**Weighted Discriminative Dictionary Learning based on Low-rank Representation**

**Heyou Chang and Hao Zheng**

Key Laboratory of Trusted Cloud Computing and Big Data Analysis, School of Information Engineering, Nanjing XiaoZhuang University, Nanjing, China

E-mail: cv_hychang@126.com, zhh710@163.com

**Abstract.** Low-rank representation has been widely used in the field of pattern classification, especially when both training and testing images are corrupted with large noise. Dictionary plays an important role in low-rank representation. With respect to the semantic dictionary, the optimal representation matrix should be block-diagonal. However, traditional low-rank representation based dictionary learning methods cannot effectively exploit the discriminative information between data and dictionary. To address this problem, this paper proposed weighted discriminative dictionary learning based on low-rank representation, where a weighted representation regularization term is constructed. The regularization associates label information of both training samples and dictionary atoms, and encourages to generate a discriminative representation with class-wise block-diagonal structure, which can further improve the classification performance where both training and testing images are corrupted with large noise. Experimental results demonstrate advantages of the proposed method over the state-of-the-art methods.

1. **Introduction**

   Low-rank representation has been widely used in computer vision and pattern recognition due to its strong robustness to the noise of corrupted observation data, such as occlusions, lighting variations and pixel corruptions. The basic idea of low-rank representation is to decompose the observation data into a sparse noise matrix and a low-rank matrix, where the low-rank matrix is often described as the linear combinations of dictionary atoms. The dictionary plays an important role on the discrimination of the representation. In the work of [1], the dictionary is pre-specified as original training set. However, this simple strategy, the training data cannot faithfully represent the test samples due to the noise, and the discriminative information contained in training data cannot be well discovered. Therefore, how to adaptively learn a dictionary from training data has attracted much attention.

   Focusing on this issue, a number of dictionary learning methods for face recognition and image classification have been proposed[2]-[6].To deal with the situation where training data are contaminated with large noises, Ma et al presented a discriminative low-rank dictionary for sparse representation (DLRD_SR)[7] algorithm by minimizing the rank of each sub-dictionary. Based on DLRD_SR, Li et al proposed discriminative dictionary learning with low-rank regularization (D^2LR^2)[8] for image classification. Obviously, these two methods only explore structural information class by class locally and cannot capture the global structure embedded in all classes. To explore the global structure, Zhang et al proposed a discriminative structured low-rank method (DSLRR)[9]. For the purpose of acquiring the structure information in the whole database, including
training data and testing data, Li et al proposed a semi-supervised low rank matrix recovery algorithm for robust face recognition by exploiting the class-wise block-diagonal structure (LR-CBDS)[10]. Although improved performance has been reported, they cannot effectively exploit the discriminative information between data and dictionary. For example, [9] constructed an ideal-code regularization term, which potentially means that all samples from the same class have the same representation. The term may bring negative effects to the representations of the data. In [10], representations of training samples and testing samples are learnt simultaneously, which is inefficient when the dataset is large or dynamic updating. Moreover, the dictionary in [10] is fixed as training samples, which also restricts the performance when training samples are heavily corrupted by noise.

In this paper, we propose weighted discriminative dictionary learning based on low-rank representation (WDDL_LRR) method. In WDDL_LRR, a weighted regularization term on the data coefficients is introduced to strengthen the discriminative ability of the dictionary. The regularization term not only contains global structure information between training data and dictionary, but also encourages that the class-specific dictionary is able to well reconstruct the samples for the same class with bad representation ability for the samples from other classes. Benefiting from the weighted regularization term, the representations learned by WDDL_LRR own class-wise block-diagonal structure, which is critical for the classification problem. Extensive experiments show that the proposed approach has better performance than existing state-of-the-art dictionary learning methods.

2. Weighted discriminative dictionary learning based on low-rank representation

With respect to the semantic dictionary $\mathbf{D}$, the optimal coefficient matrix $\mathbf{Z}$ should have class-wise block-diagonal structure as follows:

$$
\mathbf{Z}^* = \begin{bmatrix}
\mathbf{Z}_1^* & 0 & 0 \\
0 & \mathbf{Z}_2^* & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \mathbf{Z}_c^*
\end{bmatrix}
$$

The coefficient matrix $\mathbf{Z}$ should be low-rank and sparse[1]. To capture the class-wise block-diagonal structure, we introduce a discriminative regularization term, $r(\mathbf{Z}) = \| \mathbf{W} \odot \mathbf{Z} \|^2_F$, where $\odot$ means the element-wise multiplication operator and $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_N]$ is a weighted matrix. $\mathbf{W}$ is defined as

$$
\mathbf{W}(i,j) = \begin{cases} 
0, & \text{if } \mathbf{d}_i \text{ and } \mathbf{x}_j \text{ belong to the same class} \\
1, & \text{otherwise}
\end{cases}
$$

where $\mathbf{d}_i$ is the $i$-th column of $\mathbf{D}$, and $\mathbf{x}_j$ is the $j$-th column of the data matrix $\mathbf{X}$. This term encourages that $\mathbf{Z}_i^*(i \neq j)$ to be 0s and $\mathbf{Z}_i^*$ to be large values, which leads $\sum_{i \neq 1} \| \mathbf{D}_i \mathbf{Z}_i^* \|^2_F$ to be 0 and $\mathbf{Z}_1 \approx \mathbf{D}_i \mathbf{Z}_1^*$. Namely, the class-specific dictionary should be able to reconstruct the samples from the same class and doesn’t have the capacity to represent samples from other classes. This is different from the last term in[7], because the last term in[7] requires $\sum_{i=1,j \neq 1} \| \mathbf{D}_i \mathbf{Z}_j^* \|^2_F$ to be 0, which cannot lead $\mathbf{Z}_i^*(i \neq j)$ to be 0 in reverse. Besides, $\mathbf{W}$ has class-wise structure and contains the global structure information of training samples and dictionary atoms, which can improve the discriminative ability of dictionary and representation. Based on the above discussion, the final target function is formulated as follows

$$
\min_{\mathbf{D}, \mathbf{Z}, \mathbf{E}} \| \mathbf{Z} \|_* + \lambda \| \mathbf{E} \|_1 + \beta \| \mathbf{Z} \|_1 + \alpha \| \mathbf{W} \odot \mathbf{Z} \|_F^2 \quad \text{s.t. } \mathbf{X} = \mathbf{DZ} + \mathbf{E} \tag{1}
$$

where $\lambda$, $\beta$, and $\alpha$ are three parameters.

3. Optimization

To solve optimization problem (1), we introduce two auxiliary variables $\mathbf{J}$ and $\mathbf{L}$ to make the objective function separable. Problem (1) can be rewritten as :
\begin{equation}
\min_{D,Z,E} ||J||_1 + \lambda ||E||_1 + \beta ||L||_1 + \alpha ||W \odot Z||_F^2 \quad \text{s.t.} \quad X = DZ + E, Z = J, Z = L \tag{2}
\end{equation}

The optimization of (2) can be solved by ADMM[11]. The updating scheme is as follows.

**Updating J** Fix the other variables and solve the following problem

\begin{equation}
J^{t+1} = \text{argmin}_J \frac{1}{\mu^t} ||J||_1 + \frac{1}{2} ||J - (Z^t + Y^t_2/\mu^t)||_F^2 = \Phi_{\frac{1}{\mu^t}}(\Sigma)V^T
\end{equation}

where \(\Sigma V^T\) is the singular value decomposition of the matrix \(Z^t + Y^t_2/\mu^t\). \(Y^t_2\) is a Lagrange multiplier, \(\mu^t > 0\) is a penalty parameter and \(\Phi_{\frac{1}{\mu^t}}(\cdot)\) is the soft-thresholding operator defined as follows [12]

\[
\Phi_{\frac{1}{\mu^t}}(x) = \begin{cases} 
    x - \zeta, & \text{if } x > \zeta \\
    x + \zeta, & \text{if } x < -\zeta \\
    0, & \text{otherwise}
\end{cases}
\]

**Updating Z** Fix the other variables and solve the following problem

\begin{equation}
Z^{t+1} = \text{argmin}_Z \{ ||W \odot Z||_F^2 + \frac{\mu^t}{2} (||(X - E^t + Y^t_1/\mu^t)) - D^tZ||_F^2 + ||Z - (J^{t+1} - Y^t_2/\mu^t)||_F^2 + ||Z - (L^t - Y^t_3/\mu^t)||_F^2 \}
\end{equation}

In order to update \(Z\) as a whole, we denote \(||W \odot Z||_F^2 = ||Z - B^t||_F^2\), where \(B^t = Q \odot Z^t, Q = S - W, S \in R^{K \times N}\) is a matrix whose elements are all 1s. Thus, \(Z^{t+1}\) can be updated by

\begin{equation}
Z^{t+1} = \left( D^T \left( D^T + (\alpha/\mu^t + 2I) \right) \right)^{-1} \left(D^T \left( X - E^t \right) + L^t + J^{t+1} + (\alpha B^t + D^T Y^t_1 - Y^t_2 - Y^t_3)/\mu^t \right)
\end{equation}

**Updating L** Fix the other variables and solve the following problem

\begin{equation}
L^{t+1} = \text{argmin}_L \frac{\beta}{\mu^t} ||L||_1 + \frac{1}{2} ||L - (Z^{t+1} + Y^t_3/\mu^t)||_F^2 = \Phi_{\frac{\beta}{\mu^t}}(Z^{t+1} + Y^t_3/\mu^t)
\end{equation}

**Updating E** Similar with \(L\), \(E\) is updated by

\begin{equation}
E^{t+1} = \Phi_{\frac{\beta}{\mu^t}}(X - D^T Z^{t+1} + Y^t_3/\mu^t)
\end{equation}

**Updating D** The updating scheme of \(D\) is

\begin{equation}
D^{t+1} = \gamma D^t + (1 - \gamma) \frac{1}{\mu^t} (Y^t_1 + \mu^t (X - E^{t+1})) Z^{t+1} (Z^{t+1} Z^{t+1T})^{-1}
\end{equation}

where \(\gamma > 0\) is a parameter that controls the updating step. It is experimentally set 0.1.

4. Experiments and results

We evaluate our approach on three public databases, in which both training and testing images are corrupted, including illumination changes, occlusions and block noises. Our approach is compared with several latest related works, including FDDL[6], DLRRD SR[7], D^2LSR^3[8], DSLR[9], SC-LCP[10] and DODL[11]. The parameters \(\alpha, \lambda, \text{and} \beta\) in the dictionary learning phase are set as 0.5, 0.1 and 0.01 respectively. We repeat each experiment 5 times and report the average accuracy.

4.1. Classification

We use a linear classifier for classification as in [9][10]. The linear classifier is learnt as follows

\[
\tilde{W} = \text{argmin}_W ||H - WZ||_F^2 + \kappa ||W||_F^2
\]

where \(H\) is class label matrix, \(\kappa\) is the weight of regularization term. Given the test data \(X_{test}\), the representations \(Z_{test}\) are calculated by solving (1) with \(\alpha=0\). The label for test sample \(i\) is given by

\[
l = \text{argmin}_l (s = \tilde{W}z_i)
\]

where \(s\) is the class label vector, \(z_i\) is the \(i\)-th column vector in \(Z_{test}\).

4.2. AR database
A subset from AR database [14] is used, which consists of 2600 images, corresponding to 100 individuals (50 males and 50 females). For each individual, there are two sessions, including 14 non-occluded images with different facial expressions or illumination variations and 12 occluded images (6 by sunglasses and 6 by scarves). Experiments are performed under three different scenarios.

**Sunglasses:** Seven non-occluded images and one sunglasses image from session 1 are chosen as training samples, the rest seven non-occluded images and five sunglasses images as testing.

**Scarf:** Seven non-occluded images and one scarf image from session 1 are chosen as training samples, the rest seven non-occluded images and five scarf images as testing.

**Mixed (Sunglasses + Scarf):** Seven non-occluded images, one sunglasses image and one scarf image from session 1 are chosen as training samples, the rest are used for testing.

The comparison of different methods on AR database is listed in table 1. Our approach achieves the best results and outperforms the state-of-art method by 0.4% for the sunglasses scenario, 0.9% for the scarf scenario, and 0.8% for the mixed scenario.

| Method      | Sunglasses | Scarf | Mixed |
|-------------|------------|-------|-------|
| FDDL        | 89.8       | 90.5  | 89.8  |
| DLRD_SR     | 91.7       | 89.4  | 88.6  |
| SDLR        | 86.8       | 73.7  | 74.4  |
| DODL        | 91.6       | 90.1  | 87.4  |
| D^2L_2R^2   | 91.4       | 90.4  | 89.1  |
| SCLRDL      | 89.6       | 87.2  | 84.1  |
| Ours        | 92.1       | 91.4  | 90.6  |

### Extended YaleB database

The Extended YaleB database[15] contains 2,414 frontal-face images of 38 individuals. We first resize the image size to 48×42, and then randomly select half images for each person as training images and the rest as testing images. Comparison of different methods on Extended YaleB database is listed in table 2. It can be seen that our approach outperforms other competing methods. It achieves 1.9% improvement over FDDL, and 0.5% improvement over low-rank based dictionary learning methods.

We also present visualization of some decomposition results of testing image in figure 1(a). The top four images show the original images, while the middle four images and the bottom four images are corresponding to the low-rank recovery term $DZ$ and the sparse error term $E$. It turns out that illumination changes can be removed as sparse noises and the diversity of images are preserved well. Figure 1(b) shows representations of the first five class testing samples corresponding to the first five class reconstruction bases. The learned representation has class-wise block-diagonal structure.

| Method      | Sunglasses | Scarf | Mixed |
|-------------|------------|-------|-------|
| FDDL        | 97.2       | 96.8  | 98.6  |
| DLRD_SR     | 98.2       | 95.5  | 97.8  |
| SDLR        | 96.8       | 98.6  | 95.5  |
| DODL        | 97.2       | 97.8  | 99.1  |
| D^2L_2R^2   | 95.5       | 97.8  | 99.1  |
| SCLRDL      | 97.8       | 97.8  | 99.1  |
| Ours        | 99.1       | 99.1  | 99.1  |

### ORL database

The ORL database [16] contains 400 images of 40 individuals. A randomly located block of each image is replaced with an unrelated random image. For each individual, half images are randomly chosen as the training samples, and the rest are used as testing samples. Experiments are performed under different levels of block-corruption, which is measured by the area proportion between the block and the whole image. Comparison of different methods on ORL database is listed in table 3. When images are clear, FDDL achieves the best performance. With the corruption area increasing, WDDL_LRR performs better than other methods, and achieves 0.1%, 0.8%, 1.0%, 1.5% and 5.9% improvement respectively.
5. Conclusion
This paper presented weighted discriminative dictionary learning based on low-rank representation, which has a strong discriminative capability for image classification, especially, when the images are corrupted with big noise. By taking the class-wise block-diagonal structure of representation into consideration, a weighted representation term is constructed to enhance the discriminative power of the dictionary. The WDDL_LRR was extensively evaluated on several public databases. The experimental results confirmed that WDDL_LRR outperforms some state-of-the-art methods.

Table 3. The classification accuracy (%) on ORL database.

| Corruption | FDDL | DLRD_SR | SDLR | DODL | D^L'R^2 | SCLRDL | Ours  |
|------------|------|---------|------|------|---------|--------|-------|
| 0          | 96.7 | 93.1    | 91.8 | 91.0 | 94.3    | 95.1   | 94.3  |
| 10         | 86.8 | 90.4    | 84.3 | 78.0 | 91.1    | 91.2   | 91.3  |
| 20         | 74.4 | 81.2    | 78.3 | 59.0 | 82.6    | 80.6   | 83.4  |
| 30         | 61.8 | 76.5    | 66.7 | 40.3 | 77.2    | 63.7   | 78.7  |
| 40         | 49.1 | 67.8    | 56.2 | 28.7 | 68.9    | 50.5   | 74.8  |
| 50         | 36.8 | 57.2    | 47.8 | 22.0 | 59.8    | 39.2   | 68.3  |

References
[1] Guangcan L, Zhouchen L, Shuicheng Y, Ju S, Yong Y and Yi M. 2013. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on PAMI*, 35(1), pp 171–184.
[2] Chih C, Chia W and Yu W. 2012. Low-rank matrix recovery with structural incoherence for robust facerecognition. *CVPR*.
[3] Alhussein F, Mike D and Pascal F. 2015. Dictionary learning for fast classification based on soft-thresholding. *International Journal of Computer Vision, 114*(2-3), pp 306-321.
[4] Meng Y, Dengxin D, Linlin S and Luc G. 2014. Latent dictionary learning for sparse representation based classification. *CVPR*.
[5] Zhuolin J, Zhe L and Larry S D. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on PAMI*, 35(11), pp 2651–2664.
[6] Meng Y, Lei Z, Xiangchu F and David Z. 2014. Sparse representation based fisher discrimination dictionary learning for image classification. *IJCV, 109*, pp 209–232.
[7] Long M, Chunheng W, Baihua X and Wen Z. 2013. Sparse representation for face recognition based on discriminativelow-rank dictionary learning. *CVPR*.
[8] Sheng L, Liangyue L and Yun F. 2014. Low-rank and sparse dictionary learning. *Image and Vision Computing*, pp61–85.
[9] Yangmuzi Z, Zhuolin J and Larry S D. 2013. Learning structured low-rank representations for image classification. *CVPR*.
[10] Yong L, Jing L, Zechao L, Yangmuzi Z, Hanqing L and Songde M. 2014. Learning low-rank representations with class-wise block-diagonal structure for robust face recognition. *AAAI*.
[11] Stephen B, Neal P, Eric C, Borja P and Jonathan E. 2011. Distributed optimization and
statistical learning via the alternating direction method of multipliers, *Foundation and Trends in Machine Learning*, 3 (1), pp 1-122

[12] Zhouchen L, Minming C and Yi M. 2010. The augmented lagrangemultiplier method for exact recovery of corrupted low-rank matrices. *Technical report, UIUC.*

[13] Zhen D, Mingtao P and Yunde J. 2015. Discriminative orthonormal dictionary learning for fast low-rank representation. *ICONIP.*

[14] Aleix M and Robert B. 1998. The AR face database. *Technical report, CVC.*

[15] Kuang C L, Jeffrey H and David J K. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on PAMI, 27*(5):684–698.

[16] Ferdinando S and ACHarter. 1994. Parameterisation of stochastic model for human face identification. *WACV*, 138 – 142.