Affine-Constrained Group Sparse Coding and Its Application to Image-Based Classifications

Yu-Tseh Chi[*] Mohsen Ali[*] Muhammad Rushdi[†] Jeffrey Ho

Department of Computer and Information Science and Engineering, University of Florida

[†]Department of Biomedical and Systems Engineering, Cairo University, Giza, Egypt

ychi@cise.ufl.edu moali@cise.ufl.edu mrushdi@eng.cu.edu.eg jho@cise.ufl.edu

Abstract

*This paper proposes a novel approach for sparse coding that further improves upon the sparse representation-based classification (*SRC*) framework. The proposed framework, Affine-Constrained Group Sparse Coding (**ACGSC**), extends the current* SRC *framework to classification problems with* **multiple** *input samples. Geometrically, the affine-constrained group sparse coding essentially searches for the vector in the convex hull spanned by the input vectors that can best be sparse coded using the given dictionary. The resulting objective function is still convex and can be efficiently optimized using iterative block-coordinate descent scheme that is guaranteed to converge. Furthermore, we provide a form of sparse recovery result that guarantees, at least theoretically, that the classification performance of the constrained group sparse coding should be at least as good as the group sparse coding. We have evaluated the proposed approach using three different recognition experiments that involve illumination variation of faces and textures, and face recognition under occlusions. Preliminary experiments have demonstrated the effectiveness of the proposed approach, and in particular, the results from the recognition/occlusion experiment are surprisingly accurate and robust.*

1. Introduction

Sparse representation-based classification (SRC) has been investigated in several notable recent work (e.g., [5, 26]), and despite the simplicity of the framework, the reported recognition results are quite impressive. The geometric motivation behind this approach is the assumption that data from each class resides on a low-dimensional linear subspace spanned by the training images belonging to the given class. The dictionary is obtained virtually without cost by simply stacking the training samples into a matrix $\mathbf{D}$. During testing, a test image $\mathbf{x}$ is sparse coded with respect to the dictionary by optimizing a (convex) objective function $\mathcal{E}(\mathbf{c};\mathbf{D})$ of the sparse coefficient $\mathbf{c}$ that is usually a sum of an ℓ_2-data fidelity term and a sparse-inducing regularizer:

$$\mathcal{E}(\mathbf{c};\mathbf{x},\mathbf{D}) = \|\mathbf{x}-\mathbf{D}\mathbf{c}\|^2 + \lambda\Psi(\mathbf{c}). \tag{1}$$

A plethora of sparse-inducing regularizers have been proposed in the literature (e.g., [25, 1, 20, 10]), and many of them are based on the sparse-promoting property of the ℓ_1-norm [3]. The block structure of the dictionary $\mathbf{D}$ together with the sparse coding of $\mathbf{x}$ allow one to infer the class label of $\mathbf{x}$ by examining the corresponding block components of $\mathbf{c}$, and SRC essentially looks for the sparsest representation of a test image with the hope that such a representation selects a few columns of $\mathbf{D}$ from the correct block (class). However, for many image classification problems in computer vision, the current SRC as embodied by Eq. (1) has several inherent shortcomings. The proposed affine-constrained group sparse coding (**ACGSC**) model aims to further improve the effectiveness of SRC by addressing two of these shortcomings: its generalizability and its extension for multiple input images.

In the information age, data are plentiful and in many applications, test data do not come in singles but in groups. For example, in video surveillance, a duration of merely one second would provide about 30 frames of images. Therefore, there is a need to properly generalize SRC for classification problems that require decision on a group of test data $\mathbf{x}_1, ..., \mathbf{x}_k$. On the other hand, for most computer vision applications, there does not exist a classifier that can correctly anticipate every possible variation of the test samples. For image-based face recognition in particular, these include variation in illumination, pose, expression and image alignment. In particular, subspace-based classification, for which SRC is a special case, is known to be sensitive to image misalignment [26]. Even for a small degree, misalignment can be detrimental and cause temperamental behavior of the classifier with unpredictable outcomes.

Multiple inputs, $[\mathbf{x}_1, ..., \mathbf{x}_k]$, provides a new and differ-

[*] These two authors have equal contribution to this paper.

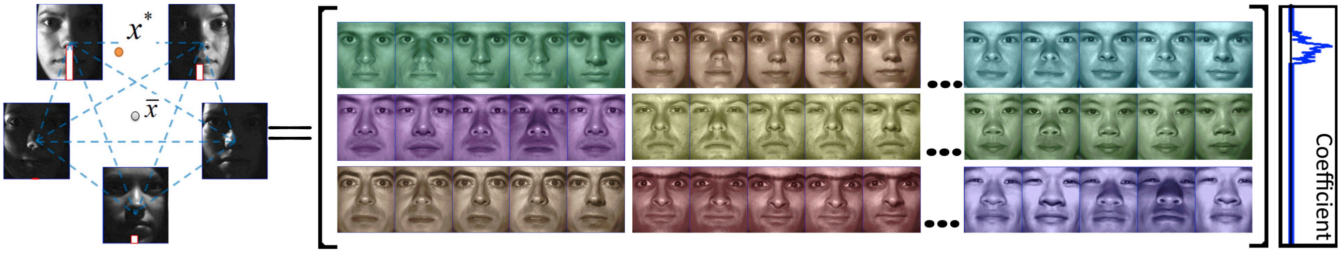


Figure 1: *Illustration of the proposed approach. Left: The convex hull formed by columns of test samples* $\mathbf{X}$*. The image corresponding to each column of* $\mathbf{X}$ *are shown and* $x^*(= \mathbf{X} \cdot \mathbf{a}^*)$ *is the solution of our proposed approach.* $\bar{x}$ *is the mean of* $\mathbf{X}$*. The magnitude of each component* $\mathbf{a}_i$ *of* $\mathbf{a}$ *is shown as the white bar in the corresponding image. Right: Illustration of the selected atoms from* $\mathbf{D}$ *and the sparse coefficients* $\mathbf{c}$ *of* x^* *shown on the right. Images of the same color are in the same block (class). (Figure best viewed in color)*

ent setting for SRC, and a straightforward application of SRC would be to apply SRC individually to each $\mathbf{x}_i$ and generate the final classification decision by pooling or voting among the individual classification results. This approach is unsatisfactory because by treating $\mathbf{x}_i$ independently, it completely ignores the (hidden) commonality shared by these obviously related data. Group sparse coding [1] (GSC) offers a solution that takes into account the potential relation among $[\mathbf{x}_1, ..., \mathbf{x}_k]$ by sparse coding all data simultaneously using the objective function

$$\min_{\mathbf{C}} \ \mathcal{E}(\mathbf{C}; \mathbf{X}, \mathbf{D}) = \|\mathbf{X} - \mathbf{DC}\|^2 + \lambda\Psi(\mathbf{C}), \qquad (2)$$

where $\mathbf{X}$ is a matrix formed by horizontally stacking together $\mathbf{x}_i$, and $\Psi(\mathbf{C})$ is an appropriate ℓ_1/ℓ_2-based regularizer. The matrix $\mathbf{C}$ of sparse coefficients can be used as in SRC to generate classification decision by applying voting or pooling across its rows. However, the effect of Ψ on the matrix $\mathbf{C}$ (and its pooling and subsequent classification) is difficult to predict and ascertain.

For classification problems in computer vision, this paper argues that a generalization of the group sparse coding, *affine-constrained group sparse coding*, using the following objective function, offers a more principled and flexible approach:

$$\mathcal{E}(\mathbf{a}, \mathbf{c}; \mathbf{X}, \mathbf{D}) = \|\mathbf{Xa} - \mathbf{Dc}\|^2 + \lambda\Psi(\mathbf{c}), \qquad (3)$$

where $\mathbf{a} = [\mathbf{a}_1 ..., \mathbf{a}_k]^\top$, is a k-dimensional vector with nonnegative components satisfying the affine constraint $\mathbf{a}_1 + ... + \mathbf{a}_k = 1$. Comparing with **GSC** as in Eq. (2), the **ACGSC** enlarges its feasible domain by including a k-dimensional vector $\mathbf{a}$. However, the feature vector $\mathbf{c}$ used in classification is in fact a vector not a matrix, and comparing with **GSC**, the classification decision based on Eq. (3) does not require pooling or voting.

Geometrically, **ACGSC** is easy to explain as it simply searches for the vector in the convex hull $\mathbf{S}$ generated by $\mathbf{x}_1, ...\mathbf{x}_k$ that can best be sparse coded using the dictionary $\mathbf{D}$. From classification viewpoint, **ACGSC** benefits from multiple inputs by using a linear generative model to "align" the test images with the dictionary elements. For example, if the test (face) images were taken under illumination conditions that are very different from the ones for the dictionary elements, sparse coding test images individually or in group can be expected to provide poor classification results. Heuristically, this can be explained as a kind of generalized misalignment (in terms of illumination effects) between the training and test images. For **ACGSC**, the coefficient $\mathbf{a}$ is used to "align" the test images with the dictionary elements, and it is precisely this online "alignment" of the input images $\mathbf{x}_i$ that provides **ACGSC** with an edge over other SRC methods. The necessity of the affine constraint can be reasoned in two ways. First, without it, Eq. 3 admits a trivial uninformative solution. Second, in most applications, the linear generative model given by $\mathbf{Xa}$ is valid only for restricted $\mathbf{a}$, and the nonnegative affine constraint proposed here is sufficiently general to provide a bounded and tractable domain for efficient optimization. We remark that for image-based applications, the test images $\mathbf{x}_i$ have nonnegative intensity values, and therefore, their convex hull would never contain the zero vector (i.e., $\mathbf{c}$ cannot be trivial). Likewise, for a typical group of data $\mathbf{x}_i$, their convex hull will not contain the zero vector. Finally, the framework embodied in Eq. 3 is sufficiently flexible to permit several interesting and useful variations, some of which will be discussed later in the paper.

The argument in favor of affine-constrained group sparse coding relies mainly on a form of sparse recovery guarantee presented in **Theorem 1** below. As will be made more precise later, this will allow us to argue that, at least in theory, the classification performance of **ACGSC** should be at least as good as the one based on **GSC** or on Eq. (1). We conclude the introduction by summarizing the three main contributions of this paper:

1. We propose a novel sparse representation-based classification framework based on affine-constrained group sparse coding, and it provides a **principled** extension of the current SRC framework to classification problems with **multiple input samples**. The resulting optimization problem can be shown to be convex and can be solved efficiently using iterative algorithms.
2. We provide theoretical analysis of the proposed frame-

work in the form of a sparse recovery result. Based on this result, we argue that in theory, the classification performance of the proposed framework is equal to or better than other existing SRC frameworks.

3. We evaluate the proposed framework using three classification experiments. The results suggest that the proposed framework does provide noticeable improvements over existing methods, particularly for difficult classification problems such as recognition with occlusions.

2. Theory and Method

Let $\mathbf{x}_1, \cdots, \mathbf{x}_k$ denote a group of input test data, and $\mathbf{D}$ is the given dictionary. We further let $\mathbf{X}$ denote the matrix $\mathbf{X} = [\mathbf{x}_1\, \mathbf{x}_2\, \cdots\, \mathbf{x}_k]$. Our proposed affine-constrained group sparse coding seeks to minimize the objective function:

$$\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c}; \mathbf{X}, \mathbf{D}) = \|\mathbf{X}\,\mathbf{a} - \mathbf{D}\,\mathbf{c}\|^2 + \lambda\, \Psi(\mathbf{c}) \qquad (4)$$

subject to the nonnegative affine constraint on the *group coefficient* $\mathbf{a}$ ($\sum_{i=1}^{k} \mathbf{a}_i = 1$, and $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_k \geq 0$). Note that in group sparse coding [1] (**GSC**), there are no group coefficients $\mathbf{a}$ and the sparse coefficients $\mathbf{c}$ are given as a matrix. A schematic illustration of the difference between the group sparse coding and our constrained version is shown in Fig. 2. The **GSC**-based classification scheme sparse codes the input feature vectors $\mathbf{x}_i$ simultaneously. While some group sparsity can be claimed for this approach based on the appropriate regularizer Ψ, it is generally difficult to provide any guarantee on the behavior of the sparse coefficient matrix $\mathbf{C}$. On the other hand, for our constrained version, the discrete set of the input vectors has been *completed* to form a convex set $\mathbf{S}$, and our approach is designed to capture any vector in this convex set that can best be sparse coded by the dictionary $\mathbf{D}$. The situation here shares some similarity with the LP-relaxation of integer programming [24] or the convexification of a non- convex program [18], in which one enlarges the feasible domain in order to achieve convexity and thereby, efficiently compute approximate solution.

We remark that the affine constraint is quite necessary in Eq. (4), and without it, there is always the trivial solution $\mathbf{a} = 0, \mathbf{c} = 0$. It is clear that the optimization problem is indeed convex and it is completely tractable as the feasible domain and objective function are both convex. We iteratively solve for $\mathbf{a}$ and $\mathbf{c}$ using gradient descent, and this scheme is guaranteed to converge. The only complication is the projection onto the simplex defined by the group coefficient constraint $\mathbf{a}_1 + ... + \mathbf{a}_k = 1$, and this step can be efficiently managed using an iterative scheme described in the supplemental material.

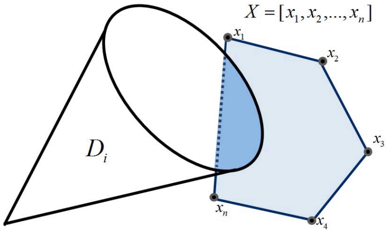


Figure 2: *Illustration of the difference between group sparse coding and constrained group sparse coding, and its effect on classification results. The cone represents the subspace spanned by a block of the dictionary* $\mathbf{D}$. *Shaded areas represent the convex hull spanned columns in* $\mathbf{X}$. *None of the* $\mathbf{x}_i$ *lie within the subspace; however some of the points on the convex hull do and the proposed algorithm is designed to capture these points.*

2.1. Theoretical Guarantee

Given a dictionary $\mathbf{D}$, a vector $\mathbf{x}$ has sparsity s if it can be written exactly as a linear combination of s columns of $\mathbf{D}$. An important result that underlies all SRC frameworks is the guarantee provided by the sparse recovery result that for a feature vector $\mathbf{x}$ with sparsity bounded from above by a constant depending on $\mathbf{D}$ [4, 6], $\mathbf{x}$ can be recovered by minimizing the ℓ^1 cost-function:

$$(\mathcal{P}_1) \quad \min_{\mathbf{c}} \|\mathbf{c}\|_1 \text{ subject to } \mathbf{D}\mathbf{c} = \mathbf{x}. \qquad (5)$$

In actual application, the above ℓ^1-program is often modified as

$$(\mathcal{P}_1^\lambda) \quad \min_{\mathbf{c}} \|\mathbf{x} - \mathbf{D}\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1, \qquad (6)$$

for a suitably chosen constant $\lambda > 0$. We remark that the two programs, while related, are in fact different, with most sparse recovery results given by minimizing $(\mathcal{P}_1)$. Let $\mathbf{x}$ be a noiseless test vector to be classified. A typical SRC method will determine its classification based on its sparse coefficients obtained by minimizing the program $(\mathcal{P}_1^\lambda)$. Compared to them, our proposed framework enlarges the optimization domain by introducing the group coefficients $\mathbf{a}$, and it is possible that with larger domain, spurious and incorrect solutions could arise. The following theorem rules out this possibility, at least when the sparse vector $\mathbf{x}$ can be **exactly** recovered by a typical SRC method and classified correctly:

Theorem 1. *Let* $\mathbf{x}$ *be a feature vector with sparsity* s *such that it can be exactly recovered by minimizing* $\mathcal{P}_1^\lambda$ *for some* λ. *We assume that* $\mathbf{x}$ *is in the convex hull* $\mathbf{S}$. *Furthermore, we assume that the global minimum of* $\mathcal{E}_{\mathbf{CGSC}}$, *given a group of input data* $\mathbf{X}$, *is unique. Then,* $\mathbf{x}$ *is the global minimum of* $\mathcal{E}_{\mathbf{CGSC}}$ *with the same* λ *and* $\mathbf{D}$ *(and* $\Psi(\mathbf{c}) = \|\mathbf{c}\|_1$*).*

Proof. The proof is straightforward and it consists of checking the sparse vector $\mathbf{x}$ also corresponds to the global minimum of $\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c})$. Since $\mathbf{x} \in \mathbf{S}$, we have

$\mathbf{x} = \mathbf{Xa}$ for some feasible $\mathbf{a}$. Since $\mathbf{x}$ is a sparse vector that can be recovered exactly by minimizing $(\mathcal{P}_1^\lambda)$ in Eq. (6), we let $\mathbf{c}$ be its sparse coefficients, and we have $\mathbf{x} = \mathbf{Dc}$. We claim that $(\mathbf{a}, \mathbf{c})$ is a global minimum of $\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{x})$ by showing that the (sub-)gradient vanishes at $(\mathbf{a}, \mathbf{c})$. First, since $\mathbf{c}$ is the global minimum for $(\mathcal{P}_1^\lambda)$ with $\mathbf{x} = \mathbf{Xa}$, and the two $\mathbf{c}$-subgradients coincide: $\nabla_{\mathbf{c}}(\mathcal{P}_1^\lambda) = \nabla_{\mathbf{c}}\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c})$. Therefore the $\mathbf{c}$-component $\nabla_{\mathbf{c}}$ of the (sub-)gradient $\nabla\mathcal{E}_{\mathbf{CGSC}}$ contains zero:

$$0 \in \nabla_{\mathbf{c}}\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c}).$$

On the other hand, since $\mathcal{E}_{\mathbf{CGSC}}$ is smooth in $\mathbf{a}$ and by direct calculation, we have $\mathbf{a}$-component of the gradient $\nabla\mathcal{E}_{\mathbf{CGSC}}$

$$\nabla_{\mathbf{a}}\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c}) = \mathbf{X}^\top\mathbf{Xa} - \mathbf{X}^\top\mathbf{Dc} = 0,$$

because $\mathbf{Dc} = \mathbf{x} = \mathbf{Xa}$. This shows that $(\mathbf{a}, \mathbf{c})$ is the global minimum of the convex function $\mathcal{E}_{\mathbf{CGSC}}(\mathbf{a}, \mathbf{c})$, regardless whether $\mathbf{x}$ is on the boundary of the convex hull $\mathbf{S}$. □

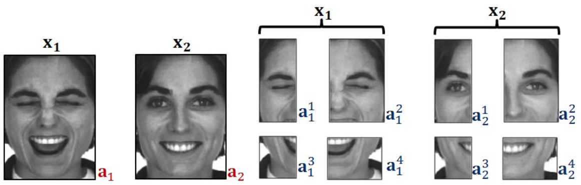
We can draw two important conclusions from the theorem. First, compared to **GSC**, our constrained version, with an enlarged feasible domain, will indeed recover the right solution if the (noiseless) solution is indeed among the input feature vectors $\mathbf{x}_1, ...\mathbf{x}_k$. Therefore, our method will not produce incorrect result in this case. However, the behavior of **GSC** in this case is difficult to predict because other (noisy) input feature vectors will affect the sparse coding of the (noiseless) input vector, and the result of the subsequence pooling based on the matrix $\mathbf{C}$ can be uncertain. Second, if there is a sparse vector $\mathbf{x}$ lying inside the convex hull $\mathbf{S}$ spanned by $\mathbf{x}_1, ...\mathbf{x}_k$, our method will indeed recover it (when the required conditions are satisfied).

2.2. Part-based ACGSC

The **ACGSC** framework based on Eq. (3) is versatile enough to allow for several variations, and here we discuss one example incorporating an image domain partition. In the standard **ACGSC**, the online reconstructed image $\mathbf{X}\,\mathbf{a}$ is simply the convex combination of the input images as the columns of $\mathbf{X}$. This model can be augmented using a known image domain partition and define the reconstructed image as a composite image that uses a different convex combination in each region of the partition. A schematic illustration of this idea is shown in Fig. 3, and this part-based **ACGSC** is particularly effective for detecting the presence of occlusions. More specifically, this requires the modification of Eq. (3)

$$\mathcal{E}(\mathbb{A}, \mathbf{c}; \mathbf{X}, \mathbf{D}) = \| \sum_i^k \mathbf{A}_i\mathbf{x}_i - \mathbf{Dc}\|^2 + \lambda\Psi(\mathbf{c}), \qquad (7)$$

where $\mathbb{A}$ is the set of all $\mathbf{A}_i$, $\mathbf{A}_i$ are diagonal matrices with nonnegative elements, k is the number of input samples, and $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th column in $\mathbf{X}$. The affine constraints on the $\mathbf{A}_i$ are $\sum_i^k \mathbf{A}_i^j = 1$ for $j = 1 \cdots d$, where $\mathbf{A}^j$ is the j-th diagonal element of $\mathbf{A}$. The resulting vector $\sum_i \mathbf{A}_i\mathbf{x}_i$ is the element-wise affine combination[1] of $\mathbf{x}_i$'s.



(a) (b)

Figure 3: *Comparison between the standard (a) and the part-based* ***ACGSC*** *(b).* ***(a)****: $\mathbf{a}_i$ are the group coefficients corresponding to the sample $\mathbf{x}_i$. The nonnegative affine constraint here is $\mathbf{a}_1 + \mathbf{a}_2 = 1$.* ***(b)****: The same input samples are split into 4 parts in the part-based approach. There are 4 nonnegative affine constraints i.e. $\mathbf{a}_1^p + \mathbf{a}_2^p = 1$ for $p = 1 \cdots 4$.*

Although Eq. (7) provides an extension of Eq. (3), it is severely under-constrained as there are $d \cdot k$ unknowns in all the $\mathbf{A}_i$. To alleviate this problem, we can further reduce the number of variables in $\mathbf{A}_i$. For example, the equation below gives only n_p different variables: $\mathbf{a}_i^j$ are scalar variables, and $\mathbf{I}_p$ are identity matrices of certain sizes,

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{a}_i^1 \cdot \mathbf{I}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_i^2 \cdot \mathbf{I}_2 & \mathbf{0} & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{a}_i^{n_p} \cdot \mathbf{I}_{n_p} \end{bmatrix}. \qquad (8)$$

This formulation of $\mathbf{A}_i$ is equivalent to splitting a sample $(\mathbf{x}_i)$ into n_p parts. Each part of $\mathbf{x}_i$ corresponds to a scalar variable $\mathbf{a}_i^p$. The size of a part is equal to the size of the corresponding $\mathbf{I}_p$. Note that each $\mathbf{I}_p$ does not necessarily have the same size. Let $\mathcal{I}_p$ denote the set of indices of the rows in $\mathbf{A}_i$ corresponding to $\mathbf{a}^p$. Eq. (7) can be rewritten as:

$$\| \begin{bmatrix} \mathbf{X}^{(\mathcal{I}_1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(\mathcal{I}_2)} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{X}^{(\mathcal{I}_{n_p})} \end{bmatrix} \begin{bmatrix} \mathbf{a}^1 \\ \mathbf{a}^2 \\ \vdots \\ \mathbf{a}^{n_p} \end{bmatrix} - \mathbf{Dc}\|^2 + \lambda\Psi(\mathbf{c})$$

$$\text{s. t.} \quad \sum_{i=1}^k \mathbf{a}_i^p = 1 \ \text{ for } p = 1 \cdots n_p \text{ and } \ \mathbf{a}_i^p \geq 0, \qquad (9)$$

where $\mathbf{a}^p = [\mathbf{a}_1^p, \mathbf{a}_2^p, \cdots, \mathbf{a}_k^p]^\intercal$ and $\mathbf{X}^{(\mathcal{I}_p)}$ are the rows in $\mathbf{X}$ that correspond to the p-th part. Because the first part of the data fidelity term is still a d-dimensional vector, optimization of $\mathbf{c}$ is the same as in Eq. (3). Although Eq. (9) and Eq. (3) have a similar structure, the former has part-structure defined on the components of $\mathbf{a}$, and in practice, the parts are specified by each individual application. For

[1] $\sum_i \mathbf{A}_i\mathbf{x}_i = \sum_i \mathsf{diag}(\mathbf{A}_i) \odot \mathbf{x}_i$, where $\odot$ denotes element-wise product.

face recognition, we can define the parts according to the image regions where the useful features such as eyes, nose and mouth are to be found.

Since the first matrix in Eq. (9) is block-diagonal, Eq. (9) can be rewritten as:

$$\sum_{p=1}^{n_p} \|\mathbf{X}^{(\mathcal{I}_p)}\mathbf{a}^p - \mathbf{D}^{(\mathcal{I}_p)}\mathbf{c}\|^2 + \lambda\Psi(\mathbf{c}), \tag{10}$$

where $\mathbf{D}^{(\mathcal{I}_p)}$ are the rows of $\mathbf{D}$ corresponding to rows of $\mathbf{X}^{(\mathcal{I}_p)}$. The vector $\mathbf{a}^p$ can then be optimized individually under the nonnegative affine constraints given in Eq. (9). Note that the indices corresponding to $\mathbf{a}^p$ in $\mathbf{A}$, as shown in Eq. (9), do not have to be contiguous. This provides us more flexibility for determining and specifying useful parts, depending on the intended application.

3. Related Work

To the best of our knowledge, a similar framework and algorithm to the one proposed in this paper has not been reported in the computer vision literature. Due to limited space, we will only summarize the major differences between our work and some of the more representative work in SRC that have appeared in the past few years. Sparse representations have been successfully applied to many vision problems such as face recognition [12], [17], [22], image classification [23], [5], [7], denoising and inpainting [14], and other areas [25]. Interestingly, one of the original motivations for using sparse representations in solving vision problems is the realization that sparse representations of visual signals are produced in the first stage of visual processing by human brain [15]. In many cases, simply replacing the original features with their sparse representations leads to surprisingly better recognition/classification results [25].

Group sparse coding was first proposed in [1] and its extension to block-structured dictionaries has been studied in [2, 19]. In these methods, group sparsity was promoted using a matrix norm that encourages features in the same group to share the same atoms in the dictionary. From classification viewpoint, [1] has two undesirable features: First, the classification is based on matrices (multiple vectors) and this increase in dimension complicates the process. Second, while promoting group sparsity, [1] does not go beyond the test data $\mathbf{x}_i$ themselves to search for potentially more useful features that are better-aligned with the dictionary. Our proposed framework addresses both shortcomings by introducing the group coefficients $\mathbf{a}$ in Eq. (3).

4. Experiments

4.1. Face Recognition with Lighting Variation

SRC-based face recognition methods have been extensively studied and the state-of-the-art results have been reported in the past few years [26, 5]. However, none of the cited work has investigated face recognition under more realistic scenario when there are large differences between the illumination conditions of the training and test images. In this experiment, we used the *cropped Extended Yale Face Database B* [11] to simulate such scenario. This database contains aligned images of 38 persons under different laboratory-controlled illumination conditions. For each subject, we chose the images with the azimuth and elevation angles of the light source $\leq 35^o$ as the training images. The rest of the images in the database, which contain significant shadows due to non-frontal illumination, were used as test images. The training images were used to simulate the well-lit images such as passport photos, and the test images were used to simulate images (e.g. from surveillance camera) that are quite different from the training image. Fig. 4 demonstrates the large differences between the training (top row) and test (bottom row) images. Unlike as in the earlier work, this experiment uses ”easier” images for training and ”harder” images for testing, and the result will demonstrate unequivocally the superior generalization capacity of our method, an important feature for classification and recognition applications.

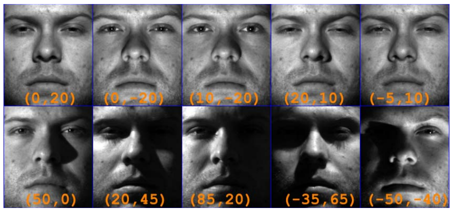


Figure 4: *Selected training (top) and test (bottom) images from Yale Face Database. Numbers in the parenthesis are, respectively, the azimuth and elevation angles of the illumination source.*

We used the training images directly as atoms of the dictionary $\mathbf{D}$ [5, 26]. Therefore, there are 38 blocks in $\mathbf{D}$, and each block contains 24 or 23 atoms[2]. The number of test images for each person is around 40. The experiment was performed as follows:

1. Reduce dimensionality of the data to 600 using PCA. Normalize the samples to have unit ℓ_2 norms.
2. Use the training images directly as atoms in $\mathbf{D}$.
3. For each subject, randomly select $n_g \in \{2, 3, \cdots, 7\}$ number of test images ($\mathbf{X}$).
4. Initialize $\mathbf{a} = [\frac{1}{n_g}, \cdots, \frac{1}{n_g}]^\intercal$ and $\mathbf{c} = \mathbf{0}$.
5. Iteratively update $\mathbf{a}$ and $\mathbf{c}$ until convergence. Determine the class label by
$$\text{label}(\mathbf{X}) = \min_i \|\mathbf{X}\mathbf{a} - \mathbf{D}_i\mathbf{c}_i\|_2, \tag{11}$$
where $\mathbf{D}_i$ and $\mathbf{c}_i$ are the i-th block of $\mathbf{D}$ and $\mathbf{c}$, respectively.

[2] There are missing images for several individuals in the database.

We repeated the above experiment ten times and the results are plotted in Fig. 5. We have compared our result with the result of simply using the mean of columns of $\mathbf{X}$ as input vector (last column of Fig. 6). We have also compared with two group-regularized sparse coding algorithms proposed by [1] [2]. In [1], the energy function is (same as Eq. (2))

$$\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{DC}\|_{\mathsf{F}}^2 + \lambda \sum_{i}^{|D|} \|\mathbf{C}_i\|_2, \tag{12}$$

where $\mathbf{C}_i$ is the i-th row of $\mathbf{C}$. Their algorithm promotes the data in a group (columns of $\mathbf{X}$) to share same dictionary atoms. In [2], block structure was added to $\mathbf{D}$ together with the group structure on the data $\mathbf{X}$. The objective function, which is still in the same form as Eq. (2), is:

$$\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{DC}\|_{\mathsf{F}}^2 + \lambda \sum_{b}^{n_b} \|\mathbf{C}_b\|_{\mathsf{F}}, \tag{13}$$

where $\mathbf{C}_b$ is a matrix containing the rows in $\mathbf{C}$ that corresponds to the b-th block of $\mathbf{D}$. Similar to the previous algorithm, the method proposed in [2] promotes the encoding of data in $\mathbf{X}$ using atoms from a few blocks of $\mathbf{D}$.

We have also compared with the results using methods proposed by Wright et. al.[26] and by Elhamifar et. al.[5]. We applied these two methods to every test sample since they do not utilize group structure on data and the results are also shown in Fig. 5[3]. The results show that the proposed method significantly outperforms all competing methods. The clear reason, as shown in Fig. 2, is that our method goes beyond the input test images and searches for a more "aligned" image on a larger domain (the convex hull spanned by $\mathbf{X}$) while the group-regularized methods can only rely on the input images that are poorly poorly "aligned" with the dictionary.

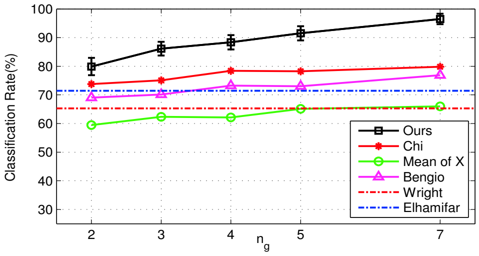


Figure 5: *Comparison of our method with other* SRC*-based methods. The λ values used in the methods, in the order of the legend, are 0.05, 0.2, 0.05, 0.1, 0.05, and 0.05, respectively.*

Fig. 6 demonstrates four groups of test samples, the actual computed coefficients $\mathbf{a}$ (white bars), the image at optimality, $\mathbf{Xa}^*$, (2nd column from the right) and the mean image (last column). The images at optimality have the lighting conditions that are more similar to the atoms in $\mathbf{D}$ (top row in Fig. 4) than the mean images *if the convex hull spanned by columns of* $\mathbf{X}$ *lies within the subspaces spanned by the blocks in* $\mathbf{D}$. The first three rows demonstrate successful examples by our methods. We were not able to classify the 1st and 3rd correctly using the mean of $\mathbf{X}$ because the overwhelming amount of shadow. The bottom row shows a failed case where none of the images contains any identifiable or distinguishable feature.

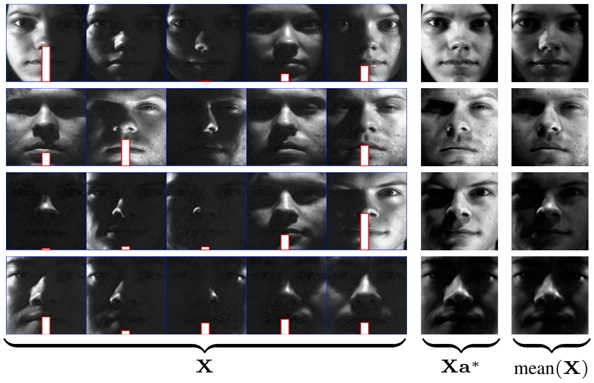


Figure 6: *Left: columns of* $\mathbf{X}$ *and the values of the computed* $\mathbf{a}$ *(white bars). The last two columns are the results of* $\mathbf{X} \cdot \mathbf{a}$ *and the mean of columns of* $\mathbf{X}$*, respectively. The first three rows*

4.2. Face Recognition with Occlusions

We have tested our proposed approach using the AR-Face database[16]. This dataset contains face images of 100 individuals. There are 14 non-occluded and 12 occluded images from each individual with different expressions and illumination variations. The occluded images contains two types of occlusions: sun-glasses and scarf covering the face from the nose down. Each occlusion type contains 6 images per person. To reduce the feature dimension, we down-sampled the images to 55 × 40 and vectorized them. We randomly selected 8 non-occluded images from each person to form the dictionary $\mathbf{D}$ with a 100-block structure. The experiment was performed as follows:

1. Randomly selected n_g test samples ($\mathbf{X}$) from the occluded images of person p. They must contain at least one from each type of occlusion.
2. Split the test images into 6 uniformly-sized and non-overlapping parts (Fig. 8(d)).
3. Initialize $\mathbf{A}_i = \mathbf{I}/n_g$ ($\mathbf{a}^p = \mathbf{1}/n_g$), $\mathbf{c} = \mathbf{0}$.
4. Iteratively optimize $\mathbf{c}$ and $\mathbf{A}_i$ ($\mathbf{a}^p$'s) using Eq. (9) and Eq. (10), respectively.
5. Determine the class label by

$$\text{label}(\mathbf{X}) = \min_i \sum_{p=1}^{n_p} \|\mathbf{X}^{(\mathcal{I}_p)}\mathbf{a}^p - \mathbf{D}_i^{(\mathcal{I}_p)}\mathbf{c}_i\|^2, \tag{14}$$

where $\mathbf{D}_i$ and $\mathbf{c}_i$ are the i-th block of $\mathbf{D}$ and $\mathbf{c}$, respectively. This equation is a modification of Eq. (11).

[3] The results for these two methods shown in Fig. 5 are worse than what were reported in the original papers ([26] and [5]). This is because, in their experiments, they randomly chose half of the dataset as atoms of $\mathbf{D}$ and the other half as test samples. Therefore their dictionaries are 50% larger than ours and the variability between training and test samples are much more limited due to the random selection of training images.

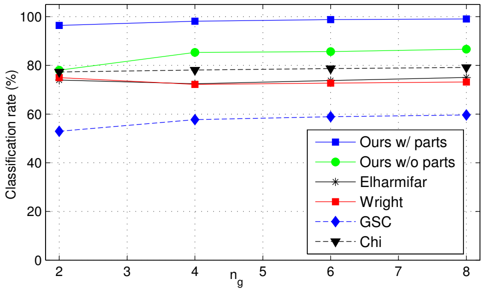


Figure 7: *Comparison of classification results. The λ values used in the methods, in the order of legends, are* $0.005, 0.005, 0.004, 0.002, 0.01,$ *and* 0.02*, respectively.*

We repeated the above procedure 20 times for each person . We compared the results with that of using standard **ACGSC**. We also compared our results with those of **GSC** [1] and of Chi et. al. [2]. Lastly, we compared with the results from directly applying regular sparse coding (Wright et. al. [26]) and block sparse coding (Elharmifar et. al. [5]) on the average image of the test samples (Fig. 8(b)). The results in Fig. 7 show that our part-wise **ACGSC** outperforms other methods by a significant margin. The standard **ACGSC** does not have significant advantage over other methods. This is due to the fact that the occlusions are present in all the test samples.

Fig. 8(d) shows the part-based group coefficients ($\mathbf{a}^p$) of the test samples after the optimization. The values of $\mathbf{a}^p$ are displayed using the colors overlaid on the corresponding parts. The largest coefficient of this specific example is around 0.6. We can see that the parts corresponding to the occluded regions have significantly lower values, i.e., our method correctly identified the occlusions. Fig. 8(e) shows the part-based affine combination of the test images. Our part-based approach was able to select the parts that are more consistent with the dictionary (training samples). Fig. 8(f) shows the reconstructed image using our method and Fig. 8(c) shows the reconstructed image by directly applying Wright's method [26] on the average image (Fig. 8(b)). Due to the occlusion of the scarf, the test samples were incorrectly matched to training images from a male subject with beard and mustache.

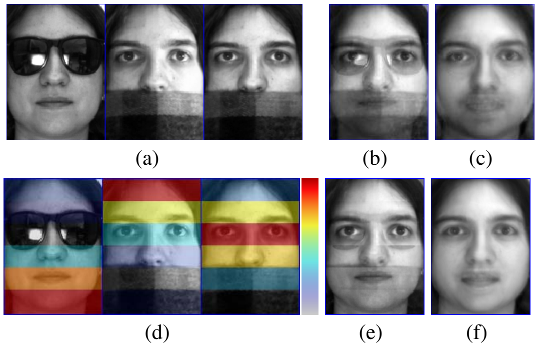


Figure 8: ***(a)*** *Input test samples.* ***(b)*** *The average of the 3 test samples.* ***(c)*** *Reconstructed image using Wright's method on the average image (b).* ***(d)*** *Weights of the part-wise group coefficients overlaid on the corresponding test samples. More redness corresponds to a larger affine weight.* ***(e)*** *The part-wise affine combination of the test samples after optimization.* ***(f)*** *Reconstructed image using our method. (Images best viewed in color)*

4.3. Texture Classification

In this experiment, we used the cropped CUReT texture [21]. This dataset contains images of 61 materials (See Fig. 9). It has a total of $61 \times 92 = 5612$ images. Each image is of size 200-by-200 pixels. For each texture category, we randomly chose 20 images as the training samples and the rest as test samples. The experiment was conducted as follows:

1. For each image, compute its SIFT [4] features over the entire image. Each image is then represented by one 128-dimensional SIFT vector.
2. Normalize the SIFT vectors to have unit ℓ_2 norm. Vector components are capped at 0.25 as any component with values > 0.25 is now set to 0.25. Normalize the vectors again.
3. Use the training samples as columns of the dictionary $\mathbf{D}$. $\mathbf{D}$ contains 1,220 columns with 61 blocks.
4. For each class, randomly choose n_g test samples ($\mathbf{X}$).
5. Iteratively update $\mathbf{c}$ and $\mathbf{a}$ in Eq. (4). λ for computing $\mathbf{c}$ is set to 0.2 through cross validation.
6. Determine the class label of $\mathbf{X}$ using Eq. (11).

We have compared our results with ones using the framework proposed by Wright et. al. [26]. Since there is no group structure defined in their framework, we computed the sparse coefficients for all the test images individually. λ for this method is set to 0.17. We have also compared our results with those from Hayman et. al. [9], Gau et. al. [8], Varma et. al. [21] and Liu et. al.[13]. The classification results are shown in Fig. 10. All results are surprisingly good for using such simple features. The result from using Wright's framework is comparable to the state-of-the-art results, and our method further improves on the result of Wright et. al. by 3.3% (classification rate is 99.67% when $n_g = 5$).

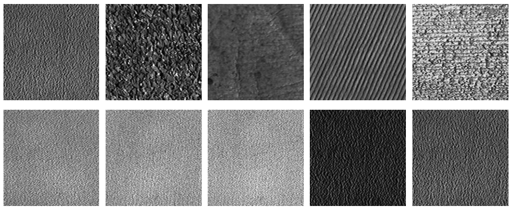


Figure 9: *Selected images from the the cropped CUReT database. Top row: 5 different types of textures. Bottom row: same texture under different illumination conditions.*

[4] We used the vl_sift package.

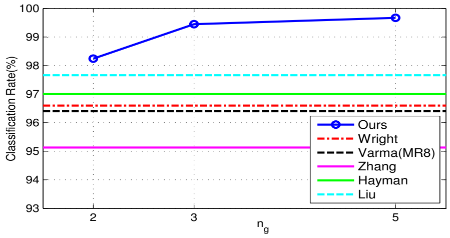


Figure 10: *Classification results from the texture classification experiment. Note that in Liu's work, 46 training images were used.*

5. Conclusions and Future Work

We have presented the novel Affine-Constrained Group Sparse Coding framework for SRC with the aim of extending the current SRC-framework to classification problems with multiple inputs. We have also presented a form of sparse recovery result and based on this result, we have argued that, at least in theory, the classification performance using the proposed method should be as good as if not better than the one using existing SRC-based methods. We have evaluated the proposed approach using three experiments that involves face recognition, texture classification and face recognition under occlusions. The preliminary experiments demonstrate the effectiveness as well as efficiency of the proposed approach. For future work, we will investigate more theoretical aspects of the approach. We believe that it is possible to obtain a stronger form of the sparse recovery result under noisy assumption, providing a better understanding of the power and limitation of the proposed algorithm. Furthermore, we will also investigate useful and effective prior for the group coefficients $\mathbf{a}$ and the resulting (usually non-convex) optimization problem.

Acknowledgement This work was partially supported by NSF Grant IIS 0916001.

References

[1] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. *Advances in NIPS*, 22:82–89, 2009.

[2] Y.-T. Chi, M. Ali, A. Rajwade, and J. Ho. Block and group regularized sparse modeling for dictionary learning. In *CVPR, 2013 IEEE Conference on*, pages 377–382, 2013.

[3] D. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[4] M. Elad. *Sparse and Redundant Representations.* Springer Verlag, 2010.

[5] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *CVPR, 2011 IEEE Conference on*, pages 1873–1879. IEEE, 2011.

[6] E. Elhamifar and R. Vidal. Block-sparse recovery via convex optimization. *Signal Process., IEEE Trans. on*, PP(99):1, 2012.

[7] S. Gao, L. Chia, and I. Tsang. Multi-layer group sparse coding for concurrent image classification and annotation. In *CVPR, 2011 IEEE Conference on*, pages 2809–2816. IEEE, 2011.

[8] Z. Guo, D. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *Image Process., IEEE Trans on*, 19(6):1657–1663, June.

[9] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh. On the significance of real-world conditions for material classification. In *ECCV (4)*, pages 253–266, 2004.

[10] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *AISTATS*, 2010.

[11] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[12] X. Li, T. Jia, and H. Zhang. Expression-insensitive 3d face recognition using sparse representation. In *CVPR 2009. IEEE Conference on*, pages 2575–2582, 2009.

[13] L. Liu and P. Fieguth. Texture classification from random features. *PAMI, IEEE Trans. on*, 34(3):574–586, March.

[14] J. Mairal. Learning multiscale sparse representations for image and video restoration (preprint). Technical report, DTIC Document, 2007.

[15] C. MarcAurelio Ranzato, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. *Advances in NIPS*, 19:1137–1144, 2006.

[16] A. Martinez and R. Benavente. The ar face database cvc technical report, no. 24. *Barcelona, Spain: Computer Vision Center, Universitat Autonoma de Barcelona*, 1998.

[17] P. Nagesh and B. Li. A compressive sensing approach for expression-invariant face recognition. In *CVPR 2009. IEEE Conference on*, pages 1518–1525. Ieee, 2009.

[18] R. T. Rockafellar. *Convex Analysis (Princeton Landmarks in Mathematics and Physics).* Princeton University Press, Dec. 1996.

[19] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. Eldar. C-hilasso: A collaborative hierarchical sparse modeling. *Signal Process., IEEE Trans. on*, 59(9):4183–4198, 2011.

[20] Z. Szabó, B. Póczos, and A. Lőrincz. Collaborative filtering via group-structured dictionary learning. *Latent Variable Analysis and Signal Separation*, pages 247–254, 2012.

[21] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1):61–81, 2005.

[22] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination. pages 597–604, 2009.

[23] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR, 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[24] L. Wolsey. *Integer Programming*. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, 1998.

[25] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI, IEEE Trans. on*, 31(2):210–227, 2009.