Multiple Pattern Classification by Sparse Subspace Decomposition

Tomoya Sakai
Institute of Media and Information Technology, Chiba University
1-33 Yayoi, Inage, Chiba, Japan
tsakai@faculty.chiba-u.jp

Abstract

A robust classification method is developed on the basis of sparse subspace decomposition. This method tries to decompose a mixture of subspaces of unlabeled data (queries) into class subspaces as few as possible. Each query is classified into the class whose subspace significantly contributes to the decomposed subspace. Multiple queries from different classes can be simultaneously classified into their respective classes. A practical greedy algorithm of the sparse subspace decomposition is designed for the classification. The present method achieves high recognition rate and robust performance exploiting joint sparsity.

1. Introduction

Classification is a task of assigning one or more class labels to unlabeled data (query data). A collection of labeled data (training data) is available for the classification. The patterns or signals to be classified are usually groups of measurement data expressed as high-dimensional vectors.

Depending on purposes, we need pattern classifiers that can answer

- a label to each of queries,
- a label to a set of queries,
- a few labels to each of queries,
- a label "invalid" to an unclassifiable query.

We develop a framework of using subspaces for all these functionalities. We regard the unlabeled data as a mixture of subspaces. The key idea is to decompose it into the subspaces of classes as few as possible. Only the classes explaining concisely the mixture are relevant to the unlabeled data. In the classification, the unlabeled data are usually supposed to belong to a few (typically one) classes. Therefore, the classification process can be interpreted as sparse decomposition of the subspace mixture.

This work is inspired by the recently developing field of compressed sensing [1, 2, 3, 4, 5] and its innovative applications to robust face recognition [6], action recognition [7], computer vision and image processing [8]. The essential idea of these works is to exploit the prior knowledge that a signal is sparse and compressible. The theory of compressed sensing is very helpful and informative for us to answer questions such as "How many measurements are enough for the pattern recognition?" and "What is the role of feature extraction?" It is worthy to explore the potential of sparse decomposition for substantial improvement of the subspace methods.

The rest of this paper is organized as follows. Section 2 provides preliminary details and definitions of subspace representation for sparse decomposition. In Section 3, we propose a classification method named *sparse subspace method*, which exploits the sparseness property for the classification tasks described above. A practical algorithm of the sparse subspace decomposition is presented in Section 4. We show some tentative evaluation results of the sparse subspace method using a face database in Section 5 before concluding in Section 6.

2. Preliminaries

Let $\mathbf{S}_k \in \mathbb{R}^{d \times n_k}$ be a matrix of training dataset of k-th class ($k = 1, \dots, C$), in which n_k labeled patterns are represented as the d-dimensional column feature vectors. We describe as follows the linear subspaces, their union, block sparsity, and sparse linear representation of a subspace. We also define a classification space where the sparsity should be encouraged.

Linear subspaces of training datasets The class subspace is defined as a vector subspace whose elements are the feature vectors of labeled data. We describe the subspace as a vector subspace in the normed space:

$$\mathcal{S}_k := \operatorname{span} \mathbf{S}_k \subset (\mathbb{R}^d, l^2). \tag{1}$$

S_k approximates the k-th class subspace. We denote the dimensionality of $\mathcal{S}_k$ by $\dim \mathcal{S}_k = \operatorname{rank} \mathbf{S}_k$.

Union of subspaces The union of subspaces is the subspace obtained by combining the feature vectors of each class.

$$\mathcal{S} := \cup_{k=1}^{C} \mathcal{S}_k = \operatorname{span} \mathbf{S} \subseteq (\mathbb{R}^d, l^2) \tag{2}$$

Here, $\mathbf{S}$ is the concatenation of $\mathbf{S}_k$ as

$$\mathbf{S} := [\mathbf{S}_1, \ldots, \mathbf{S}_C] \in \mathbb{R}^{d \times N} \tag{3}$$

and $N := \sum_{k=1}^{C} n_k$. The dimensionality of $\mathcal{S}$ is denoted by $\dim \mathcal{S} = \operatorname{rank} \mathbf{S}$.

We say that the subspaces $\mathcal{S}_k$ ($k = 1, \ldots, C$) are independent if and only if any subspace $\mathcal{S}_k$ is not a subset of the union of the other subspaces, *i.e.*, $\mathcal{S}_k \not\subset \cup_{i \neq k}^{C} \mathcal{S}_i$ for $\forall k$.

Linear representation of vector(s) Given sufficient training dataset, a d-dimensional vector $\mathbf{q}$ of unlabeled data (hereafter "query" vector) will be approximately represented as a linear combination of vectors from class subspaces.

$$\mathbf{q} = \sum_{k=1}^{C} \mathbf{S}_k \boldsymbol{\alpha}_k = \mathbf{S} \boldsymbol{\alpha} \tag{4}$$

Here, $\boldsymbol{\alpha}_k \in (\mathbb{R}^{n_k}, l^2)$ is a vector of coefficients corresponding to the k-th class, and

$$\boldsymbol{\alpha} := \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_C \end{bmatrix} \in (\mathbb{R}^N, l^2) \tag{5}$$

is the concatenation of $\boldsymbol{\alpha}_k$.

If a set of queries is given as a matrix

$$\mathbf{Q} := [\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}] \in \mathbb{R}^{d \times n}, \tag{6}$$

then we will solve

$$\mathbf{Q} = \mathbf{S} \mathbf{A}. \tag{7}$$

Here,

$$\mathbf{A} := [\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(n)}] \in \mathbb{R}^{N \times n} \tag{8}$$

is the matrix of unknown coefficients, and

$$\boldsymbol{\alpha}^{(j)} := \begin{bmatrix} \boldsymbol{\alpha}_1^{(j)} \\ \vdots \\ \boldsymbol{\alpha}_C^{(j)} \end{bmatrix} \in \mathbb{R}^N \tag{9}$$

is the concatenated vector of coefficients for the j-th query. The matrix $\mathbf{A}$ can also be described as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_C \end{bmatrix} \tag{10}$$

where

$$\mathbf{A}_k := [\boldsymbol{\alpha}_k^{(1)}, \ldots, \boldsymbol{\alpha}_k^{(n)}] \in \mathbb{R}^{n_k \times n}. \tag{11}$$

The systems of linear equations as (7) is called the problem for multiple measurement vectors (MMV), while the case of a single measurement $n = 1$ as (4) is referred to as SMV [9, 10, 11]. The query vectors correspond to the measurements in this context.

Uniqueness The solution $\boldsymbol{\alpha}$ to (4) or $\mathbf{A}$ to (7) exists if and only if

$$\mathbf{q}^{(j)} \in \mathcal{S} \ \ \forall j, \tag{12}$$

i.e., the queries lie on the union of class subspaces. For $\dim \mathcal{S} < d$, the solution does not always exist. The solution may be dense even if it exists. Most components are nonzero despite the fact that at most n class subspaces are relevant to n queries. This problem is due to invalid situation where training datasets are insufficient to identify the class, uniquely.

The actual problem we should cope with is the underdetermined case $d = \dim \mathcal{S} < N$, *i.e.*, the dimensionality of the union of subspaces is less than the total number N of training samples. Unless the training data matrices $\mathbf{S}_k$ are rank-degenerated so that $\dim \mathcal{S} < d$, the C subspaces of training data cannot be independent in the d-dimensional space. There is an infinite number of ways to express the query vector by the linear combination of the subspace bases. The underdetermined problem requires regularization to select a unique solution. A sparse solution indicating relevant classes would be preferable.

Block sparsity A vector $\boldsymbol{\xi} \in (\mathbb{R}^N, l^0)$ is called m-sparse if $||\boldsymbol{\xi}||_0 \leq m$. Here, $|| \cdot ||_0$ denotes the l^0 norm, which counts the nonzero vector components. As the support of a function is the subset of its domain where it is nonzero, the support of a vector $\boldsymbol{\xi}$ is defined as $\mathcal{T} = \{i | \xi_i \neq 0\}$. The l^0 norm is the cardinality of the support.

We define a block-wise sparsity level in a similar manner to [11]. Let $f_{\mathcal{N}}$ be a map from $\forall \mathbf{X} \in (\mathbb{R}^{N \times n}, l^F)$ to $\boldsymbol{\gamma} \in (\mathbb{R}_+^C, l^0)$ according to a list $\mathcal{N} := \{n_1, \ldots, n_C\}$ such that

$$f_{\mathcal{N}} : \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_C \end{bmatrix} \rightarrow \begin{bmatrix} ||\mathbf{X}_1||_F \\ \vdots \\ ||\mathbf{X}_C||_F \end{bmatrix} := \boldsymbol{\gamma}. \tag{13}$$

Here, $\mathbf{X}_k \in (\mathbb{R}^{n_k \times n}, l^F)$ is the k-th row block of $\mathbf{X}$ with respect to $\mathcal{N}$, and $|| \cdot ||_F$ denotes the Frobenius norm l^F. Clearly,

$$f_{\mathcal{N}} : \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_C \end{bmatrix} \rightarrow \begin{bmatrix} ||\mathbf{x}_1||_2 \\ \vdots \\ ||\mathbf{x}_C||_2 \end{bmatrix} \tag{14}$$

for $n = 1$. A vector $\mathbf{x} \in (\mathbb{R}^N, l^2)$ is called block M-sparse over $\mathcal{N}$ if $\mathbf{x}_k \neq \mathbf{0}$ for at most M indices k. The block sparsity is measured as

$$||\mathbf{x}||_{0,\mathcal{N}} := ||f_{\mathcal{N}}(\mathbf{x})||_0. \tag{15}$$

That is, $||\cdot||_{0,\mathcal{N}}$ counts the number of nonzero blocks. We measure the row block sparsity of a matrix $\mathbf{X} \in (\mathbb{R}^{N\times n}, l^F)$ over $\mathcal{N}$ as

$$||\mathbf{X}||_{0,\mathcal{N}} := ||f_{\mathcal{N}}(\mathbf{X})||_0. \tag{16}$$

A matrix $\mathbf{X}$ is row block M-sparse if $||\mathbf{X}||_{0,\mathcal{N}} \le M$.

We remark that the row block M-sparse matrix $\mathbf{X}$ can be converted into a block M-sparse vector vec$(\mathbf{X}^\top)$. Here, the operator vec transforms a matrix into a column vector by stacking all the columns of the matrix. For $\mathcal{N} := \{n_1, \ldots, n_C\}$ and $\mathcal{N}' := \{nn_1, \ldots, nn_C\}$, the block sparsity of $\mathbf{X} \in \mathbb{R}^{N\times n}$ over $\mathcal{N}$ is preserved as

$$||\mathbf{X}||_{0,\mathcal{N}} = ||\operatorname{vec}(\mathbf{X}^\top)||_{0,\mathcal{N}'}. \tag{17}$$

Sparse representation of subspace In the underdetermined case, the columns of matrix $\mathbf{S} \in \mathbb{R}^{d\times N}$ represent an overcomplete basis of $\mathbb{R}^d$ for $d < N$. Equation (4) and (7) can be consistent with infinitely many solutions $\boldsymbol{\alpha}$ and $\mathbf{A}$, respectively.

We denote the subspace of query vector(s) by $\mathcal{Q} = \operatorname{span}\mathbf{q}$ or $\operatorname{span}\mathbf{Q}$. If a possible solution $\boldsymbol{\alpha}$ or $\mathbf{A}$ is block sparse over $\mathcal{N} = \{n_1, \ldots, n_C\}$, the query subspace $\mathcal{Q}$ consists of a small minority of class subspaces corresponding to nonzero $\boldsymbol{\alpha}_k$ or $\mathbf{A}_k$. In other words, the query subspace is sparsely represented by the class subspaces. The sparsity of the subspace representation can be quantified as $||\boldsymbol{\alpha}||_{0,\mathcal{N}}$ or $||\mathbf{A}||_{0,\mathcal{N}}$.

Classification space By definition, the block sparsity $||\boldsymbol{\alpha}||_{0,\mathcal{N}}$ or $||\mathbf{A}||_{0,\mathcal{N}}$ is measured by the l^0 norm of the C-dimensional vector $\boldsymbol{\gamma} := f_{\mathcal{N}}(\boldsymbol{\alpha})$ or $f_{\mathcal{N}}(\mathbf{A})$. The components of $\boldsymbol{\gamma}$ imply the degrees of class membership. The sparser $\boldsymbol{\gamma}$ is, the more certainly the class label of each query is identified. The sparsity is properly measured by the l^0 norm. Therefore, we refer to the normed space $\mathcal{C} = (\mathbb{R}^C_+, l^0)$, where $\boldsymbol{\gamma}$ resides, as the classification space.

3. Classification based on sparse subspace representation

From the viewpoint of classification, each query vector is supposed to be composed only of vectors from the subspace of a class to which the query is classified. The subspace spanned by the query vectors should be represented as sparsely as possible by the class subspaces concerned with the queries. In our notation, the C-dimensional vector in the classification space, $\boldsymbol{\gamma} := f_{\mathcal{N}}(\boldsymbol{\alpha})$ or $f_{\mathcal{N}}(\mathbf{A})$, is intended to be sparsest. The sparsity is properly measured by the l^0 norm of $\boldsymbol{\gamma}$. Therefore, we incorporate minimization of the l^0 norm in the classification framework.

3.1. Formulation

Let $\mathbf{S} \in \mathbb{R}^{d\times N}$ be the concatenation of $\mathbf{S}_k \in \mathbb{R}^{d\times n_k}$ ($k = 1, \ldots, C$, $d = \operatorname{rank}\mathbf{S} < N = \sum_{k=1}^C n_k$), *i.e.*, the matrices of training datasets. Given the matrix $\mathbf{Q} \in \mathbb{R}^{d\times n}$ of n query vectors, we solve the l^0-minimization problem:

$$\min_{\mathbf{A}} ||\mathbf{A}||_{0,\mathcal{N}} \quad \text{subject to} \quad \mathbf{Q} = \mathbf{SA}. \tag{18}$$

Here, $\mathcal{N}$ specifies the sizes of row blocks for sparsification. Typically, $\mathcal{N} = \{n_1, \ldots, n_C\}$. The matrix $\mathbf{A}$ is released from being row-block sparse if $\mathcal{N} = \mathcal{N}_1 := \{\forall n_i = 1, i = 1, \ldots, N\} = \{1, \ldots, 1\}$.

One can rewrite the problem (18) as

$$\min_{\mathbf{A}} ||\operatorname{vec}(\mathbf{A}^\top)||_{0,\mathcal{N}'} \quad \text{subject to}$$
$$\operatorname{vec}(\mathbf{Q}^\top) = (\mathbf{S}\otimes\mathbf{I}_n)\operatorname{vec}(\mathbf{A}^\top) \tag{19}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{I}_n$ is the identity matrix of size n. The list $\mathcal{N}'$ defines the block sizes of the nN-dimensional vector vec$(\mathbf{A}^\top)$.

The l^0-minimization problem (19) is well investigated in the literature [11]. The uniqueness of the solution is guaranteed under the condition called block restricted isometry property (block RIP). Assuming $\mathbf{q}^{(j)} \in \mathcal{S}$, the RIP condition for our problem can be described as

$$\begin{aligned}(1-\delta_{M|\mathcal{N}'})||\mathbf{v}||_2^2 &\\ \le ||(\mathbf{S}\otimes\mathbf{I}_n)\,\mathbf{v}||_2^2 &\\ \le (1+\delta_{M|\mathcal{N}'})||\mathbf{v}||_2^2 &\quad \forall\mathbf{v}\in\mathbb{R}^{nN}.\end{aligned} \tag{20}$$

where $\delta_{M|\mathcal{N}'}$ is called the block-RIP constant dependent on the block sparsity M over $\mathcal{N}'$. In practice, we normalize the blocks $\mathbf{S}_k$ in order for the matrix $\mathbf{S}\otimes\mathbf{I}_n$ to satisfy the condition. The block RIP condition is less stringent than the standard RIP condition, which is widely used in the field of compressed sensing [1, 2, 3, 4, 5].

3.2. Dimensionality reduction

In (18), we assume the linear system $\mathbf{Q} = \mathbf{SA}$ to be underdetermined as $d = \operatorname{rank}\mathbf{S} < N$, and regularize it by the l^0 minimization. Actually, we do not have to deal with the queries and training data in a space of dimension $d \ge N$. The recent works in the emerging area of compressed sensing show that a small number of projections of a sparse vector can contain its salient information enough to recover the vector with regularization that promotes sparsity [1, 3, 12]. The statements in [13, 14] guaranteeing the recovery are described as follows.

Theorem 1 *Let* $\mathbf{x} := \boldsymbol{\Psi}^\top\mathbf{s}$ *be a* d*-dimensional vector represented by a* m*-sparse vector* $\mathbf{s}\in\mathbb{R}^d$ *using a basis* $\boldsymbol{\Psi}^\top \in \mathbb{R}^{d\times d}$*. Then,* $\mathbf{s}$ *can be reconstructed from a* $\hat{d}$*-dimensional vector* $\hat{\mathbf{x}} := \boldsymbol{\Phi}\mathbf{x}$ *with probability* $1 - e^{-\mathcal{O}(\hat{d})}$*. Here,* $\boldsymbol{\Phi} \in \mathbb{R}^{\hat{d}\times d}$ *is a random matrix and* $\hat{d} \ge \hat{d}_0 := \mathcal{O}(m\log(d/m))$*.*

Specially, $\hat{d} \geq 2m\log(d/\hat{d})$ holds if $m \ll d$ [15, 16]. It is also possible to recover the sparse vector $\mathbf{s}$ from a small number of projections, $\hat{\mathbf{x}}$, with overwhelming probability in more general case where $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are incoherent [15, 17, 5].

The reconstructability in Theorem 1 suggests that one can obtain the d-dimensional m-sparse solution from a much lower $\hat{d}$-dimensional vector after linear transformation. Wright *et al*. [6] showed, in their framework of face recognition based on sparse representation, that the computational cost is reduced without significant loss of recognition rate by linear transformations into lower dimensional feature spaces, such as Eigenfaces, Fisherfaces, Laplacianfaces, downsampling, and random projection. These transformations act as dimensionality reduction that preserves information for the recognition. Especially, random projection is a data-independent dimensionality reduction technique, and one can exactly recover the original d-dimensional vector. For this reason, we employ the dimensionality reduction if d is too high for computation.

3.3. Classifiers

n-to-one classifier Since the minimizer $\mathbf{A}$ for (18) is a row block M-sparse matrix, the M blocks indicate the M_C ($M_C \leq M$) classes concerned with the query subspaces. For the task of classifying all n queries into one class ($M_C = 1$), we calculate the residuals r_k of the representations by the class subspaces.

$$r_k(\mathbf{Q};\mathbf{A}) := ||\mathbf{Q} - \mathbf{S}_k\mathbf{A}_k||_F. \tag{21}$$

The residuals quantify the dissimilarities between the query subspace and the class subspaces. Note that most of the residuals are $||\mathbf{Q}||_F$ because of the sparsity. If the query subspace $\mathcal{Q}$ can be approximately represented by one of the class subspaces, the class label is identified as

$$\arg\min_k r_k(\mathbf{Q};\mathbf{A}). \tag{22}$$

This classification method achieves the same task as the mutual subspace methods [18, 19, 20] in a fundamentally different strategy. The mutual subspace methods are robust owing to the multiple queries. The robustness is further enhanced by the block sparsification in our scheme. The l^0 minimization in (18) encourages the vector of class membership degrees, $f_{\mathcal{N}}(\mathbf{A})$, to be as sparse as possible in the classification space. For the underdetermined problem with a sparse solution, the recent works in the emerging area of compressed sensing [1, 2, 3, 4] prove the exact recovery under the l^0 or l^1 regularization. Since the l^0 / l^1 minimizer is very insensitive to outliers, the sparse representation is robust compared to the conventional representations by l^2-based regularization *e.g*. PCA.

We also remark that if $n = 1$ and $\mathcal{N} = \mathcal{N}_1$, our n-to-one classification is exactly the same as the sparse representation-based classification (SRC) proposed in [6]. Our classification based on sparse subspace representation is therefore an extension of the SRC for multiple queries.

n-to-ones classifier It is also possible to classify n queries into their respective classes. We calculate $C \times n$ residual matrix whose kj-th entry measures the dissimilarity between the j-th query and its reconstruction in the k-th subspace:

$$r_k^{(j)}(\mathbf{Q};\mathbf{A}) := ||\mathbf{q}^{(j)} - \mathbf{S}_k\boldsymbol{\alpha}_k^{(j)}||_2. \tag{23}$$

Note that most of the residual entries are $||\mathbf{q}^{(j)}||_2$ because of the sparsity. If the query subspace $\mathcal{Q}$ can be approximately represented by union of a small number of class subspaces, the class label for the j-th query is identified as

$$\arg\min_k r_k^{(j)}(\mathbf{Q};\mathbf{A}). \tag{24}$$

Again, our method is expected to be robust owing to the multiple queries. Furthermore, the classes irrelevant to the queries are strongly excluded by the l^0 minimization. Therefore, the classifier (24) can detect the respective class for each query without giving the number of relevant classes.

n-to-M classifier Let us mention the potential of the sparse subspace representation for finding n-to-M relations, although we do not go into the detail of this type of multiple classification in this paper. If a query simultaneously belongs to multiple classes, the query vector is represented as a linear combination of vectors from the subspaces of the relevant classes. The residuals $r_k^{(j)}$ for such query cannot be zero, but the relevant classes are found by thresholding $r_k^{(j)}$. Thus, each of n queries is assigned to some of M classes.

Classification validity A classifier should answer "invalid" if the given query belongs to an unknown class. As suggested in [6], such an unclassifiable query is perceived to be so by measuring how the nonzero components of $\mathbf{A}$ concentrate on a single class. Wright *et al*. defined the sparsity concentration index (SCI), which quantifies the validity of the classification [6]. One may compute the SCI for each column of $\mathbf{A}$ to validate the corresponding query.

3.4. Sparse subspace method

Our classification method based on the sparse subspace representation is summarized in Algorithm 1.

Algorithm 1 Sparse subspace method (SSM)

```
Input: $\mathbf{Q} \in \mathbb{R}^{d\times n}$: matrix of $n$ queries as (6), $\mathbf{S} \in \mathbb{R}^{d\times N}$: concatenated matrix of training datasets as (3), $\mathcal{N}$: list of row block sizes;
Output: $\mathcal{L}$: set of class labels;
1 perform dimensionality reduction of $\mathbf{Q}$ and $\mathbf{S}$ if $d$ is intractably high;
2 normalize the columns of $\mathbf{S}$ to have unit $l^2$ norm;
3 decompose $\mathbf{Q}$ with respect to $\mathbf{S}$ to obtain the sparse subspace representation.
4 find the class label $\mathcal{L} = \{\arg\min_k r_k(\mathbf{Q};\mathbf{A})\}$ or
  $\mathcal{L} = \{\arg\min_k r_k^{(1)}(\mathbf{Q};\mathbf{A}),\dots,\arg\min_k r_k^{(n)}(\mathbf{Q};\mathbf{A})\}$.
```

The major concern is the sparse subspace decomposition of $\mathcal{Q}$ at Step 3. In the next section, we present a practical algorithm of the decomposition, SSD-ROMP, which efficiently and stably provides approximate solution to (18).

4. Sparse subspace decomposition

The sparse decomposition of $\mathbf{Q}$ in (18) is considered as a MMV problem whose solution is row-block sparse. The solution has two important characteristics: the column vectors $\boldsymbol{\alpha}^{(j)}$ of $\mathbf{A}$ share nonzero blocks as their support, and the block partitions are fixed by $\mathcal{N}$ in advance.

4.1. Prior work on MMV

Configuration of the nonzero entries in the solution $\mathbf{A}$ is called the joint sparsity model (JSM) [21, 22]. There are some prior works on the MMV problems with several JSMs [9, 10, 21, 22, 23, 11, 24]. Most of them [9, 10, 21, 22, 25, 23] focus on a JSM in which the column vectors $\boldsymbol{\alpha}^{(j)}$ simply share their support $\mathcal{T}$. This JSM is the special case of our row-block sparsity model with $\mathcal{N} = \mathcal{N}_1$ described in Section 3.1. Efficient algorithms for the MMV problem with this JSM have been designed as the extensions of greedy algorithms such as matching pursuit (MP) and orthogonal matching pursuit (OMP) [26, 27, 28, 29, 30]. OMP is an efficient algorithm that can recover a m-sparse vector from a $\mathcal{O}(m\log N)$-dimensional vector [30]. It iteratively selects the basis (column of $\mathbf{A}$) with the largest contribution to the current residual to reduce greedily the representation error at each iteration. The existing MP- and OMP-based algorithms for the MMV problem can be directly used for our problem with the row-block sparsity model only when $\mathcal{N} = \mathcal{N}_1$.

Eldar and Mishali [11] introduced the block sparsity model and block RIP condition applicable to MMV problems including ours. The uniqueness was guaranteed in 3.1. By l^1 convex relaxation, we can cast the vectorized version in (19) as

$$\min_{\mathbf{A}} ||\operatorname{vec}(\mathbf{A}^\top)||_{1,\mathcal{N}'} \quad \text{subject to}$$
$$\operatorname{vec}(\mathbf{Q}^\top) = (\mathbf{S}\otimes\mathbf{I}_n)\operatorname{vec}(\mathbf{A}^\top). \quad (25)$$

Here, we redefine $f_{\mathcal{N}}$ as a map from $(\mathbb{R}^{N\times n}, l^F)$ to $(\mathbb{R}_+^C, l^1)$ in the same form as (13), and define [1]

$$||\mathbf{A}||_{1,\mathcal{N}} := ||f_{\mathcal{N}}(\mathbf{X})||_1. \quad (26)$$

According to [11], this l^1 minimization problem is a second order cone problem (SOCP).

4.2. Sub-optimal algorithm

Algorithm 2 Sparse subspace decomposition (SSD-ROMP)

```
Input: $\mathbf{Q} \in \mathbb{R}^{d\times n}$: matrix of $n$ queries as (6), $\mathbf{S} \in \mathbb{R}^{d\times N}$: concatenated matrix of training datasets as (3), $\mathcal{N}$: list of row block sizes, $M_0$: sparsity level;
Output: $\mathbf{A}$: row-block sparse matrix as (10), $\mathcal{I}$: set of indices of nonzero blocks;
let the index set $\mathcal{I} := \emptyset$ and residual $\mathbf{R} := \mathbf{Q}$;
repeat
    $\mathbf{U} := \mathbf{S}^\top\mathbf{R}$;
    $\boldsymbol{\gamma} := f_{\mathcal{N}}(\mathbf{U})$;
    let $\mathcal{J}$ be a set of indices of the $M_0$ biggest components of $\boldsymbol{\gamma}$, or all of its nonzero components, whichever set is smaller;
    sort $\mathcal{J}$ in descending order of the components $\boldsymbol{\gamma}$;
    among all subsets $\mathcal{J}_0 \subset \mathcal{J}$ such that $\gamma_i \leq 2\gamma_j$ for all $i < j \in \mathcal{J}_0$, choose $\mathcal{J}_0$ with the maximal energy $||\gamma|_{\mathcal{J}_0}||_2^2 := \sum_{k\in\mathcal{J}_0}\gamma_k^2$;
    $\mathcal{I} := \mathcal{I} \cup \mathcal{J}_0$;
    for each $j$ do
        $\boldsymbol{\alpha}^{(j)} := \arg\min_{\boldsymbol{\alpha}} ||\mathbf{q}^{(j)} - \sum_{k\in\mathcal{I}}\mathbf{S}_k\boldsymbol{\alpha}||_2$;
    end for
    $\mathbf{R} := \mathbf{Q} - \sum_{k\in\mathcal{I}}\mathbf{S}_k\mathbf{A}_k$;
until $||\mathbf{R}||_F = 0$ or $\operatorname{card}\mathcal{I} \geq 2M_0$.
```

We present a practical greedy algorithm of the block sparse decomposition. Although there are optimization packages that solve the SOCP in polynomial time, we prefer a simple and efficient algorithm of the sparse recovery like the MP and OMP. As compared with the signal recovery in compressed sensing, approximate solutions may be enough for the classification purpose. Since the sparsity level is at most $\mathcal{O}(n)$ for n queries, we want the decomposition algorithm to work efficiently in the case of extreme sparseness.

[1] The norm $||\cdot||_{2,\mathcal{I}}$ defined in [11] is the same as our $||\cdot||_{1,\mathcal{N}}$, and it is actually the l^1 norm through $f_{\mathcal{N}}$ as we defined.

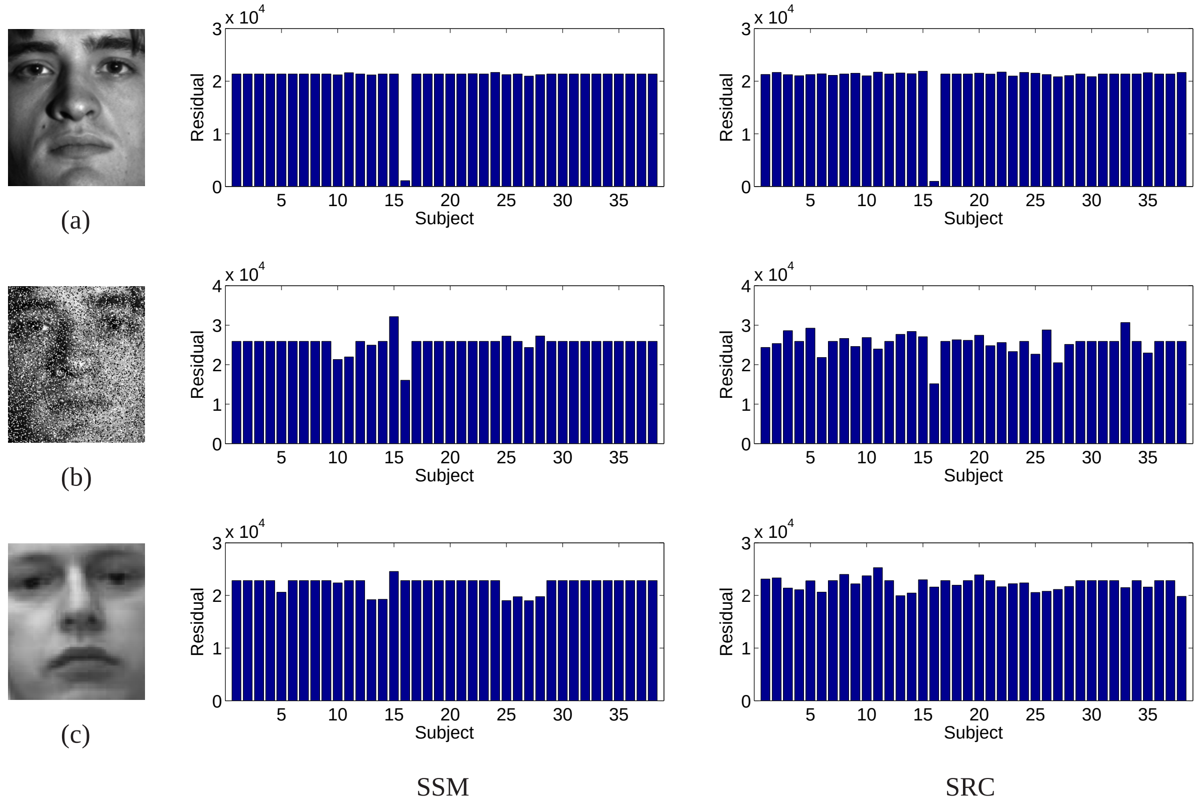

Figure 1. Examples of one-to-one classification. The first, second, and third columns respectively show the query images, residuals of the representations by SSM with respect to 38 sujects, and those by SRC. (a) A valid query image of subject #16. (b) The same query image as (a) with 30% pixels corrupted by salt and pepper noise. (c) An invalid image from unlearned face database.

We adopt the regularized OMP (ROMP) [31] because it can stably provide approximate solution from noisy queries. We modify the ROMP to seek for the nonzero row blocks of the solution as shown in Algorithm 2. This algorithm selects multiple row-blocks of $\mathbf{S}^\top \mathbf{S}\mathbf{A}$ that have comparable magnitudes measured by $f_\mathcal{N}$ at each iteration. Note that the algorithm requires the additional parameter $M_0 = \mathcal{O}(M)$ although the solution is insensitive to this parameter.

Intensive computations are the matrix multiplication at Step 3 and the least squares problem at Step 10, which cost $\mathcal{O}(nNd)$ and $\mathcal{O}(nM_0^2 d)$ time, respectively. The cost of least squares problem can be reduced to $\mathcal{O}(nM_0 d)$ by the conjugate gradient (CG) method as suggested in [31]. The total running time of Algorithm 2 is $\mathcal{O}(nM_0^2 Nd)$ or $\mathcal{O}(nM_0 Nd)$ using CG.

5. Experiments

We demonstrate our sparse subspace method (SSM) described in Algorithm 1. We perform face recognition experiments using a cropped version of the Extended Yale Face Database B [32, 33]. The database consists of 2,414 face images of 38 individuals. We randomly select half of the images of each subject for the training dataset ($n_k \approx 32$, $k = 1, \dots, 38$), and the other half for queries. Each image is expressed as a $d = 192 \times 168 = 32{,}256$ dimensional vector storing the grayscale values.

One-to-one classification Figure 1 shows examples of one-to-one classification. The SSM tries to answer a class label for a single query. We reduced the dimensionality to $\hat{d} = 1{,}024$ by the Gaussian random projection at Step 1 in Algorithm 1. We set the block sizes $\mathcal{N} = \{n_1, \dots, n_{38}\}$ and the sparsity level $M_0 = 4$ in Algorithm 2. Since SSM behaves as the SRC [6] when $\mathcal{N} = \mathcal{N}_1$, we also executed the SRC implemented with ROMP. The SSM and SRC, including the random projection, run in less than 0.2 seconds on a moderate workstation.

For the valid query image of subject #16 as Fig. 1(a), we see that only the residual r_{16} is significantly small. The SSM and SRC stably detect r_{16} as the smallest even if the query is contaminated with noise as shown in Fig 1(b) before the dimensionality reduction. We also observe in Fig 1(c) that none of the residuals can be significantly small for the invalid query (taken from the UMIST face database

[34]). In all cases, the residuals tend to be left undisturbed in SSM although the classification results are the same as SRC. This indicates that irrelevant class subspaces are ruled out by the block sparse model.

n-to-one classification For different numbers n of queries, we evaluated the recognition rate of n-to-one classifier with respect to reduced feature dimension $\hat{d}$ by Gaussian random projection. For $\hat{d} > 120$, the recognition rate increases with n and $\hat{d}$ as shown in Fig. 2. The rate is enhanced to more than 99% at $\hat{d} > 350$ with $n \geq 4$ queries. The perfect classification is achieved at $\hat{d} > 400$ with $n \geq 8$ queries. The n-to-one classifier provides better performance than the one-to-one classifier applied to each query, because the n-to-one classifier takes advantage of the joint sparsity. However, the SSM did not improve the recognition rate at low dimensions $\hat{d} < 120$. We should cope with this matter in the future work.

n-to-ones classification We also performed the n-to-ones classification. Figure 3 shows an example using the Extended Yale Face Database B. We gave the classifier five query images, three of which are taken from subject # 5 and two from # 29. These five queries are classified into their respective classes indicated by the significantly small residuals.

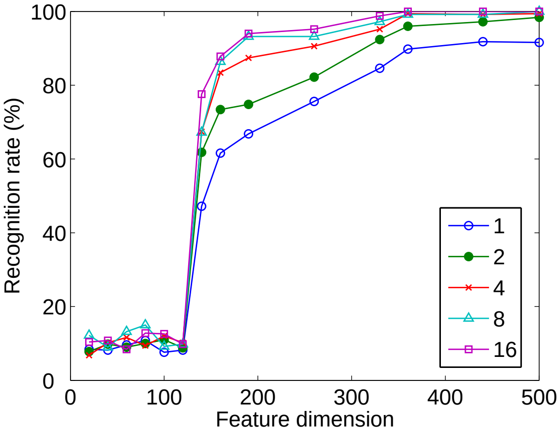

Figure 2. Recognition rates of n-to-one classifier on Extended Yale B database, with respect to feature dimension.

6. Concluding remarks

We have developed the sparse subspace method (SSM), which enables us to classify multiple queries into their respective classes, simultaneously. The SSM is based on the sparse decomposition of the query subspace. The query subspace is represented only by the relevant class subspaces. Since this sparse decomposition can be cast as the MMV problem with a row-block joint sparsity model, the uniqueness, robustness and recovery of the solution are guaranteed under the block RIP condition. We realized the block sparse decomposition by modifying the greedy algorithm ROMP. We experimentally showed that the classification of multiple queries improves the recognition rate on a face database. The joint sparsity model and the decomposition algorithm should be improved further. More detailed performance evaluation also remains in the future work.

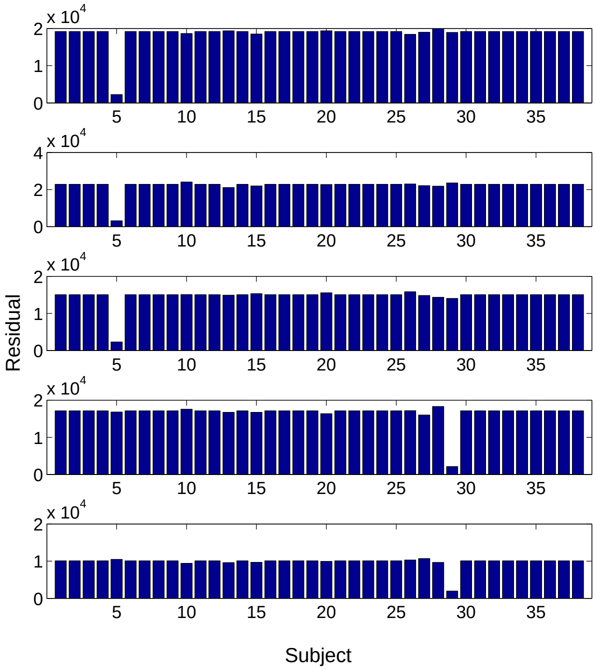

Figure 3. An example of n-to-ones classification. Residuals $r_k^{(1)}, \ldots, r_k^{(5)}$ are shown from top to bottom. Each of $n = 5$ queries is classified into one of two classes $k = 5$ and 29.

References

[1] D. Donoho, “Compressed sensing,” *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] E. J. Candès, J. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Comm. on Pure and Applied Math*, vol. 59, no. 8, pp. 1207–1223, 2006.

[4] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathematique*, vol. 346, pp. 589–592, May 2008.

[5] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, pp. 21–30, March, 2008.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[7] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, “Distributed recognition of human actions using wearable motion sensor networks,” *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, pp. 103–115, 2009.

[8] J. Wright, A. Ganesh, S. Rao, and Y. Ma, “Exact recovery of corrupted low-rank matrices by convex optimization,” in *Proc. IEEE*, 2009.

[9] J. Chen and X. Huo, “Sparse representations for multiple measurement vectors (mmv) in an over-complete dictionary,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. IV257–IV260, 2005.

[10] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, “Sparse solutions to linear inverse problems with multiple measurement vectors,” *IEEE Trans. on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[11] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a union of subspaces,” *CoRR, preprint*, vol. abs/0807.4581, 2008. informal publication.

[12] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[13] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[14] M. Rudelson, R. Vershynin, M. Rudelson, and R. Vershynin, “Geometric approach to error correcting codes and reconstruction of signals,” *Int. Math. Res. Not*, vol. 64, pp. 4019–4041, 2005.

[15] E. J. Candès, “Compressive sampling,” in *Proc. the International Congress of Mathematicians*, 2006.

[16] D. Donoho and T. J., “Counting faces of randomly projected polytopes when the projection radically lowers dimension,” *Journal of AMS*, vol. 22, no. 1, pp. 1–53, 2009.

[17] E. J. Candès and J. Romberg, “Sparsity and incoherence in compressive sampling,” *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[18] K. Maeda and S. Watanabe, “A pattern matching method with local structure (in japanese),” *IEICE Trans. Inf. and Syst.*, vol. J68-D, no. 3, pp. 345–352, 1985.

[19] O. Yamaguchi, K. Fukui, and K. Maeda, “Face recognition using temporal image sequence,” in *IEEE Int. Conf. on AFG*, pp. 318–323, 1998.

[20] K. Fukui and O. Yamaguchi, “Face recognition using multiviewpoint patterns for robot vision,” in *11th International Symposium of Robotics Research*, pp. 192–201, 2003.

[21] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, “Distributed compressed sensing,” tech. rep., Electrical and Computer Engineering Department, Rice University, 2006.

[22] M. F. Duarte, S. Sarvotham, M. Wakin, D. Baron, and R. Baraniuk, “Joint sparsity models for distributed compressed sensing,” in *Online Proc. the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2005.

[23] J. Tropp, A. Gilbert, and M. Strauss, “Algorithms for simultaneous sparse approximation. part i: Greedy pursuit,” *Signal Processing*, vol. 86, no. 3, pp. 572–588, 2006.

[24] M. Duarte, C. Hegde, V. Cevher, and R. G. Baraniuk, “Recovery of compressible signals in unions of subspaces,” in *Conference on Information Sciences and Systems*, (Baltimore, MD), 2009.

[25] X. H. Jie Chen, “Theoretical results on sparse representations of multiple-measurement vectors,” *IEEE Trans. Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[26] Y. C. Pati, R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, pp. 40–44, 1993.

[27] G. Davis, S. Mallat, and M. Avellaneda, “Greedy adaptive approximation,” *Journal of Constructive Approximation*, vol. 13, pp. 57–98, 1997.

[28] J. A. Trop, “Greed is good: Algorithmic results for sparse approximation,” *IEEE Trans. Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[29] D. Donoho, M. Elad, and V. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[30] J. A. Tropp, Anna, and C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[31] D. Needell and R. Vershynin, “Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit,” *Foundations of Computational Mathematic*, vol. 9, no. 3, pp. 317–334, 2009.

[32] A. Georghiades, P. Belhumeur, and D. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[33] K. Lee, J. Ho, and D. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[34] D. B. Graham and N. M. Allinson, “Characterizing virtual eigensignatures for general purpose face recognition,” in *Face Recognition: From Theory to Applications*, vol. 163, pp. 446–456, NATO ASI Series F, Computer and Systems Sciences, 1998.