Incoherent Dictionary Learning for Sparse Representation

Tong Lin, Shi Liu, and Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
tonglin123@gmail.com, liushi0422@gmail.com, zha@cis.pku.edu.cn

Abstract

Recent years have witnessed a growing interest in the sparse representation problem. Prior work demonstrated that adaptive dictionary learning techniques can greatly improve the performance of sparse representation approaches. Existing techniques mainly focus on the reconstructive accuracies and the discriminative power of the learned dictionary, whereas the mutual incoherence between any two basis atoms has been rarely studied yet. This paper proposes a novel method by explicitly incorporating a correlation penalty into the dictionary learning model. Experiments show that the proposed method can remarkably reduce the correlation measure of the learned dictionaries, and at the same time achieve higher classification accuracies than state-of-the-art algorithms.

1. Introduction

This paper focus on the dictionary learning problem in sparse representation framework:

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2, \\ &subject\ to\ \|\mathbf{d}_j\|_2 = 1,\ j = 1,...,K, \\ &\|\mathbf{x}_i\|_0 \le T,\ i = 1,...,N, \end{aligned} \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$ in $\mathbb{R}^{m\times N}$ is the input data matrix that contains N vectors of m dimensions, $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_K]$ in $\mathbb{R}^{m\times K}$ is the dictionary to be learned, and $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$ in $\mathbb{R}^{K\times N}$ is the sparse vector matrix. The first constraint imposes a unit l_2-norm for each basis atom in the dictionary, while the second is for the sparsity prior with a bound parameter T. The goal of dictionary learning is to adapt a nonparametric dictionary for providing better performance than using a fixed dictionary.

The earliest work in dictionary learning was the MOD[2], and the most popular method is the K-SVD[1]. Most prior works aimed at improving both reconstructive and discriminating power of the learned dictionary when class labels can be available for the input data. Zhang and Li [8] developed the D-KSVD algorithm by directly incorporating the labels of training data, while Pham and Venkatesh [6] integrated a linear classifier into the dictionary learning model. In [5] and [4], Mairal established a discriminative model via a shared dictionary and multiple class-decision functions. In [7] the relationship between two dictionaries for different classes was addressed, but the relationship among atoms of one single shared dictionary had not been investigated yet. In this paper, our purpose is to develop a discriminative dictionary learning model that maximizes the incoherence of basis atoms in one single output dictionary. Experimental results on two face databases and one digits database demonstrate that the proposed method outperforms K-SVD and D-SKVD algorithms by providing higher classification accuracies and meanwhile reducing the correlation of the basis atoms among the learned dictionaries.

The paper is organized as follows. In Section 2 we propose the Incoherent Dictionary Learning(IDL) method and describe the algorithm implementation. Section 3 shows the experimental results and Section 4 concludes the paper.

2. The Proposed Algorithm

2.1. Incoherent Dictionary Learning (IDL) model

We first define the correlation measure of a dictionary $\mathbf{D}$ as:

$$cor(\mathbf{D}) = \|\mathbf{D}^T\mathbf{D} - \mathbf{I}\|_F^2, \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{K\times K}$ is an identity matrix. Clearly, the correlation measure is zero for a dictionary $\mathbf{D}$ whose columns are orthonormal, and in this case we say $\mathbf{D}$ most incoherent. It is worth noting that this definition is not a trivial measure, as

$$\|\mathbf{D}^T\mathbf{D} - \mathbf{I}\|_F^2 = tr((\mathbf{D}^T\mathbf{D} - \mathbf{I})^T(\mathbf{D}^T\mathbf{D} - \mathbf{I})) \tag{3}$$

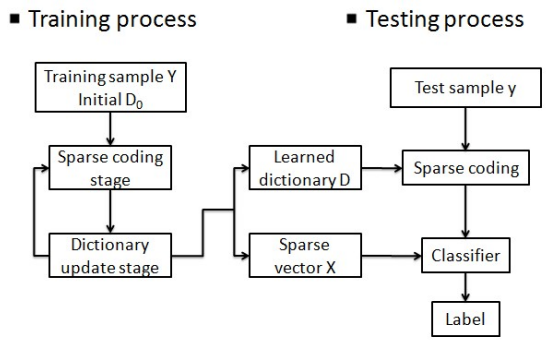

Figure 1. Flowchart of the proposed dictionary learning method for classification.

is a four degree term with respect to $\mathbf{D}$. Our goal is to maximize this incoherence to efficiently represent the input data. Fig.1 shows the flowchart of dictionary learning for supervised classification problems.

In order to get a more discriminative dictionary, we adopt the method in [6] to integrate classification errors into the dictionary learning model:

$$\begin{aligned}\min_{\mathbf{D},\mathbf{X},\mathbf{W}} &\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\lambda\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_F^2\\ &+\eta\|\mathbf{H}-\mathbf{W}\mathbf{X}\|_F^2+\beta\|\mathbf{W}\|_F^2,\\ &subject\ to\ \|\mathbf{x}_i\|_0\le T,\ i=1,...,N,\end{aligned}\tag{4}$$

where $\mathbf{W}$ in $\mathbb{R}^{M\times K}$ denotes the linear classifier matrix, $\mathbf{H}=[\mathbf{h}_1,\mathbf{h}_2,...,\mathbf{h}_N]$ in $\mathbb{R}^{M\times N}$ is the class label matrix of the training data $\mathbf{Y}$, and M is the number of classes. Note that a class label vector $\mathbf{h}_i$ for sample $\mathbf{y}_i$ is a all-zero vector except for its j-th entry being one if $\mathbf{y}_i$ belongs to the j-th class. The third term measures the classification error by using a linear classifier $\mathbf{W}$ on the sparse feature vector matrix $\mathbf{X}$, while the fourth is a regularization of $\mathbf{W}$. There are three regularization parameters (λ, η, and β) that reflect relative contributions of each term.

The unit norm constraint has been dropped in the above model (4), because the l_2-norm of each column basis atom in $\mathbf{D}$ can not be far from 1 due to the correlation penalty imposed in the second term. We argue that the column atoms need not be of unit norm exactly, and eliminating the unit norm constraint would not sacrifice the reconstructive and discriminative ability of the learned dictionary. On the other hand, the dictionary updating can be much easier to speed up the algorithm by discarding the unit norm constraint.

2.2. Algorithm implementation

As the objective function in (4) is non-convex, finding a global solution might not be available in practise. Therefore, a three-step iteration procedure is utilized (listed in Alg.1), and details of each step are described as follows.

Algorithm 1 Algorithm for supervised IDL model.

Input:

Training data Y; $\mathbf{D}$ and $\mathbf{X}$ initialized through K-SVD.

Output:

Dictionary $\mathbf{D}$, sparse vector matrix $\mathbf{X}$, and classifier $\mathbf{W}$.

Repeat:

Classifier updating step to learn $\mathbf{W}$ by fixing $\mathbf{D}$ and $\mathbf{X}$.

Dictionary learning step to learn $\mathbf{D}$ by fixing $\mathbf{W}$ and $\mathbf{X}$.

Sparse coding step to learn $\mathbf{X}$ by fixing $\mathbf{D}$ and $\mathbf{W}$.

Until: Convergence of the objective function or maximal iteration times.

3.2.1. Classifier update step. When $\mathbf{D}$ and $\mathbf{X}$ are fixed, the objective function of $\mathbf{W}$ turns out to be

$$\min_{\mathbf{W}}\eta\|\mathbf{H}-\mathbf{W}\mathbf{X}\|_F^2+\beta\|\mathbf{W}\|_F^2.\tag{5}$$

Clearly, this unconstrained optimization is a multivariate ridge regression problem which can be solved directly by setting the partial derivatives with respect to $\mathbf{W}$ to zero. The global optimal solution of (5) is

$$\mathbf{W}=\mathbf{H}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T+\gamma\mathbf{I})^{-1},\tag{6}$$

where $\gamma=\beta/\eta$.

3.2.2. Dictionary learning step. Fixing $\mathbf{X}$ and $\mathbf{W}$, the optimization problem about dictionary $\mathbf{D}$ can be reduced to

$$\min_{\mathbf{D}}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\lambda\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_F^2.\tag{7}$$

We follow the spirit of K-SVD algorithm by updating each column of basis atom $\mathbf{d}_k$ among the dictionary $\mathbf{D}$ in a random order when fixing other atoms unchanged. The unconstraint problem of $\mathbf{d}_k$ can be solved by setting the partial derivatives with respect to $\mathbf{d}_k$ to zero, with the optimal solution written as

$$\mathbf{d}_k=(\mathbf{x}_T^k\mathbf{x}_k\mathbf{I}+2\lambda\mathbf{D}_k^*\mathbf{D}_k^{*T})^{-1}\mathbf{E}_k\mathbf{x}_k,\tag{8}$$

where $\mathbf{E}_k=\mathbf{Y}-\sum_{j\in J}\mathbf{d}_j\mathbf{x}_T^j$, $\mathbf{D}_k^*=\{\mathbf{d}_j\}_{j\in J}$, $J=\{1,...,k-1,k+1,...,K\}$, and $\mathbf{x}_T^k$ is the k-th row of $\mathbf{X}$.

The procedure is repeated until all atoms in the dictionary are updated once. Note that an updated atom can play a role immediately in updating following atoms, no need to wait until all atoms finish.

3.2.3. Sparse coding step. When fixing $\mathbf{D}$ and $\mathbf{W}$, the objective function with $\mathbf{X}$ can be written as

$$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \eta\|\mathbf{H} - \mathbf{W}\mathbf{X}\|_F^2, \quad subject\ to\ \|\mathbf{x}_i\|_0 \leq T,\ i = 1, ..., N. \tag{9}$$

By grouping into a larger matrix, we have

$$\min_{X} \left\| \begin{pmatrix} \mathbf{Y} \\ \sqrt{\eta}\mathbf{H} \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\eta}\mathbf{W} \end{pmatrix} \mathbf{X} \right\|_F^2. \tag{10}$$

This formulation is just same as the sparse representation model, hence can be efficiently solved by OMP or any other sparse representation algorithm.

2.3. Classification for a new test sample

Upon finishing the dictionary learning stage, we obtain a dictionary $\mathbf{D}$ and a linear classifier $\mathbf{W}$ simultaneously. For a new test sample $\mathbf{y}$, the OMP algorithm is used to compute its sparse vector $\mathbf{x}$ based on the learned dictionary $\mathbf{D}$. The sparse vector $\mathbf{x}$ can serve as an extracted feature vector for classification. For simplicity, we directly use the learned linear classifier $\mathbf{W}$ to make the label decision:

$$\text{label}(\mathbf{y}) = \arg\max_j \{\mathbf{z}_j \mid \mathbf{z} = [\mathbf{z}_1, ..., \mathbf{z}_M]' := \mathbf{W}\mathbf{x}\}. \tag{11}$$

3. Experimental Results

The proposed IDL model is compared with K-SVD and D-KSVD algorithms on two face databases and one digital number database. In our experiments, parameters λ and η are tuned using cross-validation in a 2-D parameter space. It was observed that the linear classifier $\mathbf{W}$ tends to be well bounded in the three step iterations, so the parameter β is set to be very small (or even zero). The image preprocessing step are same as in K-SVD and D-KSVD implementations: change each image into a column vector, normalize to unit norm, and then project into a lower dimensional space using PCA.

3.1. The Extended YaleB face database

The Extended YaleB database contains about 2414 frontal face images of 38 individuals (Figure 2). We randomly split the database into two halves for training and testing. Similar to the experimental settings in [8], face images are cropped as 168×192 pixels and then reduced to 504-dim. by PCA. The dictionary size is set to have 570 basis atoms, and the sparsity prior is set as $T = 16$.

Figure 2. Example images of the Extended YaleB Database.

Table 1. Classification accuracies and average cross-correlation coefficients on the Extended YaleB database.

Method	IDL	D-KSVD	K-SVD
accuracy	95.86%	94.70%	93.54%
correlation	0.2860	0.3680	0.3816

From Table 1 we can see that the proposed method achieves higher accuracies and at the same time reduces the cross-correlation coefficient of two basis atoms among the learned dictionary. Fig.3 illustrates histograms of correlation coefficients for learned dictionaries by K-SVD, D-KSVD, and IDL, showing that IDL yields a highly incoherent dictionary.

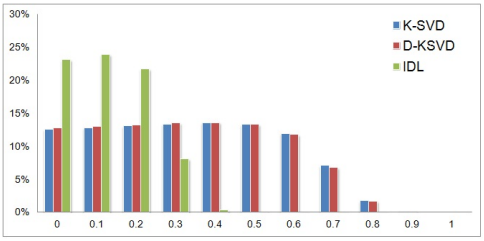

Figure 3. The histogram of correlation coefficients for each pair of atoms in the learned dictionary by K-SVD(blue), D-KSVD(red), and IDL(green).

3.2. The CAS-PEAL-R1 face database

The CAS-PEAL-R1 Face Database [3] contains 30,900 images of 1,040 individuals with varying pose, expression, accessory and lighting (PEAL) (Fig. 4). We embed each cropped face image of 80×100 pixels into a 500-dim. space. The dictionary is set to contain 700 atoms, and $T = 16$. The results are summarized in Table 2. All methods achieve high accuracies (>90%)

on expression data set, while perform very poorly on the most challenging pose variations. In comparison, K-SVD yields the lowest accuracy while IDL performs the best.

Figure 4. Sample images of CAS-PEAL-R1 Face Database.

Table 2. Accuracies on CAS-PEAL-R1 data.

Method	IDL	D-KSVD	K-SVD
Pose	45.04%	36.78%	21.49%
Expression	95.45%	94.63%	90.91%
Accessory	58.68%	50.83%	36.36%
Lighting	57.85%	43.80%	38.02%

3.3. The USPS handwritten digital number database

The USPS data contains about 5500 handwritten digital number images of 16×16 pixels. The dictionary is set to have 300 atoms, and $T = 15$. From Table 3, we can see that our method offers higher testing accuracies than K-SVD and D-KSVD, demonstrating the advantages of incoherence constraints. Fig. 5 shows examples of basis atoms in learned dictionaries by IDL and K-SVD. It is interesting to see that basis atoms in IDL dictionary exhibit great variations in gray scales and blurred shape contours of digits, due to the correlation penalty.

4. Conclusion

In this paper, we propose a new method that integrates correlation penalty into dictionary learning model. Experimental results confirmed that our method can greatly enhance the incoherence degree of dictionary and meanwhile yield higher classification accuracies. The future work is to investigate the cross-correlation between the dictionary $\mathbf{D}$ and the training data $\mathbf{Y}$.

Acknowledgments

This work was supported by the National Basic Research Program of China (2011CB302202) and the National Science Foundation of China (61075119).

Table 3. Accuracies and correlation measures on USPS data.

Method	IDL	D-KSVD	K-SVD
Acc(train)	97.49%	98.51%	100%
Acc(test)	93.85%	93.38%	92.6%
correlation	0.2746	0.3748	0.3866

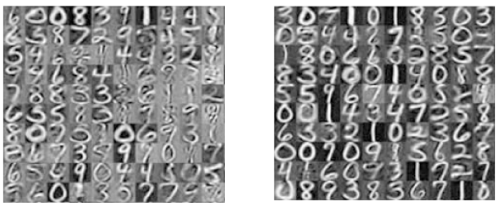

Figure 5. Dictionaries learned by IDL (left) and K-SVD (right).

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

[2] K. Engan, S. Aase, and J. Husoy. Frame based signal compression using method of optimal directions (MOD). In *Proc. IEEE Int. Symp. Circuits and Systems*, 1999.

[3] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(1):149–161, 2008.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Learning discriminative dictionaries for local image analysis. In *Proc. CVPR*, 2008.

[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Proc. NIPS*, pages 1033–1040, 2008.

[6] D. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *Proc. CVPR*, 2008.

[7] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proc. CVPR*, 2010.

[8] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *Proc. CVPR*, 2010.