and the worst classification accuracy is as high as 95%.

TABLE V
CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON THE FIFTEEN SCENE DATA SET

Alg.	Accuracy	Alg.	Accuracy
SRC[29]	91.8	TDDL [35]	92.1
CRC [32]	92.3	LatLRR [34]	91.5
LLC [33]	79.4	Robust PCA [37]	92.1
LLC* [33]	89.2	Lazebnik [27]	81.4
LRC [41]	91.9	SVM [40]	93.6
LRSIC [30]	92.4	RLR [23]	96.8
LRRC [31]	90.1	Yang [54]	80.3
SLRRC [31]	91.3	Lian [55]	86.4
Boureau [56]	84.3	LC-KSVD1 [36]	90.4
LC-KSVD2 [36]	92.9	DLSR [13]	95.9
ELM [60]	94.5	CBDS [59]	95.7
Gemert [38]	76.7	LRLR [57]	94.5
LRRR [57]	88.1	SLRR [57]	89.6
DRR	**97.88±0.23**		

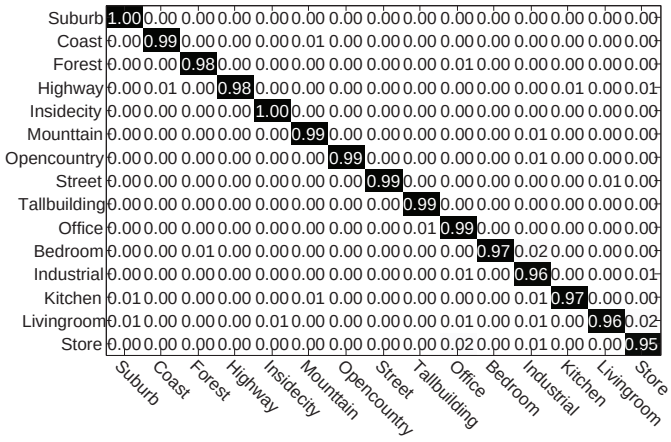


Fig. 4. Confusion matrix of our method on the Fifteen Scene Categories data set.

Fig. 5. Some images from the COIL100 data set.

C. Object Classification

The COIL100 data set contains various views of 100 objects with different lighting conditions. In our experiment, the images are converted to gray scale and resized to 32 32 pixels, and then, the robustness is evaluated on alternative viewpoints. Some image samples from this database are shown in Figure 5.

We randomly select 10, 15, 20, and 25 images per object to construct the training set, and the test set contains the rest of the images. This random selection process is repeated 30 times. We also report the mean classification accuracy and corresponding standard deviation (mean±std) of different methods, and the bold numbers suggest the best classification accuracies. The experiment results are shown in Table VI. We can see that our method outperforms the other methods. Especially, about three percentages of classification accuracy rates are improved in comparison with the rest of methods on the all cases.

D. Classification using deep learning feature

In order to test our method better, we conduct the classification experiments on the deep learning feature. We use Caltech 101 and Imagenet databases to test the classification performance of our method with deep learning feature. The deep learning features of Caltech 101(DeCAF-6) and Imagenet (DeCAF-7) are available at https://sites.google.com/site/crossdataset/home/files. Since the dimension of original feature is very high, we use PCA as a preprocessing step to preserve 98% energy of those two databases. For Caltech101 database, we randomly select 10, 15, 20, 25 and 30 samples per class for training and remaining samples for testing and we report the mean classification results over 10 random splits. Fig. 6 plots the mean classification accuracies (%) of different methods in which our method achieves the best classification results. For Imagenet database, we randomly select 71990 samples of 52 classes in our method. We randomly select

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VI
CLASSIFICATION ACCURACIES (%) OF DIFFERENT METHODS ON THE COIL100 DATA SET

Alg.	10	15	20	25
SRC [29]	80.4±0.6	86.1±0.8	89.4±0.4	91.9±0.4
CRC [32]	76.2±0.6	81.3±0.4	84.2±0.5	86.3±0.5
LLC [33]	81.6±0.8	86.9±0.4	90.2±0.4	92.5±0.5
LRC [41]	79.9±0.7	85.3±0.6	88.7±0.7	91.0±0.5
LRSIC [30]	82.3±0.3	87.7±0.4	90.1±0.6	91.2±0.3
LRRC [31]	82.7±0.5	87.0±0.6	90.2±0.4	91.8±0.5
SLRRC [31]	83.2±0.2	86.8±0.3	90.5±0.6	91.2±0.6
TDDL [35]	83.3±0.6	87.9±0.3	90.8±0.4	90.0±0.7
Robust PCA [37]	82.5±0.6	88.3±0.8	91.7±0.3	93.5±0.3
LatLRR [34]	79.6±0.5	85.3±0.4	88.4±0.4	90.7±0.4
SVM [40]	79.2±0.5	84.8±0.6	88.1±0.4	90.8±0.6
ELM [60]	81.2±0.4	85.6±0.7	89.7±0.4	92.1±0.6
RF [60]	84.3±0.5	88.3±0.5	91.1±0.5	93.3±0.5
RLR [23]	80.1±0.6	83.4±0.7	85.9±0.8	87.2±0.6
DLSR [13]	84.8±0.5	88.0±0.5	90.1±0.3	92.0±0.4
LRLR [57]	66.2±0.8	71.2±0.6	73.7±0.8	75.7±0.7
LRRR [57]	67.7±0.5	71.4±0.6	73.6±0.8	75.5±0.8
SLRR [57]	69.1±0.8	73.0±0.6	74.5±0.6	75.9±0.7
LRSI [58]	79.7±0.5	87.8±0.3	91.4±0.4	93.6±0.6
CBDS [59]	73.7±0.5	78.6±0.8	80.9±0.7	81.3±0.5
DRR	**86.2±1.1**	**90.1±0.4**	**94.0±0.4**	**95.2±0.6**

5, 10, 15, 20, 25 and 30 samples per class for training and remaining samples for testing. The mean classification results over 10 random splits is plotted in Fig. 7 in which our method also achieves the best performance. Therefore, our method has good applicability to all kinds of features.

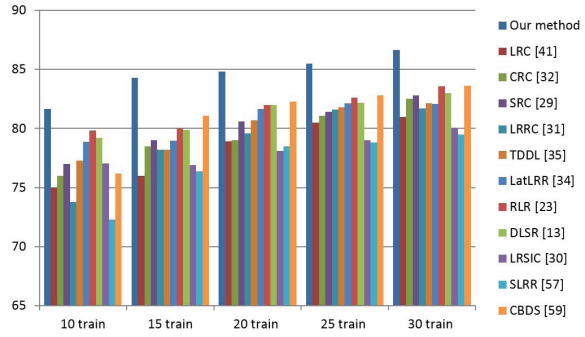


Fig. 6. Classification accuracies (%) on the deep learning features of the Caltech 101 database, in which X-axis represents the different number of training samples and Y -axis denotes the classification accuracy (%).

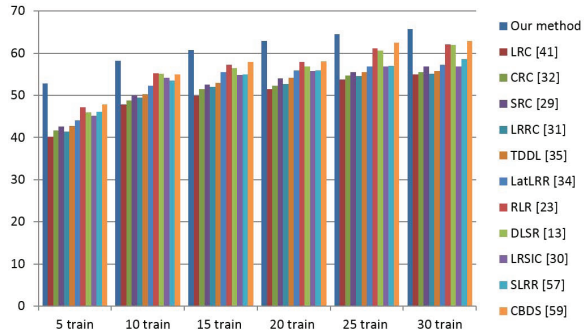
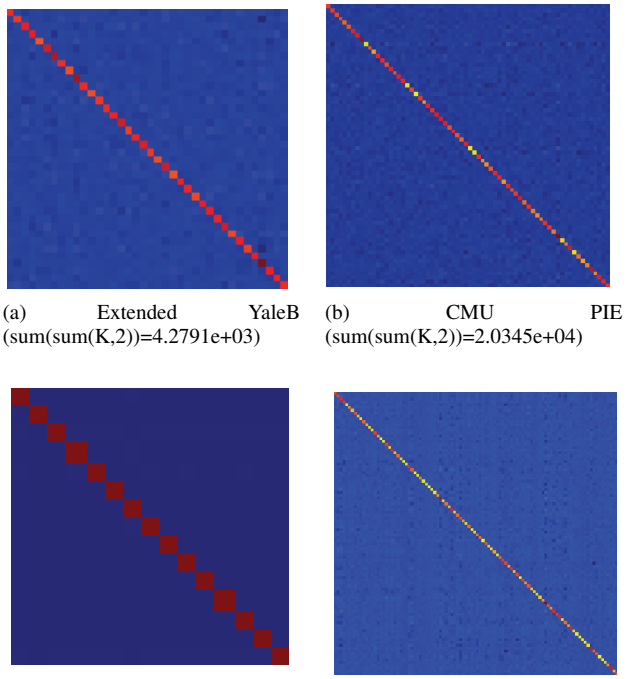


Fig. 7. Classification accuracies (%) on the deep learning features of the Imagenet database, in which X-axis represents the different number of training samples and Y -axis denotes the classification accuracy (%).



(a) Extended YaleB (sum(sum(K,2))=4.2791e+03)
(b) CMU PIE (sum(sum(K,2))=2.0345e+04)
(c) Fifteen Scene Categories (sum(sum(K,2))=113.4173)
(d) COIL100 (sum(sum(K,2))=2.1356e+04)

Fig. 8. Visualization of matrix S learned from the (a) Extended YaleB, (b) CMU PIE, (c) Fifteen Scene Categories and (d) COIL100 data sets. In each figure, the definition of K is $|W - A|$, where $|\cdot|$ represents the absolute value operation.

E. Structure analysis of S

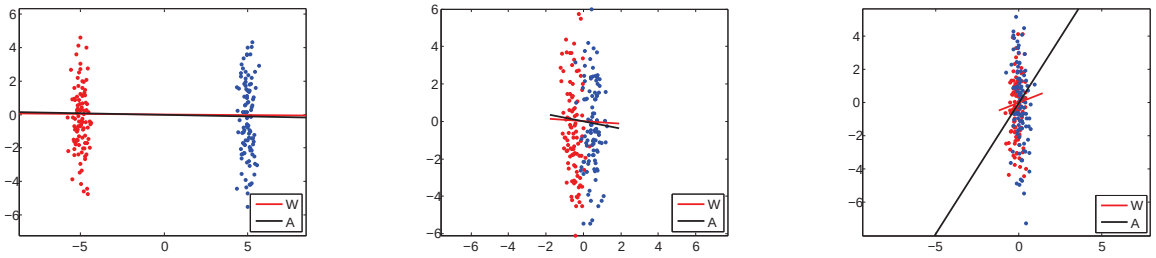
To capture the similar structure of data shared by these two transformation matrices W and A, we introduce matrix S to model some correspondences between W and A by using $\|W - AS\|_F^2$ (or called *structure-consistency*). To show such similar structure between these two matrices, we give the visualization of S on the Extended YaleB, CMU PIE, Fifteen Scene Categories and COIL100 data sets in Figure 8. From these results in Figure 8, we can see that S has block diagonal structure. In other words, elements in matrices W and A have some correspondences between classes sharing the same labels. Such correspondence is the so-called similar structure of data shared by W and A, which is very useful to guarantee that the samples sharing the same labels can be close together as much as possible. However, such similar structure does not mean W and A are completely equal. In Figure 8, we also give the distinction between them by defining $K = |W - A|$, where $|\cdot|$ represents the absolute value operation. The values of sum of elements of K are also given in Figure 8 from which we can see that W and A are completely different matrices. This indicates that matrix A indeed can share part of the responsibility of W for learning better margins.

F. Experiments on Synthetic Data

The data set for DRR is a randomly generated two-Gaussian data. In this data set, there are two classes of data which obeys the Gaussian distribution. Our goal is to find a good projection direction, i.e., W which helps to classify the two

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology



(a) Classes far away (b) Classes relatively close (c) Classes fairly close

Fig. 9. Results on the two-Gaussian synthetic data.

classes apart. The comparison results are displayed in Figure 9. Seen from Figure 9, we can find out that when these two classes are far from each other, matrices W and A are good projection directions. In this case, these matrices are nearly the same because these two classes are too far from each other. However, as the distance between these two classes lower down, A and W are still discriminative and W is still work better in classifying them. However, we also observe that A is less discriminative than W. Please note that the more parallel the transformation matrix is, the more discriminative the transformation matrix is. As the two classes become more close (fairly close), W still behaves well. The main reason is that W is closely related to the labels of training samples. In other words, W pays more attention to well classify these two classes. However, A is less discriminative. The reason may be that A sacrifices part of its discriminant ability to make W fits $Y^\circ = Y + B \odot M$ well as much as possible, i.e., the margins between different classes are enlarged as much as possible. This also means that samples sharing the same labels are close together when they are transformed into their label space. From these results in Figure 9, we can find that W keeps its high quality of projection direction all the time. There results in Figure 9 also indicates that when the classification task is difficult, matrices W and A are two completely different matrices but have similar structures from another respect, which is consistent with the results in Figure 8.

G. Experiment Analysis

From the experimental results listed in Table II, Table III, Table IV, Table V and Table VI, we have the following observations and corresponding analyses.

1) Our method performs better than all the other methods on all cases. Especially, compared with RLR, our method show its power in improving classification accuracy. For example, on the COIL100 database, when we randomly select 10 samples per class as training set, the classification accuracy of our method is about 6% higher than RLR. This indicates that the use of double transformation matrices indeed provides more freedom for solving the problem of over-fitting and achieves better margins and thus the performance is enhanced.

2) Our method and RLR usually perform better than the classical DLSR in image classification. Thus, the elimination of over-fitting problem is very important for enhancing the classification accuracy and makes algorithm has more generalization ability.

3) Although some representation based image classification methods such as sparse and low-rank representation based methods obtain good classification results, their performance is still inferior to our method. The reason may be that these methods are only to find the sparse or low-rank representation for data reconstruction. However, the best data reconstruction does not represent the best discriminate power.

4) The results in Figure 8 and Figure 9 confirm that the use of two different matrices can indeed address the problem of over-fitting more better than RLR. Thus, our method is competitive and can obtain better classification results.

H. Parameter Sensitiveness

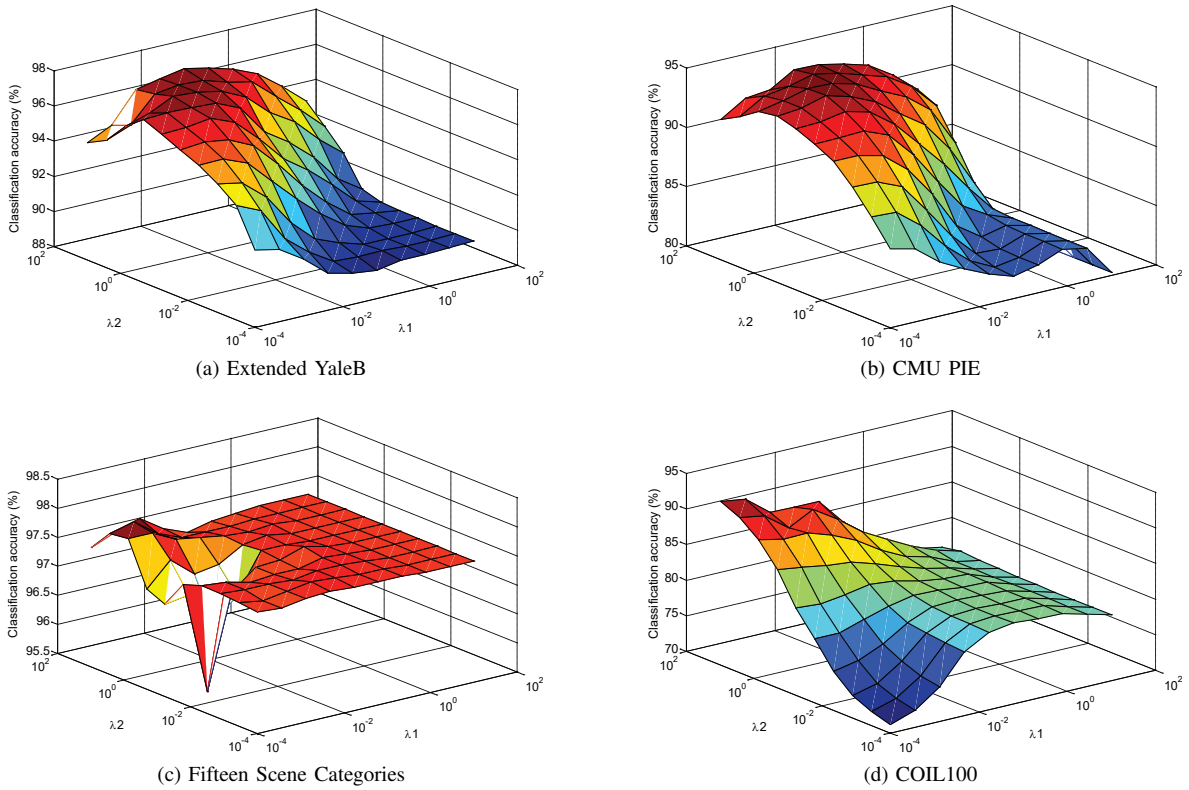
DRR requires two parameters λ_1 and λ_2 to be set in advance. In this subsection, their sensitivity is discussed. The classification accuracies variation with different parameters are plotted in Figure 10. It can be seen that the performance changes are different with respect to different data sets. However, the best classification results are always achieved with large λ_2 and small λ_1. Through tuning the parameters λ_1 and λ_2, it can be observed that the best results were achieved on the given data sets when $\lambda_1 \in [10^{-4}, 10^{-1}]$ and $\lambda_2 \in [10^{-2}, 10^{2}]$. When the value of λ_2 is too small such as $\lambda_2 \leq 10^{-3}$, the performance of DRR is very bad. This demonstrates that the term in (6) corresponding to λ_2 is more significant to learn the ideal transformation matrix W. Specifically, *structure-consistency* in our method plays a significant role in effectively address the problem of over-fitting. The value of λ_1 is small, which means that the effect of the class compactness graph is relatively less than the *structure-consistency* in our method. How to identify the optimal values of these parameters is data set dependent and still an open problem, which will be studied in our future work. In our experiments, λ_1 is firstly fixed in advance and an attempt is made to find a candidate interval where the optimal parameter λ_2 may exist. Then, by fixing the value of λ_2 in the candidate interval, the candidate interval of λ_1 is determined. Finally, the optimal parameters in the 2D candidate space of (λ_1 and λ_2) with a fixed step length are searched.

I. Convergence Study

We run our method on four data sets and plot the convergence curves of objective function values and classification

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology



(a) Extended YaleB (b) CMU PIE

(c) Fifteen Scene Categories (d) COIL100

Fig. 10. Classification accuracies (%) variation different values of parameters λ_1 and λ_2 on the (a) Extended YaleB, (b) CMU PIE, (c) Fifteen Scene Categories and (d) COIL100 data sets. For the Extended YaleB and CMU PIE databases, we randomly select 20 samples per class as training set and the remaining are used as test set. For the Fifteen Scene Categories database, we randomly select 100 samples as training samples and remaining are used as test set. For the COIL100 database, we randomly select 15 samples per class as training set and the remaining are used as test set.

accuracies with respect to the number of iterations in Figure 11. It is easy to see that the objective function value decreases as the number of iterations increases on the given data sets which indicates that our method has a good convergence property. We also can see that the proposed optimization algorithm converges fast, say 20 iterations. The curves of classification accuracies of our method on four data sets show that the classification accuracies finally reach a summit, which also confirms that the convergence property of our method is good from another respect.

J. Comparison of training and test time

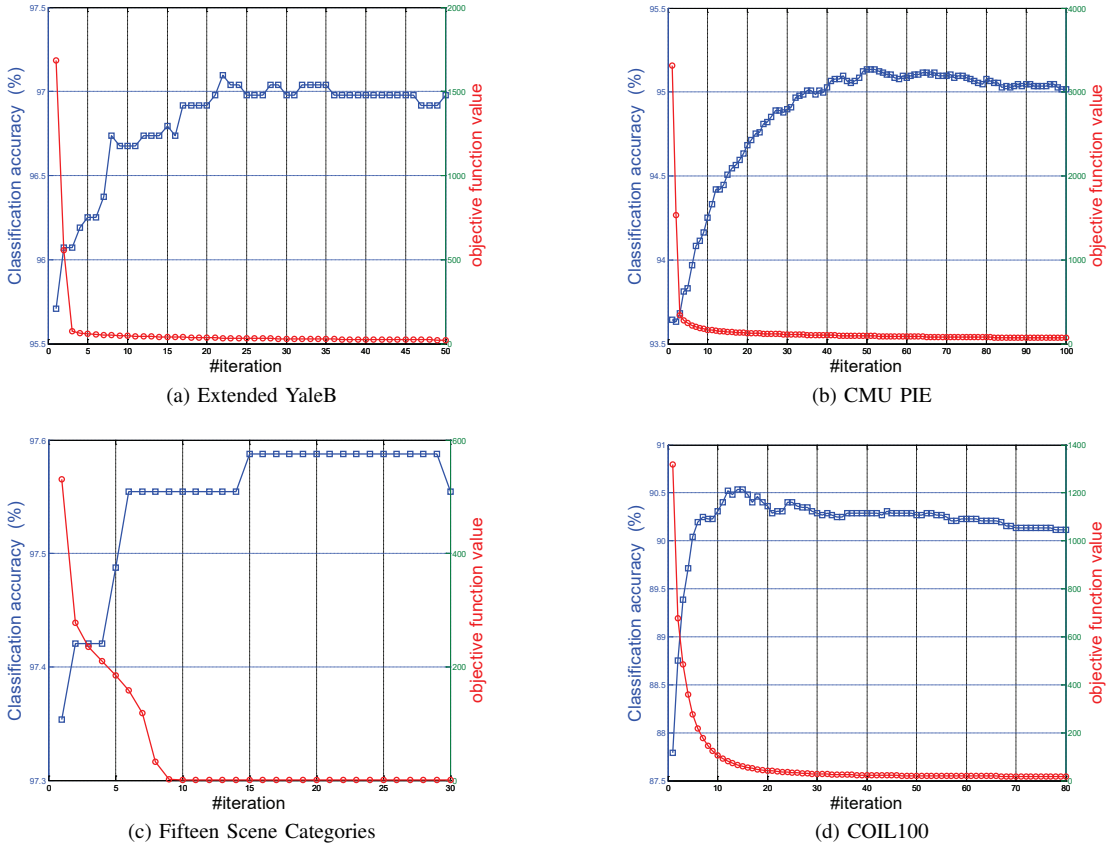
In this section, we compare the running time of DRR with those of SRC, LRRC, LRC, LLC, ELM and DLSR. Most of methods need complete training and test phases except many representation based methods such as SRC, LRC and LLC. For example, DLSR, RLR and our method need to learn a transformation matrix in the training phase and then use a linear classifier to classify test samples in the test phase. However, SRC, LRC and LLC have no training time and only have test time, since they only need to represent input test samples as a linear combination of dictionary items, then use the representation coefficients for classification. So, we respectively give the training time and test time of different methods. Table VII shows the training and test time of different methods on five data sets. We can see that the running speed of DRR is significantly faster than the representation based methods in training phase since representation based methods require a lot of time to solve an optimization problem. Especially, LRRC spends too much time to solve the rankness minimization problem and thus the training speed is too slow. Please note that LRRC is also a representation based method. As for the test time, we also find that SRC, LRC and LLC spend much time in classifying a test sample because they all need to solve the reconstruction and classification problems in test phase. DRR only needs to solve a group of linear equations in each iteration which has linear time complexity. Therefore, DRR is faster than other methods. In test phase, the running speeds of DLSR, RLR and DRR are similar since these methods use a linear classifier to classify test samples. Please note that the average test time is time cost to classify a test sample. However, classification accuracies of DLSR and RLR on these five data sets are lower than that of our method.

V. Conclusions

This paper proposes a new method called double relaxed regression (DRR) for image classification. The proposed DRR uses two different transformation matrices to address the problem of over-fitting and transform samples, which leads to two advantages 1) It provides more freedom to transform the samples into the relaxed label matrix well such that the margins between different classes are be enlarged as much as possible. 2) It well solves the problem of over-fitting. Moreover, to capture the similar structure between these two

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology



(a) Extended YaleB (b) CMU PIE

(c) Fifteen Scene Categories (d) COIL100

Fig. 11. Convergence curves of classification accuracy (%) and objective function versus iterations on the (a) Extended YaleB, (b) CMU PIE and (c) Fifteen Scene Categories and (d) COIL100 data sets. The selection of training and test samples is as the same as that in Figure 10.

TABLE VII
TRAINING TIME (S) + TEST TIME (S) ON FIVE DATA BASES (TR# AND TE# RESPECTIVELY REPRESENT TRAINING (ALL TRAINING SAMPLES) TIME AND TEST TIME (EACH TEST SAMPLE))

Alg.	CMU PIE (20)		Extended YaleB (20)		COIL100 (20)		Scene15 (100)		AR (20)	
	Tr#	Te#	Tr#	Te#	Tr#	Te#	Tr#	Te#	Tr#	Te#
SRC	none	6.3099	none	4.6932	none	6.4026	none	7.9040	none	5.5608
LRRC	89.6705	1.2004	58.8583	1.4309	70.4606	1.2056	430.0346	1.3350	33.5033	1.1128
LRC	none	0.1820	none	0.1950	none	1.2040	none	1.5030	none	0.7963
LLC	none	0.0195	none	0.0188	none	0.183	none	0.1137	none	0.1092
ELM	66.8789	0.0188	27.8910	0.0165	48.2645	0.0110	356.0056	0.0184	19.8743	0.0142
DLSR	62.6086	0.0062	23.2256	0.0073	42.7940	0.0089	348.0753	0.0055	12.5624	0.0099
RLR	58.6836	0.0038	21.6950	0.0017	38.2213	0.0072	355.2573	0.0045	10.3137	0.0050
DRR	61.5268	0.0065	22.1053	0.0033	41.1999	0.0096	351.6887	0.0060	11.3569	0.0097

transformation matrices, we propose to use the *structure-consistency* term which makes these two transformation matrices capture the resemblance between classes sharing the same labels. In this way, the samples from the same class are close together when they are transformed into the relaxed label space and thus better margins can be obtained. Promising results on seven data sets demonstrate the effectiveness of the proposed method. In our experiments, we show the comparison of running time of different methods. However, we note that the number of training samples is relatively small and thus the running speed is fast. The computation complexity $\mathcal{O}$ of DRR is relatively high ($\mathcal{O} \propto n^3$) if the size of training samples is large. Therefore, the limitation of DRR is the problem of scalability which imposes a challenging on DRR when handing large-scale samples. How to make DRR easily scale to large data is our future work. In addition, we plan to extend our method to other applications of data fitting and feature extraction (e.g., feature selection by imposing $\ell_{2,1}$-norm constraint on transformation matrix W), and so on.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of China under Grant 61772141, in part by the National Key R&D Program of China under Grant 2018YFB1003201, in part by the Guangdong Natural Science Foundation under Grants 2018B030311007, and Major R&D Project of Educational Commission of Guangdong under Grant No. 2016KZDXM052 in part by the Guangdong Provincial Natural

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

Science Foundation, under Grant 17ZK0422and in part by the Guangzhou Science and Technology Planning Project under Grants 201804010347, 201604046017 and 201604020145.

REFERENCES

[1] B. Matei and P. Meer, "Estimation of nonlinear errors-invariables models for computer vision applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1537-1552, 2006.

[2] P. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, pp. 138-156, 2000.

[3] J. Gillard and T. Iles, "Method of moments estimation in linear regression with errors in both variables," *Cardiff University, School of Mathematics*, TR, 2005.

[4] S. Choi, T. Kim, and W. Yu, "Performance Evaluation of RANSAC Family," in *BMVC*, 2009.

[5] D. Huang, R. Cabral, and F. D. Torre, "Robust Regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 363-375, 2016.

[6] T. Strutz, "Data fitting and uncertainty: a practical introduction to weighted least squares and beyond," Wiesbaden, Germany: Vieweg, 2010.

[7] S. Wold, H. Ruhe, H. Wold, and W. Dunn, III, "The collinearity problem in linear regression, the partial least squares approach to generalized inverse," *J. Sci. Stat. Comput.*, vol.5, no. 3, pp. 735-743, Jan, 1984.

[8] Y. F. Li and A. Ngom, "Nonnegative least-squares methods for the classification of high-dimensional biological data," *IEEE transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 447-456, Apr, 2013

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," emphTechnometrics, vol. 12, no. 1, pp. 55-67, 1970.

[10] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267C288, 1996.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273C297, 1995.

[12] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, "Applied Logistic Regression," New York, NY, USA: Wiley, 2013.

[13] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738-1754, Nov. 2012.

[14] F. P. Nie, H. Wang, H. Huang, and C. Ding, "Adaptive loss minimization for semi-supervised elastic embedding," *in proceedings of the twenty-third international joint conference on artificial intelligence,* Beijing, China, 2013, pp. 1565-1571.

[15] M. Belkin, P. Niyogi and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labled and Unlabeled Examples," *Journal of Machine Learning Research,* vol. 12, pp. 2399-2434, May, 2006.

[16] T. K. Paul and T. Ogunfunmi, "Study of the convergence behavior of the complex kernel least mean square algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1349C1363, Sep. 2013.

[17] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *in Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1177-1184.

[18] F. De la Torre, "A least-squares framework for component analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 6, pp. 1041-1055, Jun. 2012.

[19] X. Y. Zhang, L. F. Wang, S. M. Xiang, and C. L. Liu, "Retargeted Least Squares Regression Algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2206-2213, 2015.

[20] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "Coordinate descent method for large-scale L2-loss linear support vector machines," *J. Mach. Learn. Res.*, vol. 9, pp. 1369-1398, Jan. 2008.

[21] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265-292, 2002

[22] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1-8.

[23] X. Z. Fang, Y. Xu, X. L. Li, Z. H. Lai, W. K. Wong, and B. W. Fang, "Regularized Label Relaxation Linear Regression," *IEEE Trans. Neural Netw. Learn. Syst.*, DOI: 10.1109/TNNLS.2017.2648880, 2017.

[24] X. Z. Fang, Y. Xu, X. L. Li, Z. H. Lai, and W. K. Wong, "Learning a nonnegative sparse graph for linear regression," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2760-2771, Sep. 2015.

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

[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality constrained linear coding for image classification," *in Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3360-3367.

[27] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *in Proc. 19th IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 2169-2178.

[28] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," *in Proc. 28th Int. Conf. Mach. Learn.*, Washington, DC, USA, Jun. 2011, pp. 921-928.

[29] 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.

[30] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," *in Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2618-2625.

[31] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," *in Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 676-683.

[32] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang. (2012). "Collaborative representation based classification for face recognition." [Online]. Available: http://arxiv.org/abs/1204.2358.

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality constrained linear coding for image classification," *in Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3360C3367.

[34] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," *in Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1615-1622.

[35] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791-804, Apr. 2012.

[36] 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.

[37] E. J. Cands, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" J. ACM, vol. 58, no. 3, 2011, Art. ID 11.

[38] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," *in Proc. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 2008, pp. 696-709.

[39] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271-1283, Jul. 2010.

[40] S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel SVMs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 66-77, Jan. 2013

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

[42] Bertsekas D P, "Nonlinear programming," Athena Scientific, 1999.

[43] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664-676, 2017.

[44] D.-H. Lee, "Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks," *In International Conference on Machine Learning Workshop*, 2013.

[45] Z. Li, J. Tang, T. Mei, "Deep Collaborative Embedding for Social Image Understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, DOI 10.1109/TPAMI.2018.2852750, 2018.

[46] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep matrix factorization method for learning attribute representations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 417-429, 2017.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. of Advances in Neural Information Processing Systems*, 2012.

[48] 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. 643C660, Jun. 2001.

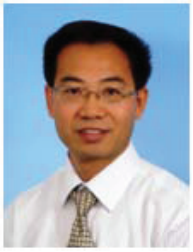
[49] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615C1618, Dec. 2003.

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2890511, IEEE Transactions on Circuits and Systems for Video Technology

[50] Nene, Sameer A., Shree K. Nayar, and Hiroshi Murase, "Columbia object image library (COIL-100)," Technical Report CUCS-005-96, 1996.
[51] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Signal Processing*, vol. 54, no. 1, pp. 4311-4322, Nov. 2006.
[52] Q. Zhang and B. Li, "Discriminative K-SVD for Dictionary Learning in Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
[53] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *in Proc. 18th IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 524-531.
[54] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *in Proc. 22nd IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1794-1801.
[55] X.-C. Lian, Z. Li, B.-L. Lu, and L. Zhang, "Max-margin dictionary learning for multiclass image categorization," *in Proc. 11th Eur. Conf. Comput. Vis.*, Barcelona, Spain, Sep. 2010, pp. 157-170.
[56] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," *in Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2559C2566.
[57] X. Cai, C. Ding, F. Nie, H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1124-1132, 2013.
[58] C. Wei, C. Chen, Y. Wang, "Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition," IEEE Trans. Image Process., vol. 23, no. 8, pp. 3294-3307, 2014
[59] Y. Li, J. Liu, H. Lu, et al., "Learning Robust Face Representation With Classwise Block-Diagonal Structure," IEEE Trans. Inf. Forensics Secur., vol. 9, no. 12, pp. 2051-2062, 2014.
[60] M. Fernandez-Delgado, E. Cernadas, S. Barro, "Do we need hundreds of classifiers to solve real world classification problem?" Journal of Machine Learning Research, vol. 15, pp. 3133-3181, Oct, 2014.

Na Han received her B.S.degree in computer science and technology at HIT in 2004. She is currently pursuing the Ph.D.degree in computer science and technology with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and machine learning.



Jigang Wu received B.Sc. degree from Lanzhou University, China in 1983, and doctoral degree from the University of Science and Technology of China in 2000. He was with the Center for High Performance Embedded Systems, Nanyang Technological University, Singapore, from 2000to 2010, as a research fellow. He was Dean, Tianjin distinguished professor of School of Computer Science and Sofware, Tianjin Polytechnic University, China, from 2010 to 2015. Now he is distinguished professor of School of Computer Science and Technology, Guangdong University of Technology. He has published more than 200 papers in IEEE TOC, TPDS, TVLSI, TNNLS, TSMC JPDC, PARCO, JSA, and international conferences. His research interests include network computing, cloud computing, machine intelligence, reconfigurable architecture. He is a member of the IEEE. He serves in China Computer Federation as technical committee member in the branch committees, High Performance Computing, Theoretical Computer Science, and Fault Tolerant Computing.

Xiaozhao Fang (S'15-M'17) received his M.S.degree in 2008, and the Ph.D.degree in computer science and technology at Shenzhen Graduate School, HIT, Shenzhen (China) in 2016. He is currently with the School of Computer Science and Technology, Guangdong University of Technology. His current research interests include pattern recognition and machine learning.

W. K. Wong received his Ph.D. degree from The Hong Kong Polytechnic University. Currently, he is with Institute of Textiles & Clothing, The Hong Kong Polytechnic University, Hong Kong and The Hong Kong Polytechnic University Shenzhen Research Institute. He has published more than fifty scientific articles in refereed journals, including IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition, International Journal of Production Economics, European Journal of Operational Research, International Journal of Production Research, Computers in Industry, IEEE Transactions on Systems, Man, and Cybernetics, among others. His recent research interests include artificial intelligence, pattern recognition, and optimization of manufacturing scheduling, planning and control.

Yong Xu (M'06-SM'15) was born in Sichuan,China,in1972.He received his B.S.degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D.degree in Pattern Recognition and Intelligence system at NUST(China) in 2005. Now he works at Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning and video analysis.

Jian Yang (M'06) received the B.S. degree in mathematics from Xuzhou Normal University, Xuzhou, China, in 1995, the M.S. degree in applied mathematics from Changsha Railway University, Changsha, China, in 1998, and the Ph.D. degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), Nanjing, China, in 2002.

In 2003, he was a Post-Doctoral Researcher with the University of Zaragoza, Zaragoza, Spain. From 2004 to 2006, he was a Post-Doctoral Fellow with the Biometrics Centre, Hong Kong Polytechnic University, Hong Kong. From 2006 to 2007, he was a Post-Doctoral Fellow with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the School of Computer Science and Technology, NUST. He has authored over 80 scientific papers in pattern recognition and computer vision. He has over 2000 ISI Web of Science and 4000 Google Scholar citations. His current research interests include pattern recognition, computer vision, and machine learning.

Dr. Yang was a recipient of the RyC Program Research Fellowship sponsored by the Spanish Ministry of Science and Technology, in 2003.

Xuelong Li (M'02-SM'07-F'12) is a full professor with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xian 710072, P. R. China.

1051-8215 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.