2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Learning Discriminative Low-rank Representation for Image Classification

Jun Li, Heyou Chang and Jian Yang

Abstract— **Low-rank representation (LRR) efficiently performs the subspace segmentation and feature extraction from corrupted data. However, there are three disadvantages in existing LRR techniques. First, the inference algorithm of LRR (as a generative model) is computationally expensive. Second, LRR ignores the discriminative information for image classification. Third, although the robust representation is implemented by recovering the low-rank components and the sparse noises, it has been limited due to the constrained assumption that noises is sparse. To solve these problems, and inspired by Denoising Autoencoders (DAE) and Contractive Autoencoders (CAE), this paper proposes a discriminative low-rank representations framework (DLRR) for image classification. We directly learn a discriminative projection dictionary that results in fast inference. Simultaneously, DLRR can obtain a robust representation from any corrupted input. Our implementation of DLRR achieves state-of-the-art results on artificial dataset and dataset of Olivetti Face Patches.**

I. Introduction

LOW-RANK representation (LRR) has received increasing attention because of their successful applications in computer vision and machine learning. In general, LRR [12], [13], [11] is to find the lowest-rank representation among all the data, which can be represented as a linear combination of the bases in a dictionary. Obviously, LRR efficiently performs the subspace segmentation and feature extraction from corrupted data. To study the insufficient and/or grossly corrupted data matrix (dictionary), [13] proposes a latent low-rank representation and inspired by matrix factorization, [14] proposes fixed-rank representation as a unified framework for unsupervised visual learning. [25] exploits the low-rank nature of particle representations for robust visual tracking. [28] proposes a novel non-negative low-rank and sparse graph for semi-supervised learning. [3] presents multi-task low-rank affinity pursuit for image segmentation. Recently, [23], [19], [26] and [27] learn low-rank representations for classification tasks[1]. However, there have two disadvantages that restrict the applications of LRR.

First, a major disadvantage with LRR (as a generative model) is that the inference algorithm is somewhat expensive. In particular, it has been limited due to prohibitive cost of calculating the low-rank representations for image classification [26], [27]. In order to make inference efficient in sparse coding, [7], [8], [5] and [20] train a parameterized nonlinear function that maps input data to the representations. Inspired by the fast sparse coding [7], [5], we directly train a projection dictionary.

Second, LRR ignores the discriminative information for image classification. The applications of LRR heavily depend on the dictionary, which usually chooses the observed data matrix itself [12] and dose not have discriminative information. In order to obtain the discrimination capabilities, the dictionary is learned by the training data reconstruction error per class [16], [15] and all training data reconstruction error [17], [26].

Motivated by these considerations, a projection dictionary is directly learned for reducing the expensive cost of inference algorithm. To have discriminative capabilities, label information from training data is incorporated into the projection dictionary learning process by adding a label-constraint term. Therefore, this paper proposes a discriminative low-rank representations framework (DLRR) for image classification.

On the other hand, it is a highly desirable property to extract invariant representations for classification tasks [4], [22], [18]. The concept of invariance implies that a representation is robust to the variant input of a class object. Many methods [13], [11], [26], [27] of LRR[2] learn the robust representations by separating both the low-rank representations and the sparse noises. From the reconstruction criterion, Denoising Autoencoders (DAE) [21], [22] can be obtained robustly from a corrupted input and recover the corresponding clean input. By penalizing the Frobenius norm of the Jacobian matrix of the representation activations with respect to the input, Contractive Autoencoders (CAE) [18] is also more invariant to the vast majority of directions orthogonal to the manifold. Such representation will yield a better performing classifier [21], [22], [18]. Here we employ two strategies: corrupted input and penalizing the Frobenius norm of the Jacobian matrix [21], [22], [18] to learn a robust representation. Therefore, we learn a robust representation by using the low rank method to train a projection dictionary from a corrupted input. Unlike LRR, we do not separate the noises. Unlike DAE and CAE, we do not recover the corresponding clean input.

The rest of this paper is organized as follows: Section II reviews three popular low-rank matrix recovery models and introduce the classification process used LRR. Section

Jun Li, Heyou Chang and Jian Yang are with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China, 219000 (email: junl.njust@gmail.com,csjyang@njust.edu.cn), (see http://www.patternrecognition.cn/ jian/).

[1]The corresponding representations from train and test sets are learned by using LRR and a classifier is constructed by the representations from training set for classification

[2]Suppose we have a data matrix, which is the superposition of a low-rank component and a sparse component [2], [11].

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

III propose a discriminative low-rank representations for image classification. Some experiments results are presented in Section IV to show its performance. Finally, Section V concludes the paper.

II. Low-rank Representation for Image Classification

In this section, we first review three popular low-rank matrix recovery models, classical principal component analysis (PCA), robust principal component analysis (RPCA) and low-rank representation (LRR). Second, we introduce the classification process used LRR.

A. Low-rank Matrix Recovery

For better capturing the intrinsic low-dimensional structure of data, we assume that the observed data matrix X is the superposition of a low-rank matrix A and a noise matrix E. If E is a small perturbation matrix, classical principal component analysis (PCA) [6] is to find the k-constraint rank matrix A by solving:

$$\min_{A} \ \|X - A\|_2 \quad s.t.\ rank(A) < k \tag{1}$$

where $\|\cdot\|$ denotes the 2-norm. If E can be arbitrary in magnitude, [2] supposes that it is a sparse matrix. Robust PCA aims at exactly recovering the low-rank A and the sparse E. It can be viewed as a regularized rank minimization problem:

$$\min_{A} \ rank(A) + \lambda\|E\|_0 \quad s.t.\ X = A + E \tag{2}$$

where $rank$ is the rank of A, $\|\cdot\|_0$ is the 0-norm and $\lambda > 0$ is a parameter. Consider that the underlying data structures are multiple low-rank subspaces, such as face recognition. LRR [11] shows that a more general rank minimization problem is as following:

$$\min_{D,Z,E} \ rank(Z) + \lambda\|E\|_0 \quad s.t.\ X = DZ + E \tag{3}$$

where D is a dictionary matrix that linearly spans the data space and Z is the lowest-rank representation of data X. Unfortunately, (3) is a highly non convex optimization problem. By relaxing the 0-norm and the rank, (3) becomes a tractable optimization problem. They are respectively replaced by the 1-norm and the nuclear norm. The optimization problem (3) of LRR is equivalent to:

$$\min_{D,Z,E} \ \|Z\|_* + \lambda\|E\|_1 \quad s.t.\ X = DZ + E \tag{4}$$

where $\|\cdot\|_*$ is the nuclear norm and $\|\cdot\|_1$ is the 1-norm. Generally, the linearized alternating direction method with adaptive penalty (LADMAP) [9], [10] is used to solve the (4).

Algorithm 1 A Linear Classifier

1: **Input:** Data Z, Label Q and Parameters τ
2: **Initialize:** all parameters θ of the DLRR
3: Given Z, Q, τ and update K by:

$$K = QZ^T(ZZ^T + \tau I)^{-1}$$

4: **return** solution K to problem (7).

Suppose the clear data $X^\dagger = X - E$, $X = DZ + E$ can be rewritten as:

$$\begin{pmatrix} X_1^\dagger, X_2^\dagger, \cdots, X_k^\dagger \end{pmatrix} = \begin{pmatrix} D_1, D_2, \cdots, D_k \end{pmatrix} \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k \end{pmatrix} \tag{5}$$

for k independent subspaces. some general properties of the minimizer to problem (5) are as followings:

Theorem 1 [11]: Assume $D_i \neq 0$ and $X_i^\dagger = D_i Z_i$ have feasible solution(s), i.e., $X_i^\dagger \in span(D_i)$. Then for all $i (1 \leq i \leq k)$,

$$Z_i = D_i^\dagger X_i^\dagger \tag{6}$$

is the unique minimizer to problem (5), where $D_i^\dagger$ is the pseudoinverse of D_i. Clearly, $rank(Z_i) = rank(X_i^\dagger)$ and Z_i is also a minimal rank solution to the problem.

B. Classification Process

To classification tasks, clearly, we can obtain the low-rank representations $Z_{training}$ and the dictionary D from the training data $X_{training}$. [27] uses the regression model to train a linear classifier:

$$K = \arg\min_{K} \|Q - KZ_{training}\|_F^2 + \tau\|K\|_F^2 \tag{7}$$

where Q is the class label matrix of data X and $\tau > 0$ is a parameter. The learning process is showed in Algorithm (1).

Next, using the D, the low-rank representations Z_{test} of test data X_{test} can be obtained by optimizing the (4). (However, solving Z_{test} is somewhat expensive.) Then label for sample i is given by:

$$l = \arg\max_{s}(s = KZ_{i_{test}}) \tag{8}$$

where s is the test label vector.

III. Discriminative Low-rank Representations for Image Classification

In this section, we propose a discriminative low-rank representations framework for image classification.

A. Motivation

Given a data X and a dictionary matrix D, solving the low-rank representations Z is somewhat expensive. From *Theorem 1*, however, we know that a closed form of Z is $A^{\dagger}X^{\dagger}$. So, we can directly learn a projection dictionary B and quickly obtain $Z = BX^{\dagger}$.

For classification tasks when all the label data points are stacked as column vectors of a matrix, the label matrix should have low rank. Thus, we add the label information to obtain the discrimination capabilities of projection dictionary.

To classification tasks the robust representation is a highly desirable property. Inspired by DAE [21], [22] and CAE [18], robust representation is obtained by training a projection dictionary from any corrupted input and penalizing the Frobenius norm of the Jacobian matrix of the representation activations.

B. Problem Statement

The key idea of DLRR is to directly learn a projection dictionary.

$$\min_{B,Z,W} \ rank(Z) + \frac{\alpha}{2}\|Z - BX\|_F^2 \quad s.t.\ Q = WZ \tag{9}$$

where α is the parameter, Q is the label matrix, B is a projection dictionary matrix and W is the projection matrix from the low-rank representations to label matrix. However, direct optimization of (9) is NP-hard. When the rank is replaced by the nuclear norm, the optimization problem (9) is equivalent to:

$$\min_{B,Z,W} \ \|Z\|_* + \frac{\alpha}{2}\|Z - BX\|_F^2 \quad s.t.\ Q = WZ \tag{10}$$

where $\|Z\|_*$ is the nuclear norm (i.e., the sum of the singular values) of Z. It approximates the rank of Z.

Although [24] also uses discriminative projection method to train LRR, it seeks a linear transformation by using the Z of (4). We directly learn the linear transformation.

C. Alternating Direction Method

Solving the optimization problem (10) by Alternating Direction Method. We first convert (10) to the following equivalent problem:

$$\begin{aligned} \min_{J,B,Z,W} \ \|J\|_* + \frac{\alpha}{2}\|Z - BX\|_F^2 \\ s.t.\ Z = J \\ Q = WZ \end{aligned} \tag{11}$$

The optimization problem (11) is convex and can be solved by various methods. For efficiency, we adopt in this paper the LADMAP [9], [10]. The augmented Lagrangian function of (11) is

$$\begin{aligned} L =& \|J\|_* + \frac{\alpha}{2}\|Z - BX\|_F^2 + \\ & tr\left(Y_1^T(Z - J)\right) + tr\left(Y_2^T(Q - WZ)\right) + \\ & \frac{\mu}{2}(\|Z - J\|_F^2 + \|Q - WZ\|_F^2) \end{aligned} \tag{12}$$

Algorithm 2 Inner loop of DLRR

1: **Input:** Data X, Projection Dictionary B, Projection Matrix W and parameter α.
2: **Initialize:** all parameters θ of the DLRR
3: **While** not converged do
4: Step 1: Given Z, Y_1, μ and update J by:

$$(U, \Sigma, V) = SVD(J - Y_1/\mu)$$
$$J = U\mathcal{S}_{\frac{1}{\mu}}(\Sigma)V$$

5: Step 2: Given $J, B, W, Q, X, Y_1, Y_2, \mu$ and update Z by:

$$Z = \left(\frac{\mu + \alpha}{\mu}I + W^TW\right)^{-1} \left(W^TQ + \frac{\alpha}{\mu}BX + J + (W^TY_2 - Y_1)/\mu\right)$$

6: Step 3 Given Z, Q, Y_2, μ, η and update W by:

$$W = (Q + Y_2/\mu)Z^T(ZZ^T + \eta I)^{-1}$$

7: Step 4: Given Z, X, η and update B by:

$$B = ZX^T(XX^T + \eta I)^{-1}$$

8: Step 5: Given Z, J, Z, Q, μ and update Y_1, Y_2 by:

$$Y_1 = Y_1 + \mu(Z - J)$$
$$Y_2 = Y_2 + \mu(Q - WZ)$$

9: **until** a stopping criterion is satisfied
10: **return** solution Z, B, W to problem (11).

The augmented Lagrangian function (12) can be rewritten as:

$$\begin{aligned} L =& \|J\|_* + \frac{\alpha}{2}\|Z - BX\|_F^2 + \\ & \frac{\mu}{2}(\|Z - J + Y_1/\mu\|_F^2 + \|Q - WZ + Y_2/\mu\|_F^2) \\ & - \frac{1}{2\mu}(\|Y_1\|_F^2 + \|Y_2\|_F^2) \end{aligned} \tag{13}$$

The function is minimized by updating each of the variables one J, Z, W, B at a time. The scheme is as follows:

$$J = \arg\min_J \frac{1}{\mu}\|J\|_* + \frac{1}{2}\|J - (Z + Y_1/\mu)\|_F^2 \tag{14}$$

$$\begin{aligned} Z =& \arg\min_Z \frac{\alpha}{2\mu}\|Z - BX\|_F^2 + \\ & \frac{1}{2}(\|Z - J + Y_1/\mu\|_F^2 + \|Q - WZ + Y_2/\mu\|_F^2) \end{aligned} \tag{15}$$

$$B = \arg\min_B \|Z - BX\|_F^2 \tag{16}$$

$$W = \arg\min_W \|Q - WZ + Y_2/\mu\|_F^2 \tag{17}$$

In order to obtain the robust representation $Z(X)$ from a training input X we propose to penalize its sensitivity to that input, measured as the Frobenius norm of the Jacobian

Algorithm 3 Outer loop of DLRR

Input: Data $\{X_1, X_2, \cdots, X_n\}$, Projection Dictionary B, Projection Matrix W and Parameters $\alpha, \zeta, \nu, \epsilon$
Initialize: all parameters θ of the DLRR
do
for batch data $X_1, X_2, \cdots, X_n$ in the training set
Step 1: compute corrupted input data:
$X = X_i . * (X_i > \nu)$
Step 2 (inner loop): solve the linearized convex optimization: $(Z^\star, B^\star, W^\star) \leftarrow$

$$\arg \min_{B,Z,W} \|Z\|_* + \frac{\alpha}{2}\|Z - B^i X\|_F^2 \quad s.t.\ Q = W^i Z$$

Step 4: update transformations:
$B^{i+1} = \zeta B^i + (1-\zeta)B^\star$
$W^{i+1} = \zeta W^i + (1-\zeta)W^\star$
until a stopping criterion is satisfied
$\|B^{i+1} - B^i\|_\infty < \epsilon$
$\|W^{i+1} - W^i\|_\infty < \epsilon$
return solution Z, B, W to problem (10).

$\mathcal{J}_Z(X)$. Formally, this penalization term is as follow:

$$\mathcal{J}_Z(X) = \|\frac{\partial Z(X)}{\partial X}\|_F^2 = \|B\|_F^2 \tag{18}$$

Penalizing $\mathcal{J}_Z(X)$ encourages the projection to the feature space to be contractive in the neighborhood of the training data. So, solving B is rewritten as

$$B = \arg\min_B \|Z - BX\|_F^2 + \eta\|B\|_F^2 \tag{19}$$

where η is a hyper-parameter controls the strength of the regularization. Similarly, solving W is rewritten as

$$W = \arg\min_W \|Q - WZ + Y_2/\mu\|_F^2 + \eta\|W\|_F^2 \tag{20}$$

The projection dictionary learning process is outlined in Algorithm (2).

D. Extracting Robust Representation from Corrupted Input

A robust representation is invariant to the corrupted data. Under the so-called manifold assumption [1], the natural high dimensional data concentrates close to a linear low-dimensional manifold in this paper. A geometric interpretation of the corrupted data is illustrated in Figure 2 of [22]. We do experiments to a simple corruption processes: a fraction ν of the elements of data X (chosen at random for each example) is forced to 0. The corruption learning process is also outlined in Algorithm (3).

IV. Experiments

In this paper we present experimental results on two datasets: a artificial dataset and a dataset of Olivetti Face Patches. Our approach is compared with LRR algorithms. **Artificial Dataset:** [11] We construct 5 independent subspaces, each of which has a rank of 10, sample 200 points of dimension 100 from each subspace, and randomly choose some points to corrupt. Using this method, the dataset has 1000 training simples and 1000 test simples. **Olivetti Face Patches:**[3] The Olivetti face dataset from which we obtain the face patches contains ten 64×64 images of each of forty different people. We construct a dataset of 7200 25×25 images by rotating ($-45°$ to $+45°$) and scaling (1.5) the original 400 images. The dataset is randomly subdivided into 3600 training images and 3600 test images. Figure (1) shows 10 randomly selected Olivetti Face and 25 randomly selected Olivetti Face Patches.

TABLE I
Inference Time (second) on LRR and DLRR

datasets	LRR	DLRR
Artificial Dataset	60	0.008
Olivetti Face Patches	837	0.17

TABLE II
Test errors on LRR and DLRR

datasets	LRR	DLRR
Olivetti Face Patches	21.69%	16.78 %

A. Comparison of Fast inference algorithm

We compare our approach with LRR. The comparative results of inference time are shown in (I). Our method is faster than LRR because DLRR only does linearly project.

B. Comparison of Classification Performance

An advantage of DLRR is that it can fast infer and has discriminative capabilities. We have also shown that it performs better than LRR. Table II compares our classification performance of DLRR to LRR. Surprisingly, for Olivetti Face Patches, our performance of DLRR achieves superior performance. Figure 2 illustrates the representations for Artificial Dataset. Figure 3 also shows the representations for Olivetti Face Patches. For comparison, DLRR has discriminative capabilities.

V. Conclusion

We propose a discriminative low-rank representations framework for image classification. This method can directly learn a projection dictionary that results in fast inference. The dictionary also has some discrimination capabilities. Inspired by DAE [21], [22] and CAE [18], moreover, robust representation is obtained by training the dictionary from any corrupted input and penalizing the Frobenius norm of the Jacobian matrix of the representation activations. Finally, experiments show that DLRR achieves state-of-the-art results.

References

[1] C. M. Bishop. In *Pattern Recognition and Machine Learning*. Springer Press, 2006.
[2] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58:1–37, 2011.

[3] http://www.cs.nyu.edu/ roweis/data/olivettifaces.mat

Fig. 1. A dataset of Olivetti Face Patches. The top line shows the raw Olivetti Face. The down line shows the Olivetti Face Patches.

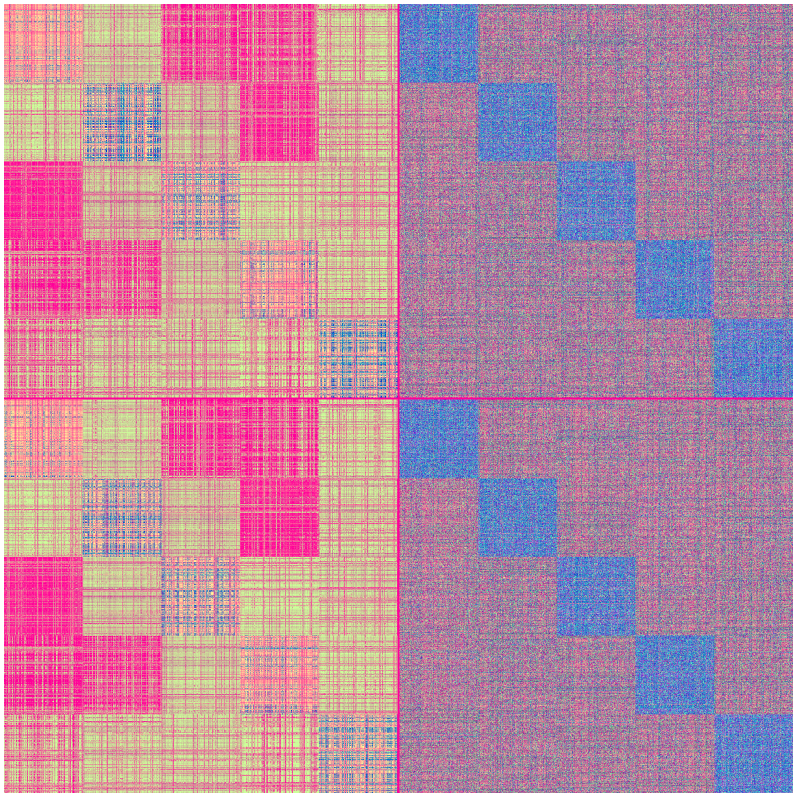

Fig. 2. Comparison of representations for training and testing samples on the Artificial Dataset. The top-left and down-left are, respectively, the representations for training and testing samples using LRR. The top-right and down-right are, respectively, the representations for training and testing samples using DLRR.

[3] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In *ICCV*, 2011.

[4] I. Goodfellow, Q. Le, A. Saxe, H. Lee, and A. Ng. Measuring invariances in deep networks. In *NIPS*, 2009.

[5] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *ICML*, 2010.

[6] I. Jolliffe. In *Principal Component Analysis*. Springer Press, 1986.

[7] K. Kavukcuoglu, M. Ranzato, and Y. LeCun. Fast inference in sparse coding algorithms with applications to object recognition. In *CBLL-TR-2008-12-01*, 2008.

[8] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun. Learning convolutional feature hierarchies for visual recognition. In *NIPS*, 2010.

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

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

[11] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35:171–184, 2013.

[12] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

[13] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *ICCV*, 2011.

[14] R. Liu, Z. Lin, F. Torrez, and Z. Su. Fixed-rank representation for unsupervised visual learning. In *CVPR*, 2012.

[15] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *CVPR*, 2012.

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

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, 2008.

[18] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders explicit invariance during feature extraction. In *ICML*, pages 473–480, 2011.

[19] Z. Shi, J. Han, T. Zheng, and S. Deng. Audio segment classification using online learning based tensor representation feature discrimination. *IEEE Trans. on Audio, Speech, and Language Processing*, 21:186–196,

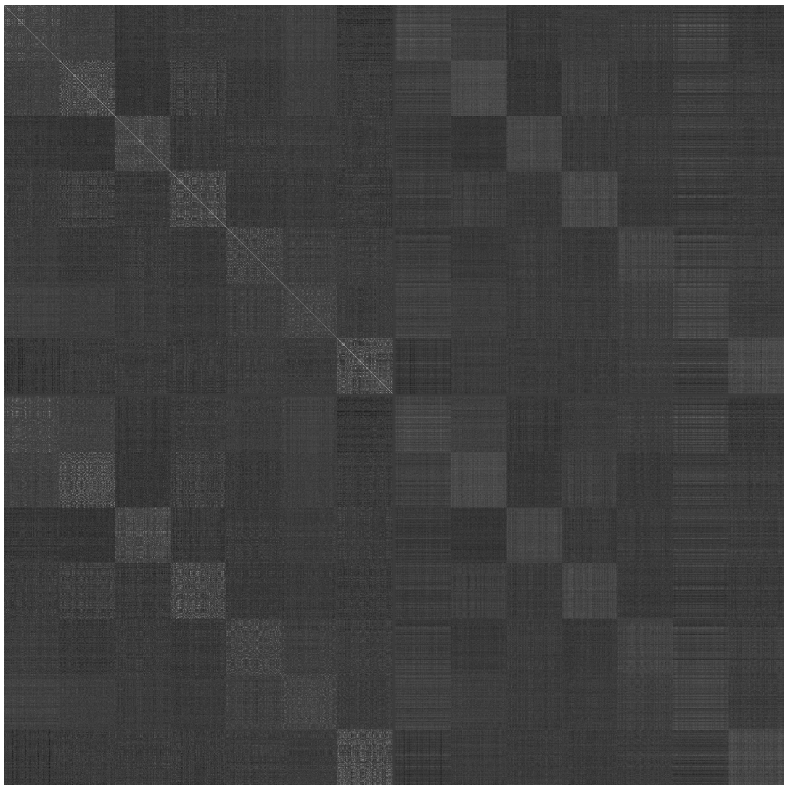

Fig. 3. Comparison of representations for training and testing samples from the first seven classes on the Olivetti Face Patches. The top-left and down-left are, respectively, the representations for training and testing samples using LRR. The top-right and down-right are, respectively, the representations for training and testing samples using DLRR.

2013.

[20] P. Sprechmann, A. M. Bronstein, and G. Sapiro. Learning efficient sparse and low rank models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PrePrints, 2013.

[21] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[23] 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. In *CVPR*, 2012.

[24] N. Zhang and J. Yang. Low-rank representation based discriminative projection for robust feature extraction. *Neurocomputing*, pages 13–20, 2013.

[25] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*, 2012.

[26] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja. Low-rank sparse coding for image classification. In *ICCV*, 2013.

[27] Y. Zhang, Z. Jiang, and L. Davis. Learning structured low-rank representations for image classification. In *CVPR*, 2013.

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