Kernel Based Weighted Group Sparse Representation Classifier

Bingxin Xu[1], Ping Guo[1,⋆], and C.L. Philip Chen[2]

[1] Image Processing and Pattern Recognition Laboratory
Beijing Normal University, Beijing, China
[2] Faculty of Science and Technology
University of Macau, Macau, China
xbing@mail.bnu.edu.cn, pguo@ieee.org, philipchen@umac.mo

Abstract. Sparse representation classification (SRC) is a new framework for classification and has been successfully applied to face recognition. However, SRC can not well classify the data when they are in the overlap feature space. In addition, SRC treats different samples equally and ignores the cooperation among samples belong to the same class. In this paper, a kernel based weighted group sparse classifier (KWGSC) is proposed. Kernel trick is not only used for mapping the original feature space into a high dimensional feature space, but also as a measure to select members of each group. The weight reflects the importance degree of training samples in different group. Substantial experiments on benchmark databases have been conducted to investigate the performance of proposed method in image classification. The experimental results demonstrate that the proposed KWGSC approach has a higher classification accuracy than that of SRC and other modified sparse representation classification.

Keywords: Group sparse representation, kernel method, image classification.

1 Introduction

Sparse representation or sparse coding has been successfully applied to many computer vision tasks, including face recognition [1], image super-resolution [2] and image classification [3]. Sparse representation classification (SRC) is a new framework which assumes that the test sample can be represented by linear combination of training samples which belong to the same class. However, SRC only uses L1-norm as regularization item to control the sparsity of linear coefficients. L1-norm treats each training sample equally and doesn't consider the cooperation of training samples from the same class. Zou [4] proves that L1-norm only selects a single sample from a group of correlated training samples. Therefore, when the test sample has a similar training sample which label is different with it, this training sample will possibly lead to wrong classification result.

[⋆] Corresponding author.

M. Kurosu (Ed.): Human-Computer Interaction, Part V, HCII 2013, LNCS 8008, pp. 236–245, 2013.

In order to solve the problem mentioned above, the group sparse classifier (GSC) is proposed by Majumdar [5]. GSC employs a L1-norm mixed L2-norm as regularization item which make coefficients belong to the same group are dense but among groups are sparse. Even though GSC considers the cooperation of training samples in the same group, it is not perfect either. The reason is that the inner L2-norm of mixed norm selects all the training samples from a particular class [6]. Actually, representing a test sample does not need all the training samples even thought they are in the same class. Due to the diversity of training samples, only some of them are more similar with test sample and these samples play an important role in linear representation.

A kernel based weighted group sparse classifier (KWGSC) is proposed in this paper to solve the shortcomings of SRC and GSC. The kernel method is successfully used in many algorithms for pattern analysis and the famous one is support vector machine (SVM). The effect of kernel method is mapping the original data into a high dimensional feature space by non-linear transformation. In order to improve the representation ability of SRC, kernel based sparse representation classification (KSRC) algorithms have been proposed in [7] [8]. However, these algorithms only replace the original space to the kernel space to compute distance. Actually, the value of kernel function which computing the inner product of a pair of data indicates the similarity of these data in the kernel space. Therefore, kernel function is directly applied to feature extraction in this paper. For the weight of each group, we also consider the influence of the similarity between test sample and members of each group.

The rest of this paper is organized as follows: In section 2, the background of this work is discussed. The proposed kernel based weighted group sparse classifier is described in section 3. Experimental design and results are presented in section 4 and the conclusion is presented in section 5.

2 Related Work

In this section, we briefly introduce the SRC method proposed in [1] and GSC method proposed in [5]. Finally, the kernel trick is reviewed.

2.1 Sparse Representation Classifier

Sparse representation classifier assumes that the training samples from a single class do lie on a subspace. Ideally, a test sample from one class can be represented by a linear combination of training samples from the same class. Specifically, given n_i training samples from the i-th class, the samples' feature vectors are stacked as columns of a matrix $\mathbf{F}_i = [\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \ldots, \mathbf{f}_{i,n_i}] \in R^{m \times n_i}$. Any new test sample $\mathbf{y} \in R^m$ from the same class can be represented as:

$$\mathbf{y} = x_{i,1}\mathbf{f}_{i,1} + x_{i,2}\mathbf{f}_{i,2} + \ldots + x_{i,n_i}\mathbf{f}_{i,n_i} = \mathbf{x}_i\mathbf{F}_i, \tag{1}$$

where $\mathbf{x}_i$ is the coefficient vector of linear representation. Since the class label of the test sample is unknown, a new matrix $\mathbf{F}$ is defined by concatenating all the classes:

$$\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_i, \ldots, \mathbf{F}_c], \tag{2}$$

where $i = 1, 2, \ldots, c$. Then the linear representation of $\mathbf{y}$ can be rewritten in terms of all the training samples as:

$$\mathbf{y} = \mathbf{F}\mathbf{x} \in R^m, \tag{3}$$

where $\mathbf{x}$ is the vector of coefficients. If $\mathbf{y}$ belongs to i-th class, the entries of $\mathbf{x}$ are expected to be zero except some of those associated with this class. Namely $\mathbf{x} = [0, 0, \ldots, \mathbf{x}_i, \ldots, 0]^T \in R^N$ and N is the total number of training samples. In SRC, the objective function is formulated as a convex programming problem:

$$\min \|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_1, \tag{4}$$

where $\|.\|_1$ denotes the L1-norm. After computing the coefficient vector, the test sample is classified according to the reconstruction error between $\mathbf{y}$ and its approximations. The i-th class approximation is computed by using only the coefficients associated with class i and assigning zeros to other entries.

2.2 Group Sparse Classifier

The group sparse classifier has the same assumption as SRC. The difference is that GSC intends to ensure all the coefficients for the correct class are selected. It means that the coefficients in the same group or class should be zeros or non-zeros simultaneously. L1-norm minimization cannot satisfy this condition, so the objective function of GSC is formulated as:

$$\min \|\mathbf{y} - \mathbf{F}\mathbf{x}\|_2 + \lambda \sum_{j=1}^{c} \|\mathbf{x}_{G_j}\|_2, \tag{5}$$

where $\mathbf{x}_{G_j}$ is the coefficients associated with group $\mathbf{G}_j$. Although GSC can ensure the group structure of coefficients, this may not be always desired. In practice, we expect most of the correct samples can be selected which are more similar to the test sample and exclude the samples which are different to the test sample even thought they are in the same class. The proposed method can solve this problem and will be detailed in section 3.

2.3 Kernel Trick

The kernel trick is a very well-know technique in machine learning [8]. It has been widely applied to pattern recognition and function approximation, such as support vector machine [9] [10], kernel-based clustering methods [11] [12] and kernel principal component analysis (KPCA) [13]. The kernel trick attracts much interest because it can easily generalize a linear algorithm to a non-linear algorithm. Actually, kernel method is a feature projection method which can map the original feature space into a higher feature space by a non-linear algorithm.

In the kernel space, the distribution of samples will be changed. The different class's samples are more separate and the same class's samples are more gathering. Usually, a Mercer kernel k can be expressed as:

$$k(x, y) = \phi(x)^T \phi(y), \tag{6}$$

where x and y are any two points, and ϕ is the implicit nonlinear mapping function associated with the kernel function k. The trick of kernel function is that we don' t need to know the expression of ϕ and just use kernel function to instead of its inner product. The commonly used kernel function are polynomial kernel: $k(x, y) = (1 + \langle x, y\rangle)^p$ and radial basis function (RBF) kernel: $k(x, y) = exp(-\|x - y\|_2^2)/\sigma^2$. p and σ are parameters need to be predefined before used.

3 Kernel-Based Weighted Group Sparse Classifier

3.1 Kernel-Induced Feature

The proposed method is different to other kernel-based methods for transforming the objective function into an inner product form. The kernel trick is directly used for feature extraction. In the kernel space, equation (6) is a distance measure between $\mathbf{x}$ and $\mathbf{y}$. This kernel-induced distance is a non-Euclidean distance measure in original data space and has been used in kernel clustering [14]. Therefore, the different value of kernel function can describe the similarity between points. Specifically, the training samples construct the dictionary $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, N is the total number of training samples. For data $\mathbf{x}_i$, the feature transformation can be formulated as:

$$\mathbf{f}_i = k(\mathbf{X}, \mathbf{x}_i) = [k(\mathbf{x}_1, \mathbf{x}_i), k(\mathbf{x}_2, \mathbf{x}_i), \ldots, k(\mathbf{x}_N, \mathbf{x}_i)]^T, \tag{7}$$

where k is predefined kernel function and different kernel function transforms the original data into a different feature space. For the feature vector $\mathbf{f}_i$, if the point $\mathbf{x}_j \in \mathbf{X}$ is closer to $\mathbf{x}_i$, the value of $k(\mathbf{x}_j, \mathbf{x}_i)$ will be larger than others. Because the dictionary $\mathbf{X}$ contains all the class samples, the kernel-induced feature vectors are discriminative in different classes.

3.2 Group Construction

In the proposed method, we construct a new group members based on the original training samples and the numbers of each group are adaptive. For each class's training samples, a kernel matrix is computed and using the minimal value of the matrix as a threshold to select the members of this group. Specifically, $\mathbf{X}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}]$ is the i-th class samples and n_i is the total number of samples in this class. The kernel matrix of i-th class is computed as follows:

$$K_i = \begin{bmatrix} k(\mathbf{x}_{i1}, \mathbf{x}_{i1}) & k(\mathbf{x}_{i1}, \mathbf{x}_{i2}) & \ldots & k(\mathbf{x}_{i1}, \mathbf{x}_{in_i}) \\ k(\mathbf{x}_{i2}, \mathbf{x}_{i1}) & k(\mathbf{x}_{i2}, \mathbf{x}_{i2}) & \ldots & k(\mathbf{x}_{i2}, \mathbf{x}_{in_i}) \\ \cdots & \cdots & \cdots & \cdots \\ k(\mathbf{x}_{in_i}, \mathbf{x}_{i1}) & k(\mathbf{x}_{in_i}, \mathbf{x}_{i2}) & \ldots & k(\mathbf{x}_{in_i}, \mathbf{x}_{in_i}) \end{bmatrix}. \tag{8}$$

For this class, a threshold is defined as:

$$m_i = \min(K_i(:)), \tag{9}$$

which used for selecting the members of this group. For a test sample $\mathbf{y}$, only the training samples which are similar to it will be selected to construct the group. Others which are different from it are considered that has little effect in representing the test sample. For example, in the i-th class, the kernel-induced feature of $\mathbf{y}$ on $\mathbf{X}_i$ is computed as $\mathbf{f}_{i,y} = k(\mathbf{X}_i, \mathbf{y})$ by equation (7). If $k(\mathbf{x}_{i,j}, \mathbf{y})$ is bigger than the threshold m_i, it represents the sample $\mathbf{x}_{i,j}$ is similar to $\mathbf{y}$ and should be selected in group i. Therefore, the number of members in each group for test sample $\mathbf{y}$ is dependent on the data. The proposed method doesn't use all the training samples as dictionary mechanically, but according to the relationship between the test sample and the training samples to select the member of each group.

3.3 Weighted Group Sparse Classifier

Assume that the new dictionary $\mathbf{G}$ is partitioned into c disjoint groups $\mathbf{G}_1$,$\mathbf{G}_2$,..., $\mathbf{G}_c$, c is the number of groups and $\mathbf{G}_i \cap \mathbf{G}_j =\varnothing$ when $i \neq j$. For different test sample, the dictionary $\mathbf{G}$ is not fixed and computed by the method in section 3.2. In each group $\mathbf{G}_i$, there are N_i training samples to represent the test data. The objective function of proposed method is defined as:

$$\min \|\mathbf{k(G,y)} \text{ - } \mathbf{k(G,G)x}\|_2 + \sum_{i=1}^{c} w_i \|\mathbf{x}_{G_i}\|_2, \tag{10}$$

where $k(\mathbf{G}, \mathbf{y})$ is the kernel-induced feature vector of $\mathbf{y}$ and the column of $k(\mathbf{G}, \mathbf{G})$ is the kernel-induced feature vector of each member of $\mathbf{G}$. $\mathbf{x}_{G_i}$ is the coefficients associated with group i and w_i is the weight of this group which represents the importance of this group for reconstruction the test sample. w_i is defined as:

$$w_i = \frac{N_i}{N} + sum(k(\mathbf{G_i}, \mathbf{y})), \tag{11}$$

where N is the total number of members in dictionary $\mathbf{G}$. The first item N_i/N describes the ratio between the numbers of group i and the total number of all the groups. If the number of samples in group i is more, w_i should be larger to indicate i-th group samples are more important to represent the test sample. The second item of equation (11) has the same effect to control the weights. If the samples in group i are similar to the test data $\mathbf{y}$, the value of this item will be larger and lead to the weight becomes larger. In order to avoid noise or outlier make the value of second item big, the first item can balance the weight to an appropriate value. After computing the coefficients, the test sample can be classified to the class that minimizes the residual:

$$\min R_i(\mathbf{y}) = \|k(\mathbf{G}, \mathbf{y}) - k(\mathbf{G}, \mathbf{G})\delta_i(x)\|_2. \tag{12}$$

$\delta_i(x)$ defines the coefficient vector which only retains the coefficients related to group i and set other entries to zero. The complete algorithm is described as follows:

1. Input: a matrix of training samples $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_c] \in \mathbf{R}^{m \times N}$ for c classes and a test sample $\mathbf{y}$.
2. Compute the kernel matrix $\mathbf{K}_i$ and threshold m_i for each class by equation (8) and (9).
3. Compute the test sample's kernel-induced feature on each class, $\mathbf{f}_{i,y} = k(\mathbf{F}_i, \mathbf{y})$.
4. Construct the group $\mathbf{G}_i$ for each class and compute the weight w_i for group $\mathbf{G}_i$ by equation (11).
5. Solve the minimization problem defined by equation (10) and compute the residual $R_i(\mathbf{y})$ of each class.
6. Output: identity($\mathbf{y}$) = arg min($R_i(\mathbf{y})$).

4 Experiments

In this section, the effectiveness of the proposed kernel-based weighted group sparse classifier (KWGSC) algorithm is carried out on three facial image dataset, namely, ORL dataset [15], Extended Yale B dataset [16] [17] and AR dataset [18]. Downsampled is used for reducing the dimension of original image. In our experiments, polynomial kernel function is used for all the dataset and the parameter p is set to 2. We compare the classification ability of KWGSC algorithm with SRC [1], GSC [5] and KSRC [8]. The optimization method used to solve the group sparse problem is alternating direction method which proposed in [19]. The experiments are carried out ten times and the average classification rate is the final result.

4.1 ORL Database

ORL database has 10 different images for each class and contains 40 distinct individuals. The images of each individual are variations in pose and facial expression. The size of each face image is 112 × 92 with 256 gray levels per pixel. The downsampling ratio is 1/8 and the resulting standardized input space dimension is 168. Fig.1 shows sample images of one person. In the experiment, five images are random selected from each individual as training samples and the rest as testing samples. Therefore, the number of images for both training and test are 200. The experimental results are presented in table 1. In order to only compare the performance of different classifier, we set the same to original feature dimension and kernel function. From table 1, it can be discovered that GSC has worse result than SRC. This is because in ORL database, the size of each group is small with only five members and the benefit of group sparsity is not significant. In according with the conclusion proposed in [20], the result is that group sparsity favors large sized groups. However, the proposed KWGSC

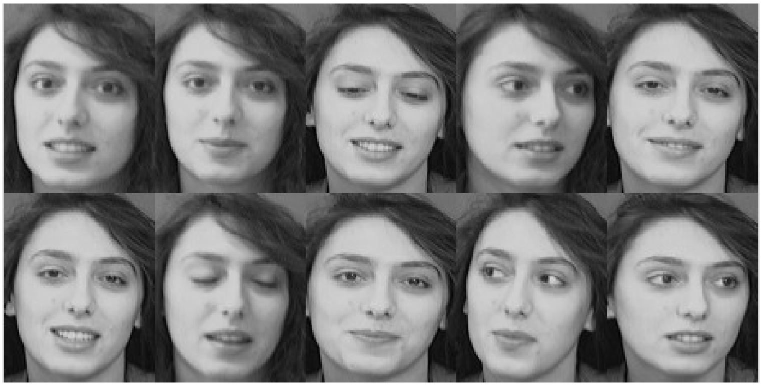

Fig. 1. Sample images of one person with different expression in ORL database

Table 1. Comparison results of various methods on the ORL database

Method	SRC	KSRC	GSC	KWGSC
Classification rate	91.5%	92.5%	87.5%	94.1%
Feature Dimension	168	168	168	168

has better result than SRC and GSC. This indicates that an appropriate weight of group can overcome the shortcoming of small sized group and improve the performance of group sparsity.

4.2 Extended Yale B Database

The Extended Yale B database consists of 2414 frontal face images of 38 individuals (about 64 images per category) captured under various laboratory controlled lighting conditions. For each category, we randomly select 32 images for training with the remaining images for testing. Therefore, the total number images for training and test is 1216 and 1198. The size of each face image is 192×168 with 256 gray levels per pixel. The downsampling ratio is 1/16 and the downsampled image is 12×11. Therefore the original standardized input space dimension is 132. Fig.2 shows sample images of one person with different lighting condition. The experimental results are presented in Table 2. The proposed method KWGSC is superior to others. Although the images in this database have much different lighting condition, KWGSC can select more important training samples automatically to represent the test image. It means that if the test image is very dark, then the training samples in the same lighting condition will be more effective than others even in the well lighting condition.

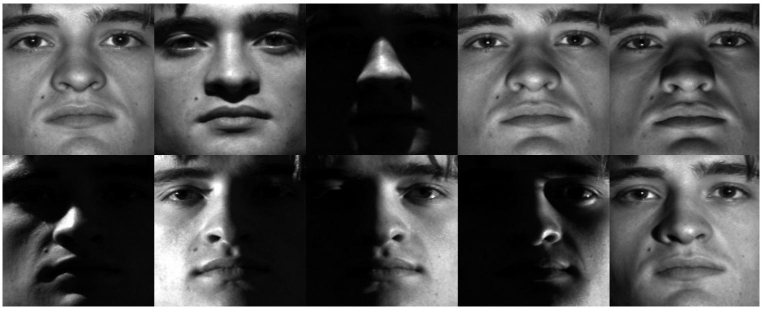

Fig. 2. Sample images of one person with different lighting condition in Extended Yale B database

Table 2. Comparison results of various methods on the Extended Yale B database

Method	SRC	KSRC	GSC	KWGSC
Classification rate	93.4%	94.8%	94.5%	95.6%
Feature Dimension	132	132	132	132

4.3 AR Database

The AR database consists of over 4,000 frontal images for 126 individuals. For each individual, 26 pictures were taken in two separate sessions [16]. These images include more facial variations including illumination change and expressions than the Extended Yale B database. The same with other methods, only a subset of the data are chosen to consist the database which contains 50 male individuals and 50 female individuals. For each individual, only 14 images with illumination change and expressions are selected. The image are cropped with dimension 165 × 120 and converted to gray scale [1]. The downsampling ratio is 1/12 and the downsampled image is 14×10. Therefore the original standardized input space dimension is 140. Fig.3 shows sample images of one person which are in different lighting condition and expression. We also partition each class averagely into training set and testing set. Table 3 lists the classification rates of each method. From Table 3, it can be concluded that KWGSC also has better result than others. The difficult of AR database is that the total number of individuals is large and the number of training samples in each class is small. However, KWGSC can avoid these problems. Although the total number of individuals is large, KWGSC excludes some training samples which have little relationship to the test sample in the group construction process.

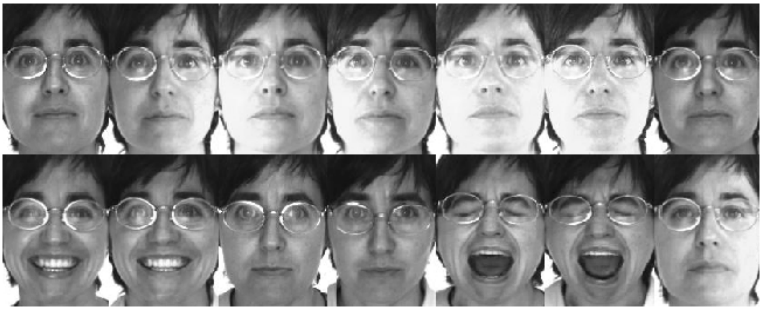

Fig. 3. Sample images of one person with different expression and lighting condition in AR database

Table 3. Comparison results of various methods on the AR database

Method	SRC	KSRC	GSC	KWGSC
Classification rate	89.4%	90.7%	91%	91.5%
Feature Dimension	140	140	140	140

5 Conclusion

In this paper, a kernel-based weighted group sparse classifier (KWGSC) algorithm is proposed. For KWGSC, samples are mapped from original feature space into a kernel-induced feature space, and then construct a new dictionary which depends on the test sample. The new dictionary contains many groups which have much relationship to test sample and excludes others based on kernel-induced distance measure. We compare the proposed method with SRC, GSC and KSRC on different facial image database. The experimental results indicate the effectiveness of KWGSC.

Acknowledgment. The research work described in this paper was fully supported by the grants from the National Natural Science Foundation of China (Project No. 90820010, 60911130513).

References

1. Wright, J., Yang, A.Y., Granesh, A.: Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
2. Yang, J.C., Wright, J., Huang, T., Ma, Y.: Image super-resolution as sparse representation and raw patches. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
3. Xu, B.X., Hu, R.K., Guo, P.: Combining affinity propagation with supervised dictionary learning for image classification. Neural Computing and Applications (in press), doi:10.1007/s00521-012-0957-7

4. Zou, H., Hastie, T.: Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 67(2), 301–320 (2005)
5. Majumdar, A., Ward, R.K.: Classification via group sparsity promoting regularization. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 861–864 (2009)
6. Majumdar, A., Ward, R.K.: Improved group sparse classifier. Pattern Recognition Letters 31(13), 1959–1964 (2010)
7. Yin, J., Liu, Z.H., Jin, Z., Yang, W.K.: Kernel sparse representation based classification. Neurocomputing 77(1), 120–128 (2012)
8. Zhang, L., Zhou, W.D., Chang, P.C., Liu, J., Yan, Z., Wang, T., Li, F.Z.: Kernel sparse representation-based classifier. IEEE Transactions on Signal Processing 60(4), 1684–1695 (2012)
9. Vapnik, V.N.: Statistical learning theory. Wiley-Interscience, New York (1998)
10. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discovery 2(2), 121–167 (1998)
11. Kin, D.W., Lee, K.Y., Lee, L.K.H.: Evaluation of the performance of clustering algorithms in kernel-iinduced feature space. Patern Recoginition 38(4), 607–611 (2005)
12. Graves, D., Pedrycz, W.: Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. Fuzzy Sets and Systems 161(4), 522–543 (2010)
13. Scholkopf, B., Smola, A.J., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
14. Chen, S.C., Zhang, D.Q.: Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. IEEE Transactions on Systems, Man, and Cybernetics, part B: Cybernetics 34(4), 1907–1916 (2004)
15. Samaria, F.S., Harter, A.C.: Parameterisation of a stochastic model for human face identification. In: Sarasota, F.S. (ed.) 2nd IEEE Workshop on Applications of Computer Vision, pp. 138–142 (1994)
16. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 643–660 (2001)
17. Lee, K., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5), 684–698 (2005)
18. Martinez, A., Benavente, R.: The AR face database. CVC Tech. Report. 24 (1998)
19. Deng, W., Yin, W.T., Zhang, Y.: Group sparse optimization by alternating direction method. Technical Report TR11-06, Department of Computational and Applied Mathematics, Rice University (2011)
20. Huang, J.Z., Zhang, T.: The benefit of group sparsity. Annals of Statistics 38(4), 1978–2004 (2010)