Learning Structured Low-rank Representations for Image Classification

Yangmuzi Zhang, Zhuolin Jiang, Larry S. Davis
University of Maryland, College Park, MD, 20742
{ymzhang,zhuolin,lsd}@umiacs.umd.edu

Abstract

An approach to learn a structured low-rank representation for image classification is presented. We use a supervised learning method to construct a discriminative and reconstructive dictionary. By introducing an ideal regularization term, we perform low-rank matrix recovery for contaminated training data from all categories simultaneously without losing structural information. A discriminative low-rank representation for images with respect to the constructed dictionary is obtained. With semantic structure information and strong identification capability, this representation is good for classification tasks even using a simple linear multi-classifier. Experimental results demonstrate the effectiveness of our approach.

1. Introduction

Recent research has demonstrated that sparse coding (or sparse representation) is a powerful image representation model. The idea is to represent an input signal as a linear combination of a few items from an over-complete dictionary D. It achieves impressive performance on image classification [29, 27, 3, 9]. Dictionary quality is a critical factor for sparse representations. The sparse representation-based coding (SRC) algorithm [27] takes the entire training set as dictionary. However, sparse coding with a large dictionary is computationally expensive. Hence some approaches [1, 27, 20, 23] focus on learning compact and discriminative dictionaries. The performance of algorithms like image classification is improved dramatically with a well-constructed dictionary and the encoding step is efficient with a compact dictionary. The performance of these methods deteriorates when the training data is contaminated (*i.e.*, occlusion, disguise, lighting variations, pixel corruption). Additionally, when the data to be analyzed is a set of images which are from the same class and sharing common (correlated) features (e.g. texture), sparse coding would still be performed for each input signal independently. This does not take advantage of any structural information in the set.

Low-rank matrix recovery, which determines a low-rank data matrix from corrupted input data, has been successfully applied to applications including salient object detection [24], segmentation and grouping [35, 13, 6], background subtraction [7], tracking [34], and 3D visual recovery [13, 31]. However, there is limited work [5, 19] using this technique for multi-class classification. [5] uses low-rank matrix recovery to remove noise from the training data class by class. This process becomes tedious as the class number grows, as in face recognition. Traditional PCA and SRC are then employed for face recognition. They simply use the whole training set as the dictionary, which is inefficient and not necessary for good recognition performance [12, 33]. [19] presents a discriminative low-rank dictionary learning for sparse representation (DLRD_SR) to learn a low-rank dictionary for sparse representation-based face recognition. A sub-dictionary D_i is learned for each class independently; these dictionaries are then combined to form a dictionary $D = [D_1, D_2, ...D_N]$ where N is the number of classes. Optimizing sub-dictionaries to be low-rank, however, might reduce diversity across items within each sub-dictionary. It results in a decrease of the dictionary's representation power.

We present a discriminative, structured low-rank framework for image classification. Label information from training data is incorporated into the dictionary learning process by adding an ideal-code regularization term to the objective function of dictionary learning. Unlike [19], the dictionary learned by our approach has good reconstruction and discrimination capabilities. With this high-quality dictionary, we are able to learn a sparse and structural representation by adding a sparseness criteria into the low-rank objective function. Images within a class have a low-rank structure, and sparsity helps to identify an image's class label. Good recognition performance is achieved with only one simple multi-class classifier, rather than learning multiple classifiers for each pair of classes [28, 21, 20]. In contrast to the prior work [5, 19] on classification that performs low-rank recovery class by class during training, our method processes all training data simultaneously. Compared to other dictionary learning methods [12, 33, 27, 25] that are very sensitive to noise in training images, our dictionary learning algorithm is robust. Contaminated images can be recovered during our dictionary learning process. The main contributions of this paper are:

- We present an approach to learn a structural low-rank and sparse image representation. By incorporating image class information, this approach encourages images from the same class to have similar representations. The learned representation can be used for classification directly.
- We present a supervised training algorithm to construct

a discriminative and reconstructive dictionary, which is used to obtain a low-rank and sparse representation for images.

- The algorithm computes a low-rank recovery for all training samples simultaneously while preserving independence across different classes in a computationally efficient manner.
- Our image classification framework is robust. It outperforms state-of-the-art methods even when training and testing data are badly corrupted.

1.1. Related Work

Sparse representation has been widely used for image classification. [26] has shown that sparse representation achieves impressive results on face recognition. The entire training set is taken as the dictionary. [29, 30] formulate a sparsity-constrained framework to model the sparse coding problem. They use a modified model to handle corruptions like occlusion in face recognition. These algorithms, however, don't learn a dictionary. The selection of the dictionary, as shown in [8], can strongly influence classification accuracy. One of the most commonly used dictionary learning method is K-SVD [1]. This algorithm focuses on the representation power of dictionaries. Several algorithms have been developed to make the dictionary more discriminative for sparse coding. In [23], a dictionary is updated iteratively based on the results of a linear predictive classier to include structure information. [12] presents a Label Consistent K-SVD (LC-KSVD) algorithm to learn a compact and discriminative dictionary for sparse coding. These methods show that performance is improved dramatically with a structured dictionary. However, if the training data is corrupted by noise, their performance is diminished.

Using low-rank matrix recovery for denoising has attracted much attention recently. Wright introduced the Iterative Thresholding Approach [26] to solve a relaxed convex form of the problem. The Accelerated Proximal Gradient Approach is described in [16, 26]. The Dual Approach in [16] tackles the problem via its dual. Applying augmented Lagrange multipliers (ALM), Lin [15] proposed RPCA via the Exact and Inexact ALM Method. Promising results have been shown in many applications [24, 35, 13, 6, 34]. Limited work, however, has applied the low-rank framework to solve image classification problems. [5] uses a low-rank technique to remove noise from training data. Denoising is implemented class by class, which gives rise to tremendous computational cost as class number increases. [19] enhances a sparse coding dictionary's discriminability by learning a low-rank sub-dictionary for each class. This process is time-consuming and might increase the redundancy in each sub-dictionary, thus not guaranteeing consistency of sparse codes for signals from the same class. [31] presents an image classification framework by using non-negative sparse coding, low-rank and sparse matrix decomposition. A linear SVM classifier is used for the final classification.

Compared to previous work, our approach effectively constructs a reconstructive and discriminative dictionary from corrupted training data. Based on this dictionary, structured low-rank and sparse representations are learned for classification.

2. Low-rank Matrix Recovery

Suppose a matrix X can be decomposed into two matrices, *i.e.*, $X = A + E$, where A is a low-rank matrix and E is a sparse matrix. Low-rank matrix recovery aims at finding A from X. It can be viewed as an optimization problem: decomposing the input X into $A + E$, minimizing the rank of A and reducing $||E||_0$.

$$\min_{A,E} rank(A) + \lambda ||E||_0 \quad s.t\ X = A + E \tag{1}$$

where λ is a parameter that controls the weight of the noise matrix E. However, direct optimization of (1) is NP-hard. [4] shows that if the rank of A is not too large and E is sparse, the optimization problem is equivalent to:

$$\min_{A,E} ||A||_* + \lambda ||E||_1 \quad s.t\ X = A + E \tag{2}$$

where $||A||_*$ is the nuclear norm (i.e., the sum of the singular values) of A. It approximates the rank of A. $||E||_0$ could be replaced with the l_1-norm $||E||_1$. As proved in [4], low-rank and sparse components are identifiable. Under fairly general conditions, A can be exactly recovered from X as long as E is sufficiently sparse (relative to the rank of A) [26]. This model assumes that all vectors in X are coming from a single subspace. [5] uses this technique to remove noise from training samples class by class; this process is computationally expensive for large numbers of classes. Moreover, structure information is not well preserved. [5] solves this problem by promoting the incoherence between different classes. A regularization term $\eta \sum_{j \neq i} ||A_j^T A_i||_F^2$ is added to function (2). It needs to be updated whenever A_j is changed. This is complicated and might not be helpful for classification.

Consider the problem of face recognition. Here, the dataset is a union of many subjects; samples of one subject tend to be drawn from the same subspace, while samples of different subjects are drawn from different subspaces. [18] proves that there is a lowest-rank representation that reveals the membership of samples. A more general rank minimization problem [18] is formulated as:

$$\min_{Z,E} ||Z||_* + \lambda ||E||_{2,1} \tag{3}$$
$$s.t\ X = DZ + E$$

where D is a dictionary that linearly spans the data space. The quality of D will influence the discriminativeness of the representation Z. [18] employs the whole training set as the dictionary, but this might not be efficient for finding a discriminative representation in classification problems.

[19] tries to learn a structured dictionary by minimizing the rank of each sub-dictionary. However, it reduces diversity in sub-dictionary, weakening the dictionary's representation power.

We will show that an efficient representation can be obtained with respect to a well-structured dictionary. Associating label information in the training process, a discriminative dictionary can be learned from all training samples simultaneously. The learned dictionary encourages images from the same class to have similar representations (*i.e.*, lie in a low-dimensional subspace); while images from other classes have very different representations. This leads to high recognition performance of our approach, as shown in the experiment section.

3. Learning Structured Sparse and Low-rank Representation

To better classify images even when training and testing images have been corrupted, we propose a robust supervised algorithm to learn a structured sparse and low-rank representation for images. We construct a discriminative dictionary via explicit utilization of label information from the training data. Based on the dictionary, we learn low-rank and sparse representations for images. Classification is carried out directly on these discriminative representations.

3.1. Problem Statement

We are given a data matrix $X = [X_1, X_2, ..., X_N]$ with N classes where X_i corresponds to class i. X may be contaminated by noise (occlusion, corruption, illumination differences, etc). After eliminating noise, samples within each class i will demonstrate similar basic structure [2, 18]. As discussed before, low-rank matrix recovery helps to decompose a corrupted matrix X into a low-rank component DZ and a sparse noise component E, *i.e.*, $X = DZ + E$. With respect to a semantic dictionary D, the optimal representation matrix Z for X should be block-diagonal [18]:

$$Z^* \triangleq \begin{pmatrix} Z_1^* & 0 & 0 & 0 \\ 0 & Z_2^* & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & Z_N^* \end{pmatrix}$$

Based on the above discussion, it is possible to learn low-rank and sparse representations for images. Low rankness reveals structure information. Sparsity identifies which class an image belongs to. Given a dictionary D, the objective function is formulated as:

$$\min_{Z,E} ||Z||_* + \lambda||E||_1 + \beta||Z||_1 \qquad (4)$$
$$s.t \quad X = DZ + E$$

where λ, β controls the sparsities of the noise matrix E and the representation matrix Z, respectively.$||.||_*$ and $||.||_1$ denotes the nuclear norm and the l_1-norm of a matrix.

The dictionary $D = [D_1, D_2, ...D_N]$ contains N sub-dictionaries where D_i corresponds to class i. Let $Z_i = [Z_{i,1}, Z_{i,2}, ...Z_{i,N}]$ be the representation for X_i with respect to D. Then $Z_{i,j}$ denotes coefficients for D_j. To obtain a low-rank and sparse data representation, D should have discriminative and reconstructive power. Firstly, D_i should ideally be exclusive to each subject i. Thus, representations for images from different classes would be different. Secondly, every class i is well represented by its sub-dictionary such that $X_i = D_i Z_{i,i} + E_i$. $Z_{i,j}$, the coefficients for D_j $(i \neq j)$, are nearly all zero.

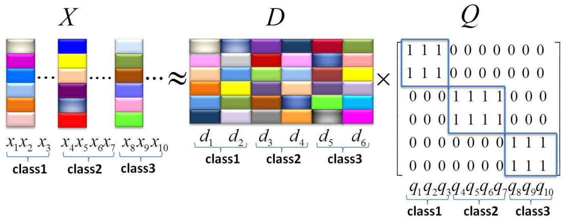


Figure 1. Optimal decomposition for classification.

We say Q is an ideal representation if $Q = [q_1, q_2, ..., q_T] \in R^{K \times T}$ where q_i, the code for sample x_i, is of the form of $[0...1, 1, 1, ...]^t \in R^K$ (K is the dictionary's size, and T is the total number of samples). Suppose x_i belongs to class L, then the coefficients in q_i for D_L are all 1s, while the others are all 0s. An example optimal decomposition for image classification is illustrated in Figure 1. Here, data $X = [X_1, X_2, X_3]$ contains images from 3 classes, where X_1 contains 3 samples x_1, x_2, x_3, X_2 contains 4 samples x_4, x_5, x_6, x_7, and X_3 contains 3 samples x_8, x_9, x_{10}. D has 3 sub-dictionaries, and each has 2 items. Although this decomposition might not result in minimal reconstruction error, low-rank and sparse Q is an optimal representation for classification.

With the above definition, we propose to learn a semantic structured dictionary by supervised learning. Based on label information, we construct Q in block-diagonal form for training data. We add a regularization term $||Z - Q||_F^2$ to include structure information into the dictionary learning process. A dictionary that encourages Z to be close to Q is preferred. The objective function for dictionary learning is defined as follows:

$$\min_{Z,E,D} ||Z||_* + \lambda||E||_1 + \beta||Z||_1 + \alpha||Z - Q||_F^2 \qquad (5)$$
$$s.t \quad X = DZ + E$$

where α controls the contribution of regularization term.

3.2. Optimization

To solve optimization problem (5), we first introduce an auxiliary variable W to make the objective function separable. Problem (5) can be rewritten as:

$$\min_{Z,E,D} ||Z||_* + \lambda||E||_1 + \beta||W||_1 + \alpha||W - Q||_F^2 \qquad (6)$$
$$s.t \quad X = DZ + E, W = Z$$

The augmented Lagrangian function L of (6) is:

$$\begin{aligned} & L(Z, W, E, D, Y_1, Y_2, \mu) \\ = \; & ||Z||_* + \lambda||E||_1 + \beta||W||_1 + \alpha||W - Q||_F^2 \\ & + < Y_1, X - DZ - E > + < Y_2, Z - W > \\ & + \frac{\mu}{2}(||X - DZ - E||_F^2 + ||Z - W||_F^2) \end{aligned} \tag{7}$$

where $< A, B >= trace(A^t B)$.

The optimization for problem (6) can be divided into two subproblems. The first subproblem is to compute the optimal Z, E for a given dictionary D. If we set $\alpha = 0$, this is exactly the optimization problem from (4). The second subproblem is to solve dictionary D for the given Z, E calculated from the first subproblem.

3.2.1 Computing Representation Z Given D

With the current D, we use the linearized alternating direction method with adaptive penalty (LADMAP)[17, 36] to solve for Z and E. The augmented Lagrangian function (7) can be rewritten as:

$$\begin{aligned} & L(Z, W, E, D, Y_1, Y_2, \mu) \\ = \; & ||Z||_* + \lambda||E||_1 + \beta||W||_1 + \alpha||W - Q||_F^2 \\ & + h(Z, W, E, D, Y_1, Y_2, \mu) - \frac{1}{2\mu}(||Y_1||_F^2 + ||Y_2||_F^2) \end{aligned} \tag{8}$$

where $h(Z, W, E, D, Y_1, Y_2, \mu)$
$= \frac{\mu}{2}(||X - DZ - E + \frac{Y_1}{\mu}||_F^2 + ||Z - W + \frac{Y_2}{\mu}||_F^2)$
The quadratic term h is replaced with its first order approximation at the previous iteration step adding a proximal term [17]. The function is minimized by updating each of the variables Z, W, E one at a time. The scheme is as follows:

$$\begin{aligned} Z^{j+1} \; = \; & \arg\min_Z ||Z||_* + < Y_1^j, X - DZ^j - E^j > \\ & + < Y_2^j, Z^j - W^j > + \frac{\mu}{2}(||X - D^j Z^j \\ & - E^j||_F^2 + ||Z^j - W^j||_F^2) \\ = \; & \arg\min_Z ||Z||_* + \frac{\eta\mu}{2}||Z - Z^j||_F^2 \\ & + < \nabla_Z h(Z^j, W^j, E^j, Y_1^j, Y_2^j, \mu), Z - Z^j > \\ = \; & \arg\min_Z \frac{1}{\eta\mu}||Z||_* + \frac{1}{2}||Z - Z^j + \big[- D^T(X - \\ & DZ^j - E^j + \frac{Y_1^j}{\mu}) + (Z - W^j + \frac{Y_2^j}{\mu})\big]/\eta||_F^2 \end{aligned} \tag{9}$$

$$\begin{aligned} W^{j+1} \; = \; & \arg\min_W \beta||W||_1 + \alpha||W - Q||_F^2 \\ & + < Y_2^j, Z - W > + \frac{\mu}{2}||Z^{j+1} - W||_F^2 \\ = \; & \arg\min_W \frac{\beta}{2\alpha + \mu}||W||_1 + \frac{1}{2}||W - (\frac{2\alpha}{2\alpha + \mu}Q \\ & + \frac{1}{2\alpha + \mu}Y_2^j + \frac{\mu}{2\alpha + \mu}Z^{j+1})||_F^2 \end{aligned} \tag{10}$$

$$\begin{aligned} E^{j+1} \; = \; & \arg\min_E \lambda||E||_1 + < Y_1^j, X - DZ^{j+1} - E > \\ & + \frac{\mu}{2}||X - DZ^{j+1} - E||_F^2 \\ = \; & \arg\min_E \frac{\lambda}{\mu}||E||_1 + \frac{1}{2}||E - (\frac{1}{\mu}Y_1^j + X \\ & - DZ^{j+1})||_F^2 \end{aligned} \tag{11}$$

where $\nabla_Z h$ is the partial differential of h with respect to Z. $\eta = ||D||_2^2$. The calculations are described in Algorithm 1.

Algorithm 1 Low-Rank Sparse Representation via Inexact ALM

```
**Input:** Data $X$, Dictionary $D$, and Parameters $\lambda$, $\beta$, $\alpha$
**Output:** $Z, W, E$
**Initialize:** $Z^0 = W^0 = E^0 = Y_1^0 = Y_2^0 = 0, \rho = 1.1, \epsilon = 10^{-7}, \mu_{max} = 10^{30}$
**while** not converged, $j \leq maxIterZ$ **do**
  fix $W, E$ and update variable $Z$ according to (9)
  fix $Z, E$ and update variable $W$ according to (10)
  fix $Z, W$ and update variable $E$ according to (11)
  update the multipliers:
    $Y_1^{j+1} = Y_1^j + \mu(X - DZ^j - E^j)$
    $Y_2^{j+1} = Y_2^j + \mu(Z^j - W^j)$
  update $\mu$:
    $\mu = \min(\mu_{max}, \rho\mu)$
  check the convergence conditions:
    $||X - DZ^j - E^j||_\infty < \epsilon, ||Z^j - W^j||_\infty < \epsilon$
**end while**
```

3.2.2 Updating Dictionary D with Fixed Z, W, E

With fixed Z, W and E, D is the only variable in this subproblem. So (7) can be rewritten as:

$$\begin{aligned} & L(Z, W, E, D, Y_1, Y_2, \mu) \\ = \; & < Y_1, X - DZ - E > + \frac{\mu}{2}(||X - DZ - E||_F^2 \\ & + ||Z - W||_F^2) + C(Z, E, W, Q) \end{aligned} \tag{12}$$

where $C(Z, E, W, Q)$ is fixed. This equation (12) is a quadratic form in variable D, so we can derive an optimal dictionary D^{update} immediately. The updating scheme is:

$$D^{i+1} = \gamma D^i + (1 - \gamma)D^{update} \tag{13}$$

γ is a parameter that controls the updating step. The dictionary construction process is summarized in Algorithm 2.

3.2.3 Dictionary Initialization

To initialize the dictionary, we use the K-SVD method. The initial sub-dictionary D_i for class i is obtained by several iterations within each training class. The input dictionary D^0 is initialized by combining all the individual class dictionaries, *i.e.*, $D^0 = [D_1, D_2, ...D_N]$.

3.3. Classification

We use a linear classifier for classification. After the dictionary is learned, the low-rank sparse representations Z of

Algorithm 2 Dictionary Learning via Inexact ALM

```
Input: Data $X$, and Parameters $\lambda$, $\beta$, $\alpha$, $\gamma$
Output: $D, Z$
Initialize: Initial Dictionary $D^0$, $\epsilon_d = 10^{-5}$
while not converged, $i \leq maxIterD$ do
    find $Z, W, E$ with respect to $D^i$ using Algorithm 1
    fix $Z, W, E$ and update D by:
        $D^{update} = \frac{1}{\mu}(Y_1 + \mu(X - E))Z^T(ZZ^T)^{-1}$
        $D^{i+1} = \gamma D^i + (1 - \gamma)D^{update}$
    check the convergence conditions:
        $||D^{i+1} - D^i||_\infty < \epsilon_d$
end while
```

training data X and Z_{test} of test data X_{test} are calculated by solving (4) separately using Algorithm 1 with $\alpha = 0$. The representation z_i for test sample i is the ith column vector in Z_{test}. We use the multivariate ridge regression model [11, 32] to obtain a linear classifier $\hat{W}$:

$$\hat{W} = \arg\min_W ||H - WZ||_2^2 + \lambda||W||_2^2 \qquad (14)$$

where H is the class label matrix of X. This yields $\hat{W} = HZ^T(ZZ^T + \lambda I)^{-1}$. Then label for sample i is given by:

$$k = arg\max_k(s = \hat{W}z_i) \qquad (15)$$

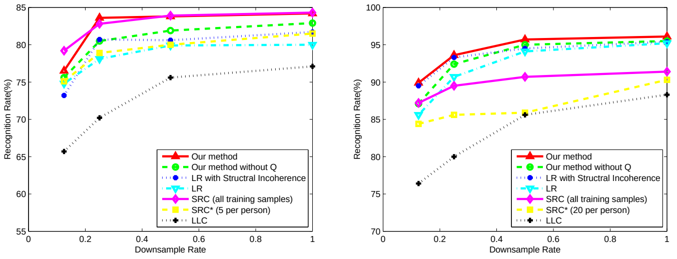
where s is the class label vector.

4. Experiments

We evaluate our algorithm on three datasets. Two face databases: Extended YaleB [10], AR [22], and one object category database: Caltech101 [14]. Our approach is compared with several other algorithms including the locality-constrained linear coding method (LLC) [25], SRC [27], LR [5], LR with structural incoherence from [5], DLRD_SR [19] and our method without the regularization term $||Z - Q||$ (our method without Q). Our method without Q involves simply setting $\alpha = 0$ in the dictionary learning process. Unlike most other image classification methods [23, 1, 29], training and testing data can both be corrupted. Our algorithm achieves state of the art performance.

4.1. Extended YaleB Database

The Extended YaleB database contains 2,414 frontal-face images of 38 people. Taken under various controlled lighting conditions, these cropped images have size 192 × 168 pixels. There are between 59 and 64 images for each person. Shadows due to different illumination conditions cause variations in this dataset. We test our algorithm on the original images as well as down-sampled images $(2, 4, 8)$. This results in data sets of feature dimension 32256, 8064, 2016 and 504. We randomly select 8 training images for each person, repeat this 5 times and report average recognition accuracy. Our trained dictionary has 5 items for each class. We repeat our experiments starting with 32 randomly selected training images and 20 dictionary items per class.



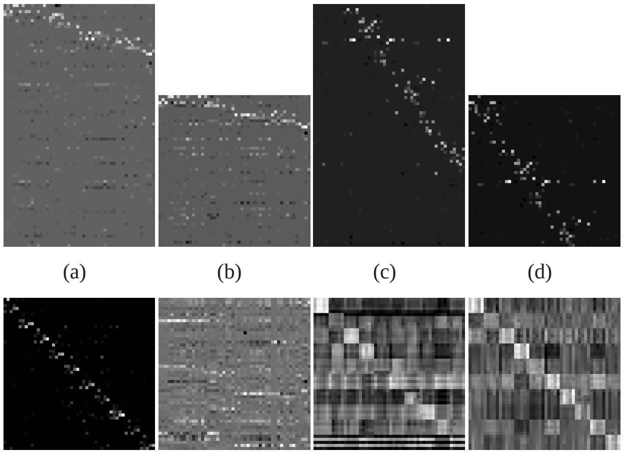
(a) n = 8

(b) n = 32

Figure 2. Performance comparisons on the Extended YaleB. n is the number of training images per person.

We compare our approach with LLC [25], SRC [27], LR [5], and LR with structural incoherence [5]. We evaluate the performance of the SRC algorithm using a full-size dictionary (all training samples). For fair comparison, we also evaluate the results of SRC and LLC using dictionaries whose sizes are the same with ours. The result for our method without Q is also calculated. The comparative results are shown in Figure 2. n is the number of training samples for each person. Our method, by taking advantage of structure information, achieves better performance than LLC, LR, LR with structural incoherence and our method without Q. It outperforms SRC when using the same-size dictionary. Our result is also comparable with [9].

Figure 3 illustrates the representations for the first ten subjects. The dictionary contains 50 items (5 for each category). The first line shows the testing images' representation based on LR and LR with structural incoherence [5]. Figures 3(a) and 3(c) are representations with the full size dictionary (all training sample). For comparison, we randomly select 5 out of 8 training samples from each class,



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3. Comparison of representations for testing samples from the first ten classes on the Extended YaleB. 5 example samples for each class. (a) LR with full-size dictionary; (b) LR with dictionary size 50; (c) LR with structural incoherence with full-size dictionary; (d) LR with structural incoherence with dictionary size 50; (e) SRC; (f) LLC; (g) Our method without Q; (h) Our method.

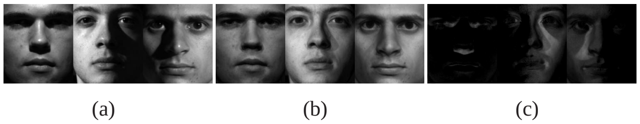


Figure 4. Examples of image decomposition for testing samples on the Extended YaleB. (a) original faces; (b) the low-rank component DZ; (c) the sparse noise component E.

and generate a 50-element dictionary. The corresponding representations are shown in Figures 3(b) and 3(d). Figures 3(e), 3(f) and 3(g) are the representations based on SRC, LLC with the same dictionary size and our method without Q. In our learned representation, Figure 3(h), images from the same class show strong similarities. This representation is much more discriminative than the others.

We present some examples of decomposition results in Figure 4. The first three images are original faces. The middle and the last three images are low-rank component(DZ) and noise component(E), respectively. We see that different illumination conditions mainly influence noise component.

We also evaluate the computation time of our approach and LR with structural incoherence [5] that trains a model class by class (Figure 5(a)) and uses SRC for classification. The training time is computed as the average over the entire training set. The testing time, which includes both encoding and classification, is averaged over all test samples. Clearly, training over all classes simultaneously is faster than class by class if discriminativeness is preserved for different classes. Our training time is twice as fast and testing is three times faster than LR with structural incoherence.

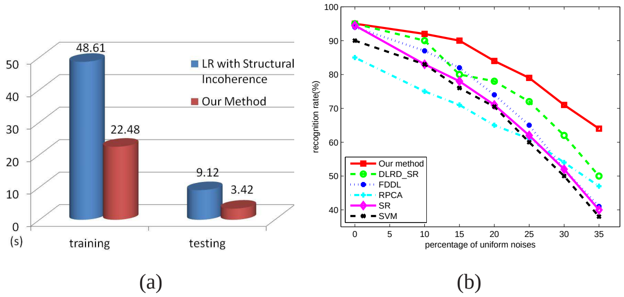


Figure 5. Experiment results. (a) Average computation time for training and testing on the Extended YaleB; (b) Recognition rates on the AR database with pixel corruption.

4.2. AR Database

The AR face database includes over 4,000 color face images of 126 individuals, 26 images for each person in two sessions. In each session, each person has 13 images. Among them, 3 are obscured by scarves, 6 by sunglasses, and the remaining faces are of different facial expressions or illumination variations which we refer to as unobscured images. Each image is 165×120 pixels. We convert the color images into gray scale and down-sample 3×3. Following the protocol in [5], experiments are run under three different scenarios:

Table 1. Recognition rates on the AR

Dimension2200	sunglass	scarf	mixed
Our Method	87.3	83.4	82.4
Our Method without Q	85.1	81.3	81.0
LR w. Struct. Incoh. [5]	84.9	76.4	80.3
LR [5]	83.2	75.8	78.9
SRC(all train. samp.) [27]	86.8	83.2	79.2
SRC*(5 per person) [27]	82.1	72.6	65.5
LLC [25]	65.3	59.2	59.9

Sunglasses: In this scenario, we consider unobscured images and those with sunglasses. We use seven unobscured images from session 1 and one image with sunglass as training samples for each person, the rest as testing. Sunglasses cover about 20% of the face.

Scarf: In this scenario, we consider unobscured images and those with scarves. We use seven unobscured images from session 1 and one image with a scarf as training samples for each person, the remainder as testing. Scarves give rise to around 40% occlusion.

Mixed (Sunglass + Scarf): In the last scenario, we consider all images together (sunglass, scarf and the unobscured). We use seven unobscured images from session 1, one image with sunglasses, and one with a scarf as training samples for each person.

We repeat our experiments three times for each scenario and average the results. Table 1 summarizes the results. We use $\alpha = 560$, $\lambda = 16$, $\beta = 15$, $\gamma = 0.1$ in our experiments. Our methods are compared with LLC [25], SRC [27], LR [5], and LR with structural incoherence [5]. For SRC, we measure the performance with two different

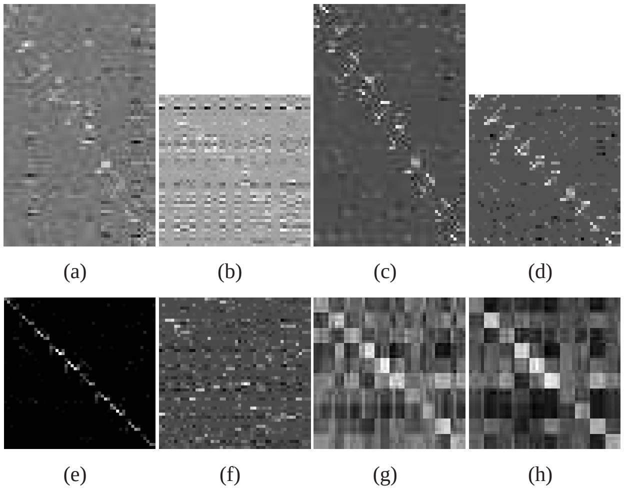
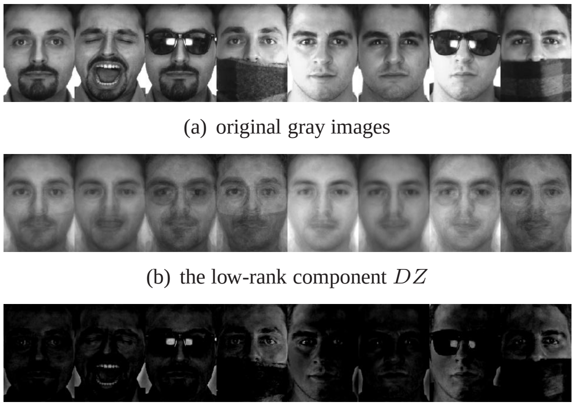


Figure 6. Comparison of representations for testing samples from the first ten classes on the AR for the sunglass scenario. 5 samples for each class. (a) LR with full-size dictionary; (b) LR with dictionary size 50; (c) LR with structural incoherence with full-size dictionary; (d) LR with structural incoherence with dictionary size 50; (e) SRC; (f) LLC; (g) Our method without Q; (h) Our method.



(a) original gray images

(b) the low-rank component DZ

(c) the sparse noise component E

Figure 7. Examples of image decomposition for testing samples from class 4 and 10 on the AR.

dictionary sizes. Our approach achieves the best results and outperforms other approaches with the same dictionary size by more than 3% for the sunglass scenario, 7% for the scarf scenario, and 2% for the mixed scenario.

We visualize the representation Z for the first ten classes under the sunglasses scenario. There are $8 \times 10 = 80$ training images and $12 \times 10 = 120$ testing images. We use 50 as our dictionary size, *i.e.*, 5 dictionary items per class. Figures 6(a) and 6(c) show the representations of LR and LR method without structural incoherence with a full-size dictionary. In Figures 6(b) and 6(d), we randomly pick 5 dictionary items for each class, and use this reduced dictionary to learn sparse codes. For comparison purposes, we also choose 50 as the dictionary size in LLC and SRC* to learn the representations shown in Figures 6(e) and 6(f). The testing images automatically generate a block diagonal structure in our method, which is absent in other methods.

Figure 7 shows image decomposition examples on the AR database. The first row shows the original gray images. The second is the low-rank component (DZ) and the third the noise component (E). Our approach separates occlusions such as sunglasses and scarves from the original images into the noise component.

Table 2. Recognition rates on the AR

Dimension2200	sunglass	scarf
Our Method	90.9	88.5
LC-KSVD [12]	78.4	63.7

In addition, we compare our results with LC-KSVD [12] using the same training samples under the sun and scarf scenarios, using unobscured images for test. The results is summarized in Table 2. Although associating label information with training process, the performance of LC-KSVD is not as good as ours since the training set is smaller and corrupted. Our approach is robust to noise like occlusion.

We also evaluate our algorithm on the corrupted AR face database following the protocol in [19]. In the experiment, seven images with illumination and expression variations from session 1 are used for training images, and the other seven images from session 2 are used as testing images. A percentage of randomly chosen pixels from each training and testing image are replaced with iid samples from a uniform distribution over $[0, V_{max}]$ as [26] did, where V_{max} is the largest possible pixel value in the image. The recognition rates under different levels of noises are shown in Figure 5(b). The results of DLRD_SR [19], FDDL [30], Robust PCA [26], SR [27], and SVM [26] are copied from [19]. Our method outperforms the other approaches. Figure 8 shows some examples of image decomposition on the AR database with 20% uniform noise.

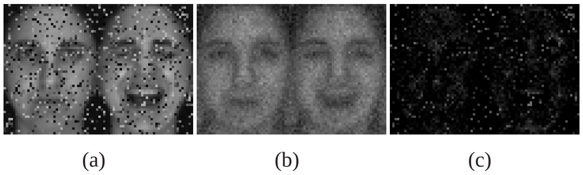


Figure 8. Examples of image decomposition for testing samples from class 95 on the AR with 20% uniform noise. (a) corrupted faces; (b) the low-rank component DZ; (c) the sparse noise component E.

4.3. Caltech101 Database

The Caltech101 database contains over 9000 images from 102 classes. 101 classes are of animals, flowers, trees, etc. and there is a background class. The number of images in each class is between 31 and 800. We evaluate our methods using spatial pyramid features and run experiments with 15 and 30 randomly chosen training images.

Figure 9 shows the representations of 15 testing samples which are randomly selected from classes $4 \sim 8$. Our representation clearly reveals structure information through representation similarity. Although the training images are visually very diverse, we are able to learn discriminative

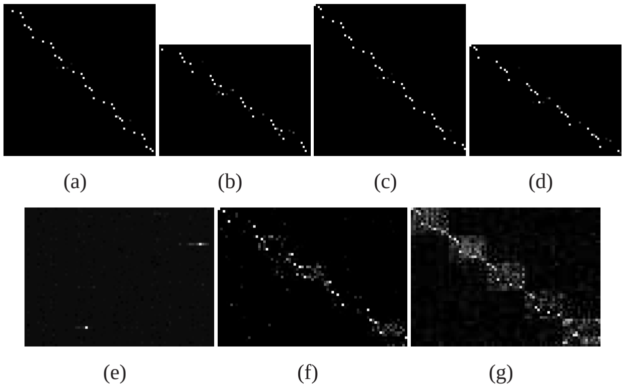


Figure 9. Comparison of representations for testing samples from class 4 to 8 on the Caltech101. 15 example samples for each class. (a) LR with full-size dictionary; (b) LR with dictionary size 55; (c) LR with structural incoherence with full-size dictionary; (d) LR with structural incoherence with dictionary size 55; (e) LLC; (f) Our method without Q; (g) Our method.

Table 3. Recognition rates on the Caltech101

number of training sample	15	30
Our Method	66.1	73.6
Our Method without Q	65.5	73.3
LR w. Struct. Incoh.[5]	58.3	65.7
LR [5]	50.3	60.1
SRC (all train. samp.) [27]	64.9	70.7
LLC [25]	65.4	73.4

representations with the constructed dictionary.

We evaluate our approach and compare it with others [5, 25, 27]. Table 3 presents classification accuracy. Our algorithm achieves the best performance. Figure 10 gives examples from classes which achieve high classification accuracy when training image is 30 per category.

(a) yin_yang, acc:100%

(b) soccer_ball, acc:100%

(c) sunflower, acc:100%

(d) Motorbikes, acc:97.7%

(e) accordion, acc:96.0%

(f) watch, acc:95.7%

Figure 10. Example images from classes with high classification accuracy of the Caltech101.

5. Conclusions

We proposed a new image classification model to learn a structured low-rank representation. Incorporating label information into the training process, we construct a semantic structured and constructive dictionary. Discriminative representations are learned via low-rank recovery even for corrupted datasets. The learned representations reveal structural information automatically and can be used for classification directly. Experiments show our approach is robust, achieving state-of-art performance in the presence of various sources of data contamination, including illumination changes, occlusion and pixel corruption.

Acknowledgement

This work was supported by the Army Research Office MURI Grant W911NF-0910383.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing*, pages 54(1):4311–4322, 2006. 1, 2, 5

[2] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces, 2001. *ICCV*. 3

[3] D. Bradley and J. Bagnell. Differential sparse coding, 2008. *NIPS*. 1

[4] E. Candès, X. Li, Y. Ma, and J. Wright. Robust pricipal component analysis? *Journal of the ACM*, 58, 2011. 2

[5] C. Chen, C. Wei, and Y. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition, 2012. *CVPR*. 1, 2, 5, 6, 8

[6] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinities pursuit for image segmentation, 2011. *ICCV*. 1, 2

[7] X. Cui, J. Huang, S. Zhang, and D. Metaxas. Background subtraction using group sparsity and low rank constraint, 2012. *ECCV*. 1

[8] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, pages 15(1)2):3736–3745, 2006. 2

[9] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation, 2011. *CVPR*. 1, 5

[10] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *TPAMI*, pages 23(6):643–660, 2001. 5

[11] G. Golub, P. Hansen, and D. O'leary. Tikhonov regularization and total least squares. *SIM J.Matri Anal. Appl.*, pages 21(1):185–184, 1999. 5

[12] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd, 2011. *CVPR*. 1, 2, 7

[13] J. Lee, B. Shi, Y. Matsushita, I. Kweon, and K. Ikeuchi. Radiometric calibration by transform invariant low-rank structure, 2011. *CVPR*. 1, 2

[14] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training samples: An incremental bayesian approach tested on 101 object categories, 2004. *CVPR Workshop on Generative Model Based Vision*. 5

[15] Z. Lin, M. Chen, and Y. Ma. The argumented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Tech. Rep. UIUC-ENG-09-2214*, 2011. 2

[16] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *IEEE Transactions on Information Theory*, 2008. 2

[17] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penality for low rank representation, 2011. *NIPS*. 4

[18] G. Liu, Z. Liu, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 2, 3

[19] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning, 2012. *CVPR*. 1, 2, 3, 5, 7

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis, 2008. *CVPR*. 1

[21] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning, 2009. *NIPS*. 1

[22] A. Martinez and R. Benavente. The ar face database, 1998. *CVC Technical Report 24*. 5

[23] D. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition, 2008. *CVPR*. 1, 2, 5

[24] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery, 2012. *CVPR*. 1, 2

[25] J. Wang, A. Yang, K. Yu, F. LV, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification, 2010. *CVPR*. 1, 5, 6, 8

[26] J. Wright, A.Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust pricipal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *submitted to Journal of the ACM*, 2009. 2, 7

[27] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, pages 31(2):210–227, 2009. 1, 5, 6, 7, 8

[28] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding, 2010. *CVPR*. 1

[29] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary, 2010. *ECCV*. 1, 2, 5

[30] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition, 2011. *CVPR*. 2, 7

[31] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition, 2011. *CVPR*. 1, 2

[32] G. Zhang, Z. Jiang, and L. Davis. Online semi-supervised discriminative dictionary learning for sparse representation, 2012. *ACCV*. 5

[33] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition, 2010. *CVPR*. 1

[34] T. Zhang, B. Ghanem, and N. Ahuja. Low-rank sparse learning for robust visual tracking, 2012. *ECCV*. 1, 2

[35] Z. Zhang, Y. Matsushita, and Y. Ma. Camera calibration with lens distortion from low-rank textures, 2011. *CVPR*. 1, 2

[36] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning, 2012. *CVPR*. 4