Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Supervised and Projected Sparse Coding for Image Classification

Jin Huang and **Feiping Nie** and **Heng Huang*** and **Chris Ding**
Computer Science and Engineering Department
University of Texas at Arlington
Arlington,TX,76019
huangjinsuzhou@gmail.com, feipingnie@gmail.com, heng@uta.edu, chqding@uta.edu

Abstract

Classic sparse representation for classification (SRC) method fails to incorporate the label information of training images, and meanwhile has a poor scalability due to the expensive computation for ℓ_1 norm. In this paper, we propose a novel subspace sparse coding method with utilizing label information to effectively classify the images in the subspace. Our new approach unifies the tasks of dimension reduction and supervised sparse vector learning, by simultaneously preserving the data sparse structure and meanwhile seeking the optimal projection direction in the training stage, therefore accelerates the classification process in the test stage. Our method achieves both flat and structured sparsity for the vector representations, therefore making our framework more discriminative during the subspace learning and subsequent classification. The empirical results on 4 benchmark data sets demonstrate the effectiveness of our method.

Introduction

Sparse representation has been extensively studied in signal processing (Candes and Wakin 2008) and computer vision (Wright et al. 2008) areas. Despite its success in the applications, the standard sparse representation framework has several limitations: (i) In the learning process, the training data are first grouped into the column-based matrix, then the sparse representation of the test image is found via solving a convex problem which minimizes the empirical loss with an ℓ_1 regularization that introduces the sparsity. However, the label of the training data is not utilized, therefore the sparsity is based solely on the structure of the individual data. (ii) Sparse representation itself does not involve any dimension reduction process, solving a high dimensional ℓ_1 minimization problem is still computational expensive nowadays. When the atoms used for the sparse decomposition are the training samples themselves, then this framework usually requires the training matrix to be overcomplete (the number of training samples is larger than the individual image dimension), therefore sparse representation so far is applied to relatively small scale data sets.

*Corresponding Author

To address the issue caused by unsupervised training, we introduced the group sparse regularity term to incorporate the label information and therefore explore the sparse structure of the representation vector. To have better data scalability, we seek a projection matrix to effectively lower down the image dimension and meanwhile preserve the sparse structure for subsequent classification. During the training stage, we develop a novel supervised subspace learning algorithm using group sparse regularization term, named Supervised and Projected Sparse Coding (SPSC), which simultaneously optimizes the sparse vector representations and the projection direction. To the best of our knowledge, our algorithm is the first projected sparse coding algorithm using the label information. Our algorithm has several explicit advantages. First, extending from sparse representation, our algorithm incorporates the label information during each iteration of the projection matrix updating, therefore making our method more discriminative. Second, through the learning of the projection matrix, we can save more time in the classification stage than the time spent for training, therefore our proposed algorithm has the potential to solve large-scale problems. Last, our method has no explicit feature selection process and is robust to contiguous image occlusions. To demonstrate the effectiveness and advantage of the proposed method for image classification, extensive experiments have been performed on the four commonly used data sets. We compare the classification results of our method with multiple related methods. In particular, the experiments show that with the same subspace dimension, our method gets better classification results than other benchmark methods.

Sparse Representation Overview

In this section, we first briefly introduce the basic background of sparse representation and necessary notations for subsequent context. Suppose we have n images of size $r \times c$, reshape them to vectors, arrange them into columns of a training sample matrix

$$A = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathbb{R}^{m \times n}$$

where $m = r * c$ is the length of $\mathbf{x}_i$. Sparse representation based classification model assumes $n \gg m$, *i.e.*, A is an overcomplete system. Given a new input image $\mathbf{y}$, Sparse representation based classification model computes a repre-

sentation vector

$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T \in \mathbb{R}^n \quad (1)$$

to satisfy

$$\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i \alpha_i = A\boldsymbol{\alpha} \quad (2)$$

Since A is overcomplete, to seek a sparse solution, it is natural to solve

$$\min_{\boldsymbol{\alpha}} \|A\boldsymbol{\alpha} - y\|_2^2 + \lambda_0 \|\boldsymbol{\alpha}\|_0 \quad (3)$$

where ℓ_0 pseudo norm counts the number of non-zero elements in $\boldsymbol{\alpha}$ and $\lambda_0 > 0$ is the controlling parameter.

Recent discovery in (Donoho 2004; Candes and Tao 2006) found that the sparse solution in Eq. (3) could be approximated by solving the ℓ_1 minimization problem:

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1, \ s.t. \ A\boldsymbol{\alpha} = \mathbf{y} \quad (4)$$

or the equivalent penalty version:

$$\min_{\boldsymbol{\alpha}} \|A\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 \quad (5)$$

This ℓ_1 problem can be solved in polynomial time by standard linear programming methods (Chen, Donoho, and Saunders 1998).

While sparse representation has been applied successfully in the areas mentioned in the introduction, it also has two major limitations. 1) No label information has been used for the learning of $\boldsymbol{\alpha}$. 2) Poor data scalability, so far no efficient tool to solve large scale high dimension ℓ_1 minimization problem.

Supervised Group Sparse Coding

When applied to image classification, an important limitation of sparse representation is that the useful label information is not used in the classification. Note that in Eq. (5), the calculation of $\boldsymbol{\alpha}$ is totally based on the individual data structure without the label information, the non-zero elements in $\boldsymbol{\alpha}$ are although sparse(flat sparse), not so condensed. In order to take advantage of the training data labels, in this paper, we introduce the group ℓ_1 norm, which is defined as follows: assuming there are c classes in the total n images, each class has n_k images for $k = 1, \cdots, c$, the representation vector $\boldsymbol{\alpha}$ can be written into Eq. (6) considering its label information.

$$\boldsymbol{\alpha} = (\alpha_{1,1}, \cdots, \alpha_{1,n_1}, \cdots, \alpha_{c,n_1}, \cdots, \alpha_{c,n_c}), \quad (6)$$

We assume

$$A = (A_1, \cdots, A_c), \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_c)^T,$$

where

$$A_k = (\mathbf{x}_{k1}, \cdots, \mathbf{x}_{kn_k}), \quad \boldsymbol{\theta}_k = (\alpha_{k1}, \cdots, \alpha_{kn_i})^T,$$

The corresponding group sparse coding framework could be written in a compact way as follows:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|A\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \gamma \|\boldsymbol{\theta}\|_g \quad (7)$$

where

$$\|\boldsymbol{\theta}\|_g = \sum_{k=1}^{c} \|\boldsymbol{\theta}_k\|_2 \quad (8)$$

denotes the group ℓ_1 norm throughout this paper and $\gamma > 0$ is the parameter. It is easy to see $\|.\|_g$ is well defined and essentially same as the penalty term of the group lasso model proposed in (Yuan and Lin 2006), except here we do not take the group size into account for simplicity. With the group ℓ_1 regularization term, the learned $\boldsymbol{\theta}$ explores the sparse structure of $\boldsymbol{\theta}$ as it also encodes the label information. Fig. (1) illustrates the framework. It can be expected that our proposed method would be more discriminative than the sparse representation framework if both used for image classification. We will give the detailed algorithm for solving Eq. (7) in the subsequent subsection.

Group Sparse Coding Algorithm

One of the most important contributions in this paper is our new algorithm to solve the group ℓ_1 problem. The Eq. (7) can be written explicitly in the following form:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|A\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \gamma \sum_{i=1}^{c} \|\boldsymbol{\theta}_i\|_2 \quad (9)$$

Taking the derivative with respect to $\boldsymbol{\theta}$ and setting to 0, we get the following equation:

$$A^T A\theta - A^T y + \gamma D\theta = 0 \quad (10)$$

where D is a block diagonal matrix and defined by

$$\begin{bmatrix} \frac{1}{2\|\boldsymbol{\theta}_1\|_2} I_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{2\|\boldsymbol{\theta}_c\|_2} I_{n_c} \end{bmatrix}$$

From this, we obtain

$$\boldsymbol{\theta} = (A^T A + \gamma D)^{-1} A^T \mathbf{y} \quad (11)$$

Because D depends on $\boldsymbol{\theta}$, the convergence of this algorithm is unclear at this point. We need to give an algorithm that is indeed convergent. We present an iterative algorithm to solve this problem. The detailed algorithm is given in Algorithm 1.

A few notes about this algorithm:

Step(A0): The Eq. (9) is a convex problem, therefore the initialization in (A0) is quite flexible given enough number of iterations, it is guaranteed that Algorithm 1 will converge to global optimum. In practice, we start with $\boldsymbol{\theta}$ the identity matrix.

Step(A2): The updating formula for $\boldsymbol{\theta}$ is a closed form solution and relatively simple, which makes the algorithm easy to implement.

Convergence criteria: we terminate the iteration when the value of the objective function $\frac{\|obj^{(t+1)} - obj^{(t)}\|_2}{\|obj^{(t)}\|_2} < 10^{-4}$ or the maximum number of iterations reached, which is 50. We want to point out that 50 is generally far more than the actual iterations needed to get the algorithm converged (at least true for our experiments in this paper). Indeed, we compared our

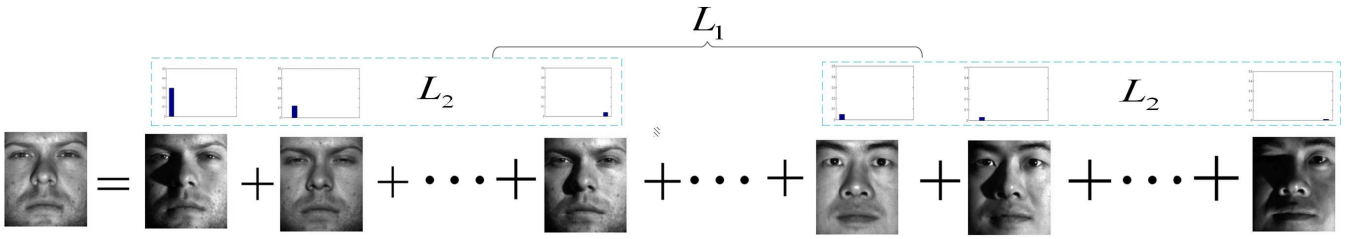

Figure 1: The illustration of our group sparse coding method. We use ℓ_2-norm for coefficients of images from the same subject, and impose ℓ_1-norm between different subjects. Thus, in group sparse coding process, the images from the same group prefer to have the large or small weights together, *i.e.* we expect the images from the correct class can dominant the sparse coding.

Algorithm 1: The Group Sparse Coding Algorithm

(A0): t=0. Initialize $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \cdots, \boldsymbol{\theta}_c^{(t)})^T$ **repeat**

(A1): Calculate the diagonal matrix

$$D^{(t)} = \begin{bmatrix} \frac{1}{2\|\boldsymbol{\theta}_1^{(t)}\|_2} I_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{2\|\boldsymbol{\theta}_c^{(t)}\|_2} I_{n_c} \end{bmatrix}$$

(A2): $\boldsymbol{\theta}^{(t+1)} = (A^T A + \gamma D^{(t)})^{-1} A^T \mathbf{y}$

(A3): t=t+1

until *Converge*

method with the gradient projection method in SLEP (Liu, Ji, and Ye 2009) to solve Eq. (9), for the data in this paper, it takes gradient projection method over 30,000 iterations to get the same converged objective function value that our method gets within 20 iterations and our method runs about 4 times faster.

Parameter Setting: overall the method is not sensitive to the parameter γ as long as within the reasonable threshold. For images with pixel value ranging from 0 to 255, $\gamma = 100$ is a good starting point for tuning.

There are several other ways to solve ℓ_1 problems, such as gradient projection (Figueiredo, Nowak, and Wright 2007; Kim et al. 2007), homotopy (Asif and Romberg 2009), iterative shrinkage-thresholding (Wright, Nowak, and Figueiredo 2008), proximal gradient (Beck and Teboulle 2009), iterative re-weighted (Nie et al. 2010; Cai et al. 2011), and augmented Lagrange multiplier (Bertsekas 2003). The comparison of these methods is beyond the scope of this paper.

Convergence Analysis

In step (A2) of the algorithm, assume $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ for notation simplicity in the proof and the updated $\boldsymbol{\theta}$ is $\tilde{\boldsymbol{\theta}}$.

Theorem 1. *The objective function value in each iteration will decrease.*

Proof. According to step (A2), we know that

$$\tilde{\theta} = \frac{1}{2} \min_{\theta} \|A\theta - y\|_2^2 + \gamma \theta^T D^{(t)} \theta$$

therefore

$$\frac{1}{2}\|A\tilde{\boldsymbol{\theta}} - \mathbf{y}\|_2^2 + \gamma \tilde{\boldsymbol{\theta}}^T D^{(t)} \tilde{\boldsymbol{\theta}} \leq \frac{1}{2}\|A\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \gamma \boldsymbol{\theta}^T D^{(t)} \boldsymbol{\theta} \quad (12)$$

Because we can easily prove

$$\|\tilde{\boldsymbol{\theta}}\|_2 - \frac{\|\tilde{\boldsymbol{\theta}}\|_2^2}{2\|\boldsymbol{\theta}\|_2} \leq \frac{\|\boldsymbol{\theta}\|_2}{2} = \|\boldsymbol{\theta}\|_2 - \frac{\|\boldsymbol{\theta}\|_2^2}{2\|\boldsymbol{\theta}\|_2}, \quad (13)$$

we have

$$\gamma \sum_{i=1}^{c} \|\tilde{\boldsymbol{\theta}}_i\|_2 - \gamma \sum_{i=1}^{c} \frac{\|\tilde{\boldsymbol{\theta}}_i\|_2^2}{2\|\boldsymbol{\theta}_i\|_2} \leq \gamma \sum_{i=1}^{c} \|\boldsymbol{\theta}_i\|_2 - \gamma \sum_{i=1}^{c} \frac{\|\boldsymbol{\theta}_i\|_2^2}{2\|\boldsymbol{\theta}_i\|_2}$$

which is equivalent to

$$\gamma \sum_{i=1}^{c} \|\tilde{\boldsymbol{\theta}}_i\|_2 - \gamma \tilde{\boldsymbol{\theta}}^T D^{(t)} \tilde{\boldsymbol{\theta}} \leq \gamma \sum_{i=1}^{c} \|\boldsymbol{\theta}_i\|_2 - \gamma \boldsymbol{\theta}^T D^{(t)} \boldsymbol{\theta} \quad (14)$$

Adding the inequalities (12) and (14) in both sides, we arrive at:

$$\frac{1}{2}\|A\tilde{\boldsymbol{\theta}} - \mathbf{y}\|_2^2 + \gamma \sum_{i=1}^{c} \|\tilde{\boldsymbol{\theta}}_i\|_2 \leq \frac{1}{2}\|A\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \gamma \sum_{i=1}^{c} \|\boldsymbol{\theta}_i\|_2 \quad (15)$$

since the objective function would decrease at each iteration or have converged according to our criteria and bounded below, the algorithm would converge. □

Supervised and Projected Sparse Coding

Although the group sparse coding method works well for the classification task, it still needs to be accelerated to overcome the efficiency issue as mentioned in the sparse representation. Given the high-dimensional data, one way is to apply classic dimension reduction methods such as PCA and LDA first, then apply the Group Sparse Coding to do the classification using projected data. However, such projected subspace is *unlikely* to be optimal for Group Sparse Coding method due to the decoupling of the two steps involved.

We propose to tightly integrate these two steps, the dimension reduction and classification, together in a unified framework. The critical part here is how to learn a proper projection matrix W. Here we propose a consistent way to get W, utilizing the class label information as discussed in preceding parts.

The objective function of our framework to learn W is the following

$$\min_{W,\boldsymbol{\beta}_i} \sum_{i=1}^{n} \tfrac{1}{2}\|W^T A_{-i}\boldsymbol{\beta}_i - W^T\mathbf{x}_i\|_2^2 + \gamma\|\boldsymbol{\beta}_i\|_g \quad s.t. \quad W^TW = I \tag{16}$$

where

$$\boldsymbol{\beta}_i = (\beta_1, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T \in \mathbb{R}^{n-1} \tag{17}$$

$$A_{-i} = (\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \cdots, \mathbf{x}_n) \in \mathbb{R}^{m\times(n-1)} \tag{18}$$

Here $\mathbf{x}_i$s are individual images from the training matrix A. The main motivation here is to simultaneously optimize the sparse vector representation for individual images and seek optimal projection matrix in terms of the representation residue sum, therefore find the optimal projected subspace for the subsequent classification. After we learned W, we can do the classification with Group Sparse Coding again in the embedded space.

Supervised Projected Sparse Coding Algorithm

Initialization The training of W has two stages. The first stage is to do the initialization of $\boldsymbol{\beta}_i^{(0)}$

$$\boldsymbol{\beta}_i^{(0)} = (\beta_1^{(0)}, \cdots, \beta_{i-1}^{(0)}, \beta_{i+1}^{(0)}, \cdots, \beta_n^{(0)})^T \in \mathbb{R}^{n-1} \tag{19}$$

from

$$\min_{\boldsymbol{\beta}_i^{(0)}} \frac{1}{2}\|A_{-i}\boldsymbol{\beta}_i^{(0)} - \mathbf{x}_i\|_2^2 + \gamma\|\boldsymbol{\beta}_i^{(0)}\|_g \tag{20}$$

for each training image $\mathbf{x_i}$, where A_{-i} defined in Eq.(18) represents the training matrix *without* i-th image, this can be solved using the algorithm given in Section .

After we learned $\boldsymbol{\beta}_{-i}^{(0)}$s, we can find the initial projection matrix $W^{(0)}$ according to the following objective function:

$$\min_{W^T} \sum_{i=1}^{n} \frac{1}{2}\|W^T A_{-i}\boldsymbol{\beta}_{-i}^{(0)} - W^T\mathbf{x}_i\|_2^2 \quad s.t. \quad W^TW = I \tag{21}$$

Here $W^{(0)} \in \mathbb{R}^{m\times k}$ and k is the pre-specified subspace dimension. The solution of $W^{(0)}$ is given by the eigenvectors corresponding to the k smallest eigenvalues of S, where $S = \sum_{i=1}^{n}(A_{-i}\boldsymbol{\beta}_i^{(0)} - \mathbf{x}_i)(A_{-i}\boldsymbol{\beta}_i^{(0)} - \mathbf{x}_i)^T$. So $W^{(0)}$ is the initial projection matrix to minimize S, the following section is about how to iteratively update W to find the optimal projection direction to minimize the sum of representation residuals for the training data.

Algorithm for Learning W After the initialization of W and $\boldsymbol{\beta}_i$s, we employ an iterative approach to solve Eq. (16). The first term in Eq. (16) is the sum of individual representation residuals in the subspace, the second term ensures the group sparsity of $\boldsymbol{\beta}_i$s. We solve Eq. (16) using an iterative approach that repeats the following two steps:

Step A. With W fixed, we get the updated β_is by solving Eq. (16). The objective function Eq. (16) becomes:

$$\min_{\boldsymbol{\beta}_{-i}} \sum_{i=1}^{n} \tfrac{1}{2}\|\tilde{A}_{-i}\boldsymbol{\beta}_i - \tilde{\mathbf{x}}_i\|_2^2 + \gamma\|\boldsymbol{\beta}_i\|_g \quad \tilde{A}_{-i} = W^TA_{-i} \quad \tilde{\mathbf{x}}_i = W^T\mathbf{x}_i \tag{22}$$

This involves n independent quadratic problems containing $\boldsymbol{\beta}_i$, which can be solved by Algorithm 1.

Step B. Solve W while fixing $\boldsymbol{\beta}_i$s. The objective function Eq. (16) becomes:

$$\min_{W} Tr\ W^TSW, \quad s.t. \quad W^TW = I \tag{23}$$

where $S = \sum_{i=1}^{n}(A_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i)(A_{-i}\boldsymbol{\beta}_i - \mathbf{x}_i)^T$

The solution of W is given by the eigenvectors corresponding to the k smallest eigenvalues of S.

Convergence. This iterative algorithm is guaranteed to converge because in both steps (A,B), the objective function is decreased. We set the convergence criteria of W as

$$\frac{\|W^{(t)}W^{(t)^T} - W^{(t-1)}W^{(t-1)^T}\|_1}{\|W^{(t-1)}W^{(t-1)^T}\|_1} < \varepsilon \tag{24}$$

where $W^{(t)}$ is the W at the t-th iteration and ε is a small tolerance which is usually set to $\varepsilon = 10^{-4}$.

Classification The class label prediction is given by the following equation.

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\left\|W^TA\boldsymbol{\theta} - W^T\mathbf{y}\right\|_2^2 + \gamma\left\|\boldsymbol{\theta}\right\|_g$$
$$\min_{k}\ r_k(\mathbf{y}) = \|W^T\mathbf{y} - W^TA_k\boldsymbol{\theta}_k\|_2 \tag{25}$$

Here $\mathbf{y}$ is the test image and $\boldsymbol{\theta}$ is the same value used for training for simplicity. It can be observed that classification equation (25) is consistent with Eq. (16), the W from Eq. (16) is well designed for the projected space classification.

Summary of Our Framework

There are a few main advantages of our method that we want to summarize here. First, different from sparse representation, our Group Sparse Coding method successfully incorporates the label information and therefore becomes a supervised method, in Algorithm 1 we further derive the solution to iterative update the representation vectors to find the global solution. Next, in our Projected Group Sparse Coding framework, with the unified objective function and simultaneous optimization strategy, the optimal projection matrix is iteratively updated. The experimental comparison between our algorithm and gradient projection method also demonstrates our method runs much faster.

Learning W using ℓ_1 Norm

There is another interesting discovery about the distinction of ℓ_1 norm and group ℓ_1 norm. If we replace the group ℓ_1 with the ℓ_1 norm throughout the training function, to be specific, replace Eq. (16) by the following equation:

$$\min_{W,\boldsymbol{\beta}_i} \sum_{i=1}^{n} \|W^TA_{-i}\boldsymbol{\beta}_i - W^T\mathbf{x}_i\|_2^2 + \lambda_2\|\boldsymbol{\beta}_i\|_1 \tag{26}$$

where $\lambda_2 > 0$ is the appropriate parameter. If using almost identical procedure except now involving ℓ_1 minimization

problems, we find the above plausible model fails due to low classification performance, the reason is that W would diverge from the supposed projection direction after enough long time of iterations in Eq.(26). This demonstrates the effectiveness of the label information and the group sparse regularization term from another point of view.

Experiments

Four benchmark datasets are used in our experiments, including extended YaleB (Georghiades, Belhumeur, and Kriegman 2001), subset of CMU PIE (contains only C05, C07, C09, C27, C29) (Sim, Baker, and Bsat 2002), USPS Handwritten Digit, and UCI Semeion Handwritten Digit (Buscema and MetaNet 1998). We use the cropped images from them, resize them to 16×16 pixels images to satisfy the over-complete assumption for sparse representation.

In this section, we want to demonstrate the classification performance for both Group Sparse Coding (GSC) and SPSC. We will also include supervised translation-invariant sparse coding (SSC) for classification comparison for non-embedding scenario, as SSC requires reasonable size batch and therefore quite difficult to apply to low dimension cases.

Unless otherwise specified, the experiment setting throughout this section is as follows. We randomly select n images from each class, equally split them into training set and test set and then do the two-fold cross validation on the same data instance for each method. Since each classification method has one or more parameters to be tuned, in order to compare these methods fairly, we run these methods under different parameters setting for each instance and the best one for each method has been recorded. The reported result is the average accuracy in percentages of 10 independent such experiments on each data set.

For k in kNN, it is tuned by searching the grid $\{1, 2, 3, \ldots, 10\}$. For our method Group Sparse Coding and SRC, the regularization parameters are tuned by searching the grid $\{1, 2, 2^2, \ldots, 2^{10}\}$ for images with pixel values from 0 to 255, adjusted accordingly for normalized images. For c in SVM linear model, we search from 10^{-8} and double it each time for 40 different values. We choose to use linear kernel for simplicity due to the subspace structure of face images and there is a justification on page 9 in (Wright et al. 2008). These parameters are set via searching for optimal setting. For SSC, we use the suggested value in (Yang, Yu, and Huang 2010).

Comparison to Different Classification Methods

In this part, we want to demonstrate the classification performance of the GSC method. We randomly sample 30 images each class from YaleB, 20 each class from PIE, 100 each class from USPS and 100 each class from Semeion, apply these classification methods on the same data instances and summarize the average accuracy results in Table 1. It can be observed that GSC outperforms other methods on all data sets except PIE. In particular, this indicates encoding the label information during the sparse coding can further improve the classification results at most cases.

Table 1: Classification Methods Accuracy Comparison

Methods	YaleB	PIE	USPS	Semeion
GSC	**91.9** ± 0.2	86.9 ± 0.1	**93.4** ± 0.1	**92.8** ± 0.1
SSC	91.7 ± 0.2	**87.3** ± 0.2	93.1 ± 0.1	92.6 ± 0.1
SRC	91.4 ± 0.2	85.0 ± 0.1	92.8 ± 0.1	92.6 ± 0.1
SVM	78.6 ± 0.3	60.2 ± 0.2	92.9 ± 0.2	91.8 ± 0.2
KNN	39.5 ± 0.4	29.3 ± 0.4	90.6 ± 0.3	85.8 ± 0.4

Classification in Subspace

In this part, we compare our Projected Group Sparse Coding method with different combinations of three dimension reduction methods and three classification methods, in total 9 different combinations. Dimension reduction methods are PCA, LDA, neighborhood component analysis (NCA) (Goldberger et al. 2004) and Laplacian preserving projection (LPP), classification methods are KNN, SVM (linear kernel) and SRC. It is well known that PCA and LPP belong to the unsupervised category, we include PCA here since it is very popular and a good benchmark, as to LPP, we implemented the supervised version of LPP in (Cai, He, and Han 2007), in their paper, Cai et al. integrate the label information into the projection matrix by searching the nearest neighbors of each training sample among the points sharing the same label. In the experiment section, when we mention LPP, we are referring to the supervised LPP and we set its parameters using the suggested values by the authors.

Classification with Varying Sample Size In this part, we experiment on the influence of different number of training samples for the classification performance. LDA method can reduce image size to a valid dimension up to number of class minus one. We set the projection dimension close to the number of class trying to be fair. In YaleB, we set the subspace dimension to 40; in PIE, we set it to 80; in USPS and Semeion, we set it to 20. We use SPSC to represent our method in the tables and figures thereafter. From Table 2, it can be observed that our method outperforms the other methods including those methods that use SRC classifier.

Classification with Partial Occluded Images

In this part, we want to demonstrate our method is robust to occlusions and compare the classification performance with other methods when the images are partially occluded. There are a few papers discussing the solutions to the image occlusions. Martinez *et al.* (Martinez 2002) choose to use blocks of features and analyze the local match with a probabilistic measure while Pentland *et al.* (Pentland, Moghaddam, and Starner 1994) used multiple eigenfaces to select fixed feature. However, when the location of the occlusion is unpredictable, these methods are less likely to succeed here. SRC has demonstrated its robustness dealing with image occlusions comparing to conventional methods. We want to show our method has even more advantages. Table 3 summarizes the classification methods' performance as occlusion size grows, the sample size n is 30 images each class in YaleB, 20 images each class in PIE, 100 images each class in USPS and Semeion, our method restricted the subspace di-

(a) YaleB

Methods/Samples	20	30	40
SPSC	$\mathbf{80.7 \pm 0.3}$	$\mathbf{83.0 \pm 0.4}$	$\mathbf{90.3 \pm 0.6}$
LDA+SRC	73.4 ± 1.7	82.5 ± 0.8	89.4 ± 0.5
LDA+KNN	72.3 ± 1.6	81.8 ± 0.9	86.8 ± 1.0
LDA+SVM	73.9 ± 1.0	82.1 ± 0.2	87.4 ± 0.7
LPP+SRC	77.9 ± 0.6	82.1 ± 0.9	85.1 ± 0.2
LPP+KNN	77.1 ± 1.1	82.5 ± 0.2	85.0 ± 1.4
LPP+SVM	77.1 ± 1.1	81.3 ± 0.4	84.6 ± 0.9
NCA+SRC	75.7 ± 1.2	81.5 ± 0.6	87.7 ± 0.3
NCA+KNN	75.3 ± 1.4	82.3 ± 0.4	85.8 ± 0.9
NCA+SVM	75.4 ± 1.2	81.8 ± 0.3	86.2 ± 0.8
PCA+SRC	73.8 ± 0.7	79.2 ± 0.7	84.6 ± 0.9
PCA+KNN	31.8 ± 0.2	38.2 ± 1.7	40.2 ± 1.3
PCA+SVM	73.9 ± 1.0	76.4 ± 0.8	82.3 ± 1.3

(b) PIE

Methods/Samples	20	30	40
SPSC	$\mathbf{79.4 \pm 0.1}$	$\mathbf{97.0 \pm 0.3}$	$\mathbf{98.3 \pm 0.3}$
LDA+SRC	78.2 ± 0.3	96.0 ± 0.3	97.4 ± 0.1
LDA+KNN	76.8 ± 2.4	94.6 ± 0.8	97.0 ± 0.2
LDA+SVM	77.1 ± 0.7	95.5 ± 0.9	96.8 ± 0.3
LPP+SRC	77.4 ± 0.1	90.5 ± 1.1	95.3 ± 0.7
LPP+KNN	76.9 ± 1.3	82.5 ± 0.8	95.6 ± 0.1
LPP+SVM	78.9 ± 0.9	93.8 ± 0.7	95.7 ± 0.4
NCA+SRC	77.2 ± 0.1	90.5 ± 1.0	95.4 ± 0.6
NCA+KNN	76.8 ± 2.2	93.8 ± 1.2	95.5 ± 0.3
NCA+SVM	78.3 ± 1.2	94.8 ± 0.9	96.4 ± 0.3
PCA+SRC	57.7 ± 1.3	90.5 ± 1.9	89.8 ± 0.9
PCA+KNN	29.8 ± 2.5	69.0 ± 2.3	72.1 ± 0.7
PCA+SVM	60.7 ± 2.8	81.0 ± 0.2	92.7 ± 0.2

(c) USPS

Methods/Samples	60	80	100
SPSC	$\mathbf{91.9 \pm 0.6}$	$\mathbf{92.5 \pm 0.5}$	$\mathbf{92.8 \pm 0.4}$
LDA+SRC	45.5 ± 2.2	62.7 ± 1.7	73.8 ± 0.3
LDA+KNN	47.3 ± 3.1	65.2 ± 2.0	74.0 ± 0.4
LDA+SVM	46.3 ± 2.4	65.5 ± 2.7	76.1 ± 2.1
LPP+SRC	89.3 ± 3.0	89.9 ± 0.1	90.1 ± 1.0
LPP+KNN	90.9 ± 1.1	89.6 ± 0.4	90.1 ± 0.6
LPP+SVM	89.8 ± 0.6	89.7 ± 0.5	89.7 ± 0.6
NCA+SRC	89.2 ± 2.9	89.9 ± 0.1	90.1 ± 0.9
NCA+KNN	90.9 ± 1.0	89.7 ± 0.4	90.2 ± 0.4
NCA+SVM	89.7 ± 0.7	89.8 ± 0.6	89.8 ± 0.5
PCA+SRC	86.3 ± 1.3	85.9 ± 0.3	85.9 ± 0.8
PCA+KNN	89.5 ± 0.3	89.4 ± 0.6	90.0 ± 0.6
PCA+SVM	90.2 ± 2.4	91.3 ± 0.4	91.5 ± 0.3

(d) Semeion

Methods/Samples	60	80	100
SPSC	$\mathbf{84.4 \pm 0.2}$	$\mathbf{86.4 \pm 0.3}$	$\mathbf{88.8 \pm 0.4}$
LDA+SRC	39.6 ± 1.1	61.8 ± 0.1	69.1 ± 1.5
LDA+KNN	42.7 ± 0.2	67.9 ± 0.7	76.2 ± 0.7
LDA+SVM	44.4 ± 1.1	69.1 ± 1.0	77.2 ± 0.4
LPP+SRC	74.5 ± 0.5	79.9 ± 1.0	82.4 ± 2.1
LPP+KNN	76.5 ± 1.0	80.1 ± 1.5	83.2 ± 1.1
LPP+SVM	78.8 ± 0.8	82.5 ± 0.5	84.5 ± 0.2
NCA+SRC	77.6 ± 0.4	82.3 ± 0.3	85.4 ± 1.9
NCA+KNN	79.2 ± 0.9	82.4 ± 1.1	86.2 ± 1.3
NCA+SVM	78.6 ± 0.9	82.8 ± 0.4	85.1 ± 0.2
PCA+SRC	83.1 ± 0.4	85.6 ± 0.3	86.9 ± 0.1
PCA+KNN	83.8 ± 2.5	85.8 ± 1.1	87.9 ± 1.1
PCA+SVM	83.8 ± 1.3	86.1 ± 0.6	88.3 ± 0.3

Table 2: Classification Performance on Different Data Sets

(a) YaleB

size	3	4	5	6	7
SPSC	**85.4**	**84.0**	**83.4**	**78.6**	**77.9**
SSC	85.1	82.8	81.3	78.3	77.6
SRC	82.4	80.8	79.9	78.2	77.4
SVM	68.1	58.3	58.9	54.7	48.7
KNN	27.8	19.7	15.4	16.1	15.1

(b) PIE

size	3	4	5	6	7
SPSC	77.1	75.2	**72.8**	**70.7**	**66.1**
SSC	**78.2**	**75.7**	72.6	69.4	65.7
SRC	76.8	74.3	72.4	68.9	64.9
SVM	42.7	37.1	32.3	28.5	25.8
KNN	20.5	14.3	12.1	11.8	10.8

(c) USPS

size	3	4	5	6	7
SPSC	**91.2**	**91.0**	**89.3**	**88.4**	**84.6**
SSC	90.6	88.5	87.3	82.7	78.4
SRC	90.6	88.6	87.2	82.8	78.3
SVM	89.7	90.1	87.0	86.5	82.4
KNN	89.3	87.0	85.3	80.3	77.0

(d) Semeion

size	3	4	5	6	7
SPSC	**88.2**	**88.0**	**87.2**	**86.1**	**85.2**
SSC	87.2	85.6	83.3	79.3	77.4
SRC	84.0	84.2	82.8	79.0	77.0
SVM	86.8	87.4	86.0	84.6	83.8
KNN	82.8	81.2	76.6	75.2	73.8

Table 3: Classification Performance on Partial Occluded Different Data Sets

mension to 120 in all four data sets. It can be observed that our method outperforms other methods in most cases. One thing to note in these tables is that due to the randomness in both the images selection and occlusion position, the average classification accuracy is not a monotone decreasing curve as the block size grows, nevertheless, it still can be concluded that our method has better performance overall in dealing with the random contiguous occlusion.

Conclusion

In this paper, we propose a new Supervised Group Sparse Coding framework based on the introduction of the label information during the learning of the representation vector, such that the conventional sparse representation framework is turned into a supervised method. We also proposed a novel algorithm to solve group ℓ_1 problem. What is more, we unify the dimension reduction and sparse vector learning into one objective function and do the simultaneous optimization. Experimental results on four benchmark datasets demonstrate that the proposed method outperforms related classification methods.

Acknowledgements

This work was partially funded by NSF CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965. The first draft of this paper was finished in 2011.

References

Asif, M., and Romberg, J. 2009. Dynamic updating for ℓ_1 minimization.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage thresholding algorithm for linear inverse problems. In *SIAM Journal of Imaging Sciences*, volume 2, 183–202.

Bertsekas, D. 2003. *Nonlinear programming*. Athena Scientific.

Buscema, M., and MetaNet. 1998. The theory of independent judges, in substance use and misuse.

Cai, X.; Nie, F.; Huang, H.; and Ding, C. 2011. Multi-class $l_{2,1}$-norm support vector machine. In *ICDM*, 91–100.

Cai, D.; He, X.; and Han, J. 2007. Spectral regression for efficient regularized subspace. In *ICCV*, 1–8.

Candes, E., and Tao, T. 2006. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory* 52(12):5406–5425.

Candes, E., and Wakin, M. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine* 25:21–30.

Chen, S.; Donoho, D.; and Saunders, M. 1998. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 20:33–61.

Donoho, D. 2004. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math* 59:797–829.

Figueiredo, M.; Nowak, R.; and Wright, S. 2007. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. 586–597.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 2001. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on PAMI* 23(6):643–660.

Goldberger, J.; Roweis, S.; Hinton, G.; and Salakhutdinov, R. 2004. Neighbourhood components analysis. In *NIPS*.

Kim, S.; Koh, K.; Lustig, M.; Boyd, S.; and Gorinevsky, D. 2007. An interior-point method for large-scale ℓ_1-regularized least squares. In *IEEE Journal of Selected Topics in Signal Processing*, volume 1, 606–617.

Liu, J.; Ji, S.; and Ye, J. 2009. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University.

Martinez, A. 2002. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *IEEE Transactions on PAMI* 24(6).

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *NIPS*.

Pentland, A.; Moghaddam, B.; and Starner, T. 1994. View-based and modular eigenspaces for face recognition. In *CVPR*, 84–91.

Sim, T.; Baker, S.; and Bsat, M. 2002. The cmu pose, illumination, and expression (pie) database.

Wright, J.; Yang, Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2008. Robust face recognition via sparse representation. 31(2):210–227.

Wright, S.; Nowak, R.; and Figueiredo, M. 2008. Sparse reconstruction by separable approximation. *ICASSP* 57(7):2479–2493.

Yang, J.; Yu, K.; and Huang, T. 2010. Supervised translation-invariant sparse coding. In *CVPR*, 3517–3524.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68:49–67.