

Connecting the Dots: Image Classification via Sparse Representation from a Constrained Subspace Perspective

Liang Liao¹, Stephen John Maybank², Haichang Ye¹, Xin Liu³ and Xinqiang Wang¹

¹School of Electronics and Information Engineering Zhongyuan University of Technology, Zhengzhou, P. R. China

²Birkbeck college University of London, London, UK

³Information Center, Yellow River Conservancy Commission, Zhengzhou, P. R. China

Abstract—We consider the problem of classifier design via sparse representation based on a constrained subspace model. We argue that the data points in the linear span of the training samples should be constrained in order to yield a more accurate approximation to the corresponding data manifold. For this purpose, the constrained set of data points is formulated as a union of affine subspaces in the form of affine hulls spanned by training samples. We argue that the intrinsic dimension of the affine subspaces should be equal to that of data manifold. Thus, a classifier based on this model has a high classification accuracy similar to that of the conceptual NM (Nearest Manifold) classifier. Based on this model, we connect the dots of some classical classifiers including NN (Nearest Neighbor), NFL (Nearest Feature Line), NS (Nearest subspace) and the recently emerged state-of-the-art SRC (Sparse Representation Classifiers) and interpret the mechanism of SRC and Yang's variant of the SRC using the constrained subspace perspective. Experiments on the Extended Yale B database for image classification corroborate our claims and demonstrate the possibility of a proposed classifier called NCSC-CSR which has higher classification accuracy and robustness.

Keywords—sparse representation; constrained subspace; manifold approximation

I. INTRODUCTION

An $m \times n$ image with 256 grayscales for each pixel has 256^{mn} pixel configurations, but only a few correspond to particular classes of objects. Since image representation usually has a high redundancy, it is well accepted that the images of a particular class usually belong to a data manifold with low intrinsic dimension (id). Briefly speaking, when images are represented as vectors in a feature space of high dimension, the manifold of an image class is a nonlinear geometric structure which describes the distribution of observed image samples and can serve as a universal set for this class. If the geometrical properties of manifold can be learned from a finite number of observed data samples, it can be exploited for pattern recognition and for many other applications in computer vision [1] [2].

On the other hand, due to the high redundancy of image representations, a sparse representation technique can be employed for these applications [3]-[8]. For example, the recently emerged state-of-the-art algorithm SRC (Sparse

Representation Classifier), proposed by J. Wright et al., exploits the sparseness of a vector solution to an underdetermined linear system of equations for classifier design and yields impressive results for robust face recognition [3], [4]. Although there have been some papers on the mechanism of sparse representation and inference [8]- [9] and some interesting work that challenges these efforts [11], a simple and intuitive interpretation of the mechanism of sparse representation for inference is still unavailable.

We are particularly interested in the issue from the viewpoint of manifold approximation, intrinsic dimension and constrained least-squares and seek the constrained sparse solution vector for a linear system of equations. Based on these discussions, we propose a model called *Constrained Subspace* and establish a classifier framework connecting the dots of NN (Nearest Neighbor), NFL (Nearest Feature Line), NS (Nearest Subspace), the recently emerged SRC and some of its variants. We then offer an intuitive interpretation of the mechanism of sparse representation as applied to image classification.

The remainder of this paper is organized as follows. In Section II, some background material on NM (Nearest Manifold), NN, NFL, NS and intrinsic dimension is briefly introduced. In Section III, a constrained subspace model and a novel classifier called NCSC-CSR, based on the constrained subspace model, are proposed. The constrained subspace model establishes the connection between NN, NFL, NS and NM from the perspective of manifold approximation. In Section IV, an intuitive interpretation of the mechanism of SRC and Yang's method (a variant of SRC) for image classification is offered. Some experimental results obtained by applying the relevant classifiers to the Extended Yale B database are given in Section V. The classification accuracy and the robustness of the classifiers are assessed. The experiment results corroborate our claims. Finally, Section VI concludes this paper and suggests some possible future work.

II. BACKGROUND

We describe some classical classifiers including NN, NFL, NS in a generalized framework from the viewpoint of manifold approximation. Before our detailed discussion, some background material is given.

A. NM: Nearest Manifold

Given a query sample \mathbf{y} belonging to one of K classes, if the data manifolds $\mathcal{M}_1, \dots, \mathcal{M}_K$ of the K classes are known, the classifier NM (Nearest Manifold) classifies \mathbf{y} as follows.

$$\begin{cases} \text{class}(\mathbf{y}) = \operatorname{argmin}_{i \in \{1, 2, \dots, K\}} r_i(\mathbf{y}) \\ r_i(\mathbf{y}) = \min_{\mathbf{x} \in \mathcal{M}_i} \|\mathbf{y} - \mathbf{x}\|_2 \end{cases} \quad (1)$$

where \mathbf{x} is the sample on \mathcal{M}_i and $r_i(\mathbf{y})$ is the distance of \mathbf{y} to class i (i.e., the projection distance to \mathcal{M}_i) for all $i = 1, \dots, K$.

Since \mathcal{M}_i serves as the universal set of samples in class i for all i , if and only if an uncorrupted query sample \mathbf{y} belongs to class i , we have $r_i(\mathbf{y}) = 0$. Thus, by seeking the minimum distance of \mathbf{y} to a particular manifold, NM classifies \mathbf{y} in a simple way.

B. NN, NFL, NS and Intrinsic Dimension

The NM classifier is only conceptual because in practice only a limited number of samples are available and it is not possible to construct the data manifold accurately. Thus, the three classical classifiers NN, NFL [12] and NS are widely used as alternatives to NM.

Given query sample \mathbf{y} and training samples belonging to K classes, NN, NFL and NS classify \mathbf{y} as follows.

$$\begin{cases} \text{class}(\mathbf{y}) = \operatorname{argmin}_{i \in \{1, 2, \dots, K\}} r_i(\mathbf{y}) \\ r_i(\mathbf{y}) = \min_{\mathbf{x} \in \mathbb{M}_i} \|\mathbf{y} - \mathbf{x}\|_2 \end{cases} \quad (2)$$

where \mathbb{M}_i is a *training superset* of class i for the particular classifier in the corresponding question.

Given N_i training samples (vectors) of class i denoted by $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(N_i)}$ for all i , the training superset \mathbb{M}_i for the NN classifier is a point set defined as follows.

$$\mathbb{M}_i = \left\{ \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(N_i)} \right\}. \quad (3)$$

For the NFL classifier, \mathbb{M}_i is a set of lines defined as follows.

$$\mathbb{M}_i = \left\{ \lambda \mathbf{x}_i^{(a)} + (1 - \lambda) \mathbf{x}_i^{(b)} \mid a, b \in \{1, \dots, N_i\} \right\}. \quad (4)$$

For the NS classifier, \mathbb{M}_i is the linear subspace defined as follows.

$$\mathbb{M}_i = \left\{ \mathbf{A}_i \boldsymbol{\alpha}_i \mid \boldsymbol{\alpha}_i \in \mathbb{R}^{N_i} \right\} \quad (5)$$

where $\mathbf{A}_i = \left[\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(N_i)} \right]$ is the base matrix.

Thereinafter, for notational clarity, we respectively denote \mathbb{M}_i in equations 3, 4 and 5 by \mathbb{M}_i^{NN} , $\mathbb{M}_i^{\text{NFL}}$ and \mathbb{M}_i^{NS} .

Since NN, NFL, NS and NM employ the same strategy of the nearest point to classify a query sample, we argue that \mathbb{M}_i^{NN} , $\mathbb{M}_i^{\text{NFL}}$ and \mathbb{M}_i^{NS} are approximations to \mathcal{M}_i , but none of them is necessarily an accurate approximation to \mathcal{M}_i in terms of *intrinsic dimension* (id), where intrinsic dimension is defined as the minimum number of parameters needed to describe the data structure such that the fundamental properties of the data are preserved [13].

Since \mathbb{M}_i^{NN} is a point set and $\mathbb{M}_i^{\text{NFL}}$ is a set of lines, the id of \mathbb{M}_i^{NN} and $\mathbb{M}_i^{\text{NFL}}$ is respectively 0 and 1. Moreover, if the column vectors of \mathbf{A}_i are linear independent, the id of \mathbb{M}_i^{NS} is equal to N_i , namely, the number of vectors in a basis for the subspace \mathbb{M}_i^{NS} .

III. FROM CONSTRAINED SUBSPACE TO NCSC-CSR

A. Constrained Subspace Model

From equations 3, 4 and 5, it is not difficult to see that

$$\mathbb{M}_i^{\text{NN}} \subset \mathbb{M}_i^{\text{NFL}} \subset \mathbb{M}_i^{\text{NS}}, \quad \forall i. \quad (6)$$

Since \mathbb{M}_i^{NN} and $\mathbb{M}_i^{\text{NFL}}$ are subsets of \mathbb{M}_i^{NS} , we call them *constrained subspaces* with id respectively equal to 0 and 1.

Based on equations 3, 4 and 5, a more generalized constrained subspace $\mathbb{M}_i^{\text{NCSC}}$ can be formulated using a constrained ℓ_0 -norm sparse representation as follows.

$$\mathbb{M}_i^{\text{NCSC}} = \left\{ \mathbf{A}_i \boldsymbol{\alpha}_i \mid \begin{array}{l} \boldsymbol{\alpha}_i \in \mathbb{R}^{N_i}, \\ \mathbf{1}_{N_i}^T \boldsymbol{\alpha}_i = 1 \text{ and } \|\boldsymbol{\alpha}_i\|_0 \leq \kappa \leq N_i \end{array} \right\} \quad (7)$$

where $\mathbf{1}_{N_i}$ is the column vector of length N_i , whose entries are all equal to 1 and κ , a sparsity parameter of $\boldsymbol{\alpha}_i$, is a fixed integer.

When $\kappa = 1$, $\mathbb{M}_i^{\text{NCSC}}$ becomes \mathbb{M}_i^{NN} . When $\kappa = 2$, $\mathbb{M}_i^{\text{NCSC}}$ becomes $\mathbb{M}_i^{\text{NFL}}$. It follows from equation (7) that

$$\text{id}(\mathbb{M}_i^{\text{NCSC}}) = \kappa - 1. \quad (8)$$

The set $\mathbb{M}_i^{\text{NCSC}}$ can be rewritten as a union of a series of affine hulls $\mathcal{H}_{i,j}$ spanned by κ samples as follows.

$$\begin{cases} \mathbb{M}_i^{\text{NCSC}} = \bigcup_{j=1}^{\binom{N_i}{\kappa}} \mathcal{H}_{i,j} \\ \mathcal{H}_{i,j} = \left\{ \mathbf{A}_i^{(j)} \boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^\kappa \text{ and } \mathbf{1}_\kappa^T \boldsymbol{\beta} = 1 \right\} \end{cases} \quad (9)$$

where $\mathbf{A}_i^{(j)}$ is the base matrix with κ columns which are chosen as the j -th κ -combination of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_i}$ (the column order of $\mathbf{A}_i^{(j)}$ does not matter).

Note that the affine hull is also mathematically referred to as linear manifold. The essence of $\mathbb{M}_i^{\text{NCSC}}$ is the use of a series of linear manifolds to approximate the nonlinear set \mathcal{M}_i .

In our previous works [14], [15], we have proposed a generalized classifier framework called NCSC (Nearest Constrained Subspace Classifier), which classifies a vector \mathbf{y} as follows.

$$\begin{cases} \text{class}(\mathbf{y}) = \operatorname{argmin}_{i \in \{1, 2, \dots, K\}} r_i(\mathbf{y}) \\ r_i(\mathbf{y}) = \min_{\mathbf{x} \in \mathbb{M}_i^{\text{NCSC}}} \|\mathbf{y} - \mathbf{x}\|_2 \end{cases} \quad (10)$$

In equation (10), $r_i(\mathbf{y})$ can also be rewritten in the form of a constrained the-least-squares problem as follows.

$$\begin{aligned} r_i(\mathbf{y}) &= \min_{\alpha_i \in \mathbb{R}^{N_i}} \|\mathbf{y} - \mathbf{A}_i \alpha_i\|_2 \\ \text{subject to } &\|\alpha_i\|_0 \leq \kappa \text{ and } \mathbf{1}_{N_i}^T \alpha_i = 1 \end{aligned} \quad (11)$$

NCSC includes NN and NFL as special cases with $\kappa = 1$ and 2 respectively. There is a close relationship between NCSC and NS in which $\|\alpha_i\|_0 = N_i$ and the constraint $\mathbf{1}_{N_i}^T \alpha_i = 1$ is removed.

In order to make NCSC a more accurate approximation to NM, we assume the intrinsic dimension of the employed constrained subspace should at least match the intrinsic dimension of the corresponding manifold. This assumption naturally leads to the problem of estimating the intrinsic dimension. Although there are some papers on this issue [16]-[19], their estimates often differ because of the different models and algorithms that are employed, and also because of different choices of the observation parameter{s}. A comprehensive comparison and evaluation of method for estimating the id has still to be made. We leave the problem of intrinsic dimension estimators to a possible future work and adopt a trial-and-error approach to tune the intrinsic dimension parameter κ of NCSC.

B. Global NCSC and COMP

NN, NFL, NS (equation 2) and NCSC as formulated in equation 10 are all the class-wise methods which calculate the distance of a query sample to all classes. If a query sample \mathbf{y} lies near to the borders between several classes (i.e., data manifolds), then, the distances $r_i(\mathbf{y})$ are similar and the classification accuracy is reduced.

Interestingly, the state-of-the-art SRC offers an interesting approach in which all classes collaborate and compete with each other to produce a sparse solution to a linear system [3], [20]. This approach can extend a class-wise classifier to a ‘‘class-intelligent’’ one. Briefly speaking, the approach is to solve the following ℓ_1 -norm minimization problem, whose solution is also the sparsest (with the minimal ℓ_0 -norm) for most large underdetermined system of linear equations [21], [22].

$$\alpha^* = \operatorname{argmin}_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\alpha \quad (12)$$

where $N = \sum_{i=1}^K N_i$ is the number of training samples for all the classes and $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_K]$.

When α^* is obtained, SRC classifies \mathbf{y} as follows.

$$\begin{cases} \text{class}(\mathbf{y}) = \operatorname{argmin}_{i \in \{1, 2, \dots, K\}} r_i(\mathbf{y}) \\ r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_i(\alpha^*)\|_2 \end{cases} \quad (13)$$

where $\delta_i : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is the feature function, which returns a sparse vector, whose nonzero entries are the entries of α^* associated with the training samples of class i . More mathematically, denoting the index set of the training samples of class i in \mathbf{A} by \mathcal{S}_i and δ_N^j as the sparse column vector of length N with entry j equal to 1 and all other entries equal to 0, then, given $\alpha \in \mathbb{R}^N$, $\delta_i(\alpha)$ is defined as follows.

$$\delta_i(\alpha) = \sum_{j \in \mathcal{S}_i} \operatorname{diag}\{\delta_N^j \alpha^T\} \quad (14)$$

Using this strategy of optimizing a sparse vector in \mathbb{R}^N (equation 12) with the help of $\delta_i(\cdot)$ rather than directly optimizing K sparse vectors in \mathbb{R}^{N_i} for all i , SRC outperforms in terms of classification accuracy a wide range of class-wise classifiers. Note that the most important reason that SRC optimizes a sparse vector in \mathbb{R}^N rather than K sparse vectors in \mathbb{R}^{N_i} is that equation 12 is designed for an underdetermined linear system, in that the dimension of a target sparse vector is far larger than the row number of \mathbf{A} .

Similarly, it is convenient to extend the local (class-wise) NCSC (equations 10 and 11) to a global (class-intelligent) classifier using

$$\begin{aligned} \alpha^* &= \operatorname{argmin}_{\alpha \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{A}\alpha\|_2 \\ \text{subject to } &\|\alpha\|_0 \leq \kappa \text{ and } \mathbf{1}_N^T \alpha = 1 \end{aligned} \quad (15)$$

Equation 15 is a typical problem of combination optimization. To avoid such a N-P hard problem, we offer a greedy algorithm called COMP (Constrained Orthogonal Matching Pursuit) for approximately solving for α^* in equation 15.

Before a detailed discussion, let's see how to remove constraint $\mathbf{1}_N^T \alpha = 1$ in equation 15.

Denote $\mathbf{A} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ and $\mathbf{A}^- = \mathbf{A} - \mathbf{x}_j \mathbf{1}_N^T$ where \mathbf{x}_j is any element in the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, then, the solution of equation 15 is given as follows.

$$\begin{cases} \alpha^* = \beta^* + (1 - \mathbf{1}_N^T \beta^*) \delta_N^j \\ \beta^* = \operatorname{argmin}_{\beta \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{x}_j - \mathbf{A}^- \beta\|_2 \\ \text{subject to } \|\beta\|_0 \leq \kappa - 1 \end{cases} \quad (16)$$

With the constraint $\mathbf{1}_N^T \alpha = 1$ removed, the problem for solving α^* in equation 15 becomes the problem for solving β^*

in equation 16, but the problem for solving β^* is still N-P hard. To avoid such a N-P hard problem, we propose an optimizer called COMP (Constrained Orthogonal Matching Pursuit), an extension to OMP (Orthogonal Matching Pursuit), given as algorithm 1, for obtaining an approximation to α^* in equation 16.

We denote the set of indices of non-zero entries of α by a support set \mathcal{S} , namely, $\text{support}(\alpha) = \mathcal{S}$. We also define x_j , called the principle support vector, to be the vector nearest to \mathbf{y} in that

$$\begin{aligned} x_j &= \underset{x_k \in \{x_1, x_2, \dots, x_N\}}{\text{argmin}} \left(\|\mathbf{y}\|_2^2 - \frac{(x_k^T \mathbf{y})^2}{\|x_k\|_2^2} \right) \\ &\equiv \underset{x_k \in \{x_1, x_2, \dots, x_N\}, \lambda \in \mathbb{R}}{\text{argmin}} \|\mathbf{y} - \lambda x_k\|_2 \end{aligned} \quad (17)$$

Then, using the same strategy as SRC with the help of the feature function $\delta_i(\cdot)$, given a query sample, N training samples of K classes and a intrinsic dimension parameter κ , we propose a global (class-intelligent) classifier NCSC-CSR in Algorithm 2.

Algorithm 1 $\alpha^* = \text{COMP}(\mathbf{A}, \mathbf{y}, \kappa, \epsilon)$ — COMP solution of equation (15) (or equivalently equation (16)).

Input: Query sample \mathbf{y} , $\mathbf{A} = [x_1, x_2, \dots, x_N]$, parameter $\kappa > 1$ and a predefined small tolerance constant $\epsilon > 0$.

Output: $\alpha^* \in \mathbb{R}^N$ subject to $\|\alpha^*\|_0 \leq \kappa$ and $\mathbf{1}_N^T \alpha^* = 1$.

- 1: Calculate x_j as in equation (17);
 - 2: Support set $\mathcal{S} \leftarrow \{j\}$ and its complementary set $\mathcal{S}_C \leftarrow \{1, 2, \dots, N\} \setminus \{j\}$;
 - 3: $\mathbf{A}^- \leftarrow \mathbf{A} - x_j \mathbf{1}_N^T \equiv [x_1 - x_j, \dots, x_N - x_j]$;
 - 4: $\mathbf{r} \leftarrow \mathbf{y} - x_j$;
 - 5: **while** $\|\mathbf{r}\|_2 > \epsilon$ **and** $|\mathcal{S}| < \kappa$ **do**
 - 6: $k^* \leftarrow \underset{k \in \mathcal{S}_C}{\text{argmin}} \left(\|\mathbf{r}\|_2^2 - \frac{((x_k - x_j)^T \mathbf{r})^2}{\|x_k - x_j\|_2^2} \right)$;
 - 7: $\mathcal{S} \leftarrow \mathcal{S} \cup \{k^*\}$;
 - 8: $\mathcal{S}_C \leftarrow \mathcal{S}_C \setminus \{k^*\}$;
 - 9: Solve the following problem
 $\beta^* \leftarrow \underset{\beta \in \mathbb{R}^N}{\text{argmin}} \|\mathbf{y} - x_j - \mathbf{A}^- \beta\|_2$
subject to $\text{support}(\beta) = \mathcal{S} \setminus \{j\}$;
 - 10: $\mathbf{r} \leftarrow \mathbf{y} - x_j - \mathbf{A}^- \beta^*$;
 - 11: **end while**
 - 12: **return** $\alpha^* \leftarrow \beta^* + (1 - \mathbf{1}_N^T \beta^*) \delta_N^j$;
-

Algorithm 2 NCSC-CSR — Nearest Constrained Subspace Classifier via Constrained Sparse Representation

Input: Query sample \mathbf{y} , training samples x_1, x_2, \dots, x_N of K classes, parameter κ satisfying $\kappa \leq N$ and a predefined small tolerance constant $\epsilon > 0$.

Output: Class label of \mathbf{y} .

- 1: $\mathbf{A} \leftarrow [x_1, x_2, \dots, x_N]$;
 - 2: $\alpha^* \leftarrow \text{COMP}(\mathbf{A}, \mathbf{y}, \kappa, \epsilon)$;
 - 3: $r_i \leftarrow \|\mathbf{y} - \mathbf{A} \delta_i(\alpha^*)\|_2 \quad \forall i = 1, 2, \dots, K$;
 - 4: **return** $\text{class}(\mathbf{y}) \leftarrow \underset{i \in \{1, 2, \dots, K\}}{\text{argmin}} r_i$;
-

IV. SPARSE REPRESENTATION FROM CONSTRAINED SUBSPACE PERSPECTIVE

A number of papers provide explanations for the success of SRC. For example, Zhang et al [20] contend that the collaboration between the training samples of candidate classes contributes the success of SRC. In another interesting work, Yang et al [8] propose a variant of SRC by adding a constraint $\mathbf{1}_N^T \alpha = 1$ to equation 12. Namely

$$\begin{aligned} \alpha^* &= \underset{\alpha \in \mathbb{R}^N}{\text{argmin}} \|\alpha\|_1 \\ \text{subject to } &\mathbf{y} = \mathbf{A} \alpha \text{ and } \mathbf{1}_N^T \alpha = 1. \end{aligned} \quad (18)$$

It is reported that Yang's method based on equation 18 slightly outperforms SRC in terms of classification accuracy. We argue that the mechanism of Yang's method is to find the local linear manifold (as a patch from a constrained subspace) with the minimal intrinsic dimension, which the query sample \mathbf{y} happens to lie on.

On the other hand, since the given samples are usually limited rather than overcomplete, NCSC-CSR, which connects the dots with NN, NFL, NS and NM, relaxes the condition $\mathbf{y} = \mathbf{A} \alpha$ and resorts to the constrained least-squares approach with a ℓ_0 -norm sparsity parameter κ matched with the intrinsic dimension of the data manifold (i.e., $\kappa = \text{id} + 1$, where id is the intrinsic dimension of the data manifold) to seek the nearest distance to query sample \mathbf{y} .

We contend that if κ is well-tuned, the classification accuracy of NCSC-CSR can be higher than the accuracies of SRC and Yang's method.

V. EXPERIMENTS AND VERIFICATIONS

In this section, we evaluate the relevant classifiers in terms of classification accuracy and robustness on the publicly available Extended Yale B database for supervised image classification.

The Extended Yale B database contains 2,414 frontal-face images which belong to 38 classes (individuals), with roughly 64 images per class. We use the cropped version of the database, in which each image is cropped to the size 192×168 , for our classification experiments [23].

Although image features play an important role in applications of image recognition and computer vision, in this paper, we are only interested in evaluating particular classifiers rather image features. To be focused, we only use simple image features for our experiments. For this reason, we down-sample the experimental images using ratios of 1/32, 1/24, 1/16 and 1/8. Thus, the extracted image vectors are respectively 30, 56, 120 and 504 dimensional.

A. Comparison of Classification Accuracies

In this section, we evaluate the classification accuracies of the relevant classifiers. In order to get stable classification accuracies from a sufficient number of query samples, our experiments contain multiple rounds of classification. In each round, 1,216 images (32 images/class \times 38 classes) are

randomly selected as the training samples. The remaining 1,198 images are retained as the query samples (roughly 32 images per class).

In order to avoid unnecessary perturbations of the accuracy, in each round, the training and query samples are preserved and downsampled with different ratios for all the relevant classifiers. After 10 rounds, the classification accuracy of each classifier is given as follows. $accuracy = \frac{w}{W}$ where w is the total number of correctly classified test samples and W is the total number of query samples, which is equal to 11980 (1198 samples/round \times 10 rounds).

Using different feature dimensions of image samples, the overall classification accuracies of NN on all query samples in 10 rounds, and other statistics of each round are given in table I.

Although it is computationally efficient, NN is a weak classifier with classification accuracies not larger than 0.627 in Table I. Other relevant classifiers have much higher classification accuracies. Since a large number of query samples which NN can correctly classify only makes the accuracies of stronger classifiers consistently high and difficult to distinguish from each other, we use the results of NN as a borderline in the following experiments. More specifically, in each round of classification, we keep the incorrectly classified query samples by NN as the *NN-challenging* samples. Then, we use the *NN-challenging* samples to query other classifiers, which include SRC, Yang's method (and their variants) and our NCSC-CSR with a range of values of κ . It is found that given a specific feature dimension and its corresponding *accuracy* in Table I, the total number of the *NN-challenging* query samples is equal to 1,198 samples/round \times 10 rounds \times (1 - *accuracy*).

TABLE I. OVERALL CLASSIFICATION ACCURACIES OF NN FOR 10 ROUNDS ON THE EXTENDED YALE B DATABASE WITH DIFFERENT FEATURE DIMENSIONS. ACCURACY= OVERALL CLASSIFICATION ACCURACY FOR 10 ROUNDS; STD=STANDARD DEVIATION OF THE ACCURACY OF EACH ROUND FOR 10 ROUNDS; VAR =VARIANCE OF THE ACCURACY OF EACH ROUND FOR 10 ROUNDS

statistics	dimension			
	30	56	120	504
<i>accuracy</i>	0.525	0.575	0.627	0.676
<i>std</i>	0.012	0.008	0.010	0.010
<i>var</i> ($\times 10^{-5}$)	15.448	6.619	10.858	9.572

Using the *NN-challenging* samples of the Extended Yale B database with different feature dimensions, Table II gives the classification accuracies of SRC (equation 12) and Yang's method (equation 18) with ℓ_1 -norm optimization as well as SRC with OMP as the optimizer and Yang's method with COMP as the optimizer.

To implement Yang-L1, we employ the scheme proposed by Yang et al [8] to remove the constraint $\mathbf{1}_N^T \alpha = 1$. More specifically, by setting

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix}, \quad \hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ \mathbf{1}_N^T \end{bmatrix}, \quad (19)$$

TABLE II. OVERALL ACCURACIES OF SRC AND YANG'S METHOD ON THE NN-CHALLENGING SAMPLES OF THE EXTENDED YALE B IMAGE DATABASE. SRC-L1=SRC VIA ℓ_1 -NORM OPTIMIZER; YANG-L1 = YANG'S METHOD VIA ℓ_1 -NORM OPTIMIZER; SRC-OMP=SRC VIA OMP; YANG-COMP = YANG'S METHOD VIA COMP

method	dimension			
	30	56	120	504
SRC-L1	0.561	0.727	0.841	0.905
Yang-L1 [†]	0.577	0.740	0.841	0.906
SRC-OMP	0.257	0.378	0.663	0.829
Yang-COMP	0.431	0.601	0.747	0.833

The strongest classifier in terms of accuracy this table

equation 18 can be rewritten as the following ℓ_1 -norm optimization problem, which can be solved with the same tool employed by SRC.

$$\alpha^* = \operatorname{argmin}_{\alpha \in \mathbb{R}^N} \|\alpha\|_1 \text{ subject to } \hat{\mathbf{y}} = \hat{\mathbf{A}}\alpha. \quad (20)$$

For a comparison purpose, Table I also gives the accuracies of two variants respectively called SRC-OMP and Yang-COMP. Among them, SRC-OMP is a variant of SRC with OMP as an approximate optimizer for equation 12. Yang-COMP is a variant of Yang's method with COMP (Algorithm 1) as the optimizer for equation 18.

Since Yang's method satisfies the condition $\mathbf{y} = \mathbf{A}\alpha$ and it is always the case in Algorithm 1 (COMP) that $|\mathcal{S}|$ is not larger than N , the column number of \mathbf{A} , it follows that in using COMP as the optimizer for equation 18, Yang-COMP is a special case of NCSC-CSR with $\kappa = N = \sum_{i=1}^K N_i$ where in this experiment, $N = 1216$.

An observation from Table II is that Yang-L1 outperforms SRC-L1 and Yang-COMP outperforms SRC-OMP in terms of classification accuracy, although not always significantly. This observation is consistent with the claim in [8]. We argue that the advantage of Yang's method is due to the virtues of the constrained subspace model, which seeks a local linear manifold, rather than a local patch of subspace, which has the nearest distance to query samples.

Another observation from Table II is that SRC-OMP is outperformed by SRC-L1 and Yang-COMP is outperformed by Yang-L1 in terms of accuracy. We argue it is primarily due to the following two reasons: (1) In Yang-COMP, κ is not tuned. (2) OMP and COMP are the greedy methods based on the least-squares and are apt to converge to a local minimum rather than the global one, which is obtained by the ℓ_1 -norm optimizer.

Although deficient as sparsity optimizers, OMP and COMP have their strengths, in that they do not require the training samples to be overcomplete while ℓ_1 -norm optimizer needs the number of training samples to be large enough in order to have a solution to equation 12. The fact that OMP and COMP require fewer training samples makes them more applicable in scenarios where the number of training samples are limited and SRC-L1 and Yang-L1 have no solutions.

But we are more interested in whether NCSC-CSR with a tuned κ can outperform SRC-L1 or even Yang-L1 in terms of

accuracy, and how κ affects the classification accuracy of NCSC-CSR.

For this purpose, we evaluate NCSC-CSR with a range of values of κ by repeating the same experiments summarized in Table II on the *NN-challenging* query samples. Figure 1 gives the accuracy curves of NCSC-CSR with different values of κ over dimensions of 30, 56, 120 and 504.

It can be seen from Figure 1 that NCSC with $\kappa = 5$ yields the lowest classification accuracy on each feature dimension. On the other hand, the curves of NCSC respectively with $\kappa = 10, 15, 50, 100, 200$ overlap with each other. Based on this observation, we argue when κ is large enough, to yield the optimal classification accuracies for NCSC-CSR, κ of NCSC-CSR should be micro-tuned in the appropriate range.

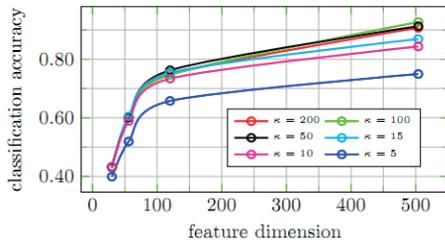


FIGURE I. ACCURACY CURVES OF NCSC-CSR RESPECTIVELY WITH $\kappa = 5, 10, 15, 50, 100, 200$, OVER FEATURE DIMENSIONS OF 30, 56, 120 AND 504 ON THE *NN-CHALLENGING* SAMPLES

Table III gives more details about the classification curves in Figure 1. For comparison purposes, it also gives the results of NCSC-CSR with $\kappa \in \{2, 400, 1216\}$. Note that the results of NCSC-CSR with $\kappa = 1216$ are exactly the results of Yang-COMP. On the other hand, NCSC-CSR with $\kappa = 2$ is a variant of NFL.

Table III indicates that the highest accuracy, corresponding to the tuned κ , depends on the feature dimension. More specifically, on feature dimensions 30, 56, 120 and 504, the highest classification accuracies are respectively 0.433, 0.605, 0.762 and 0.927. The corresponding tuned κ is respectively equal to (or near to) 10, 15, 50 and 100. This observation from Table III that the tuned κ is an increasing function of feature dimension is consistent with the intuition that a dataset with a higher feature dimension usually possesses higher intrinsic dimension.

It is noted that for the given feature dimension of 504, NCSC-CSR with $\kappa = 100$ yields so far the highest classification accuracy of 0.927, which is higher than the highest accuracy of SRC-L1 (0.906) and Yang-L1 (0.906) with exactly the same experiment settings. This shows the possibility of a stronger classifier with a tuned local sparsity optimizer (e.g., NCSC-CSR) outperforming, in terms of classification accuracy, SRC and its variants with ℓ_1 -norm optimizer.

TABLE III. CLASSIFICATION ACCURACIES OF NCSC-CSR WITH DIFFERENT VALUES OF κ OVER DIFFERENT FEATURE DIMENSION ON THE *NN-CHALLENGING* SAMPLES OF THE EXTENDED YALE B DATABASE

κ	dimension			
	30	56	120	504
1216 [†]	0.431	0.601	0.747	0.833
400	0.431	0.601	0.747	0.846
200	0.431	0.601	0.747	0.907
100	0.431	0.601	0.747	0.927*
50	0.431	0.601	0.762	0.913
15	0.433	0.605	0.754	0.869
10	0.433	0.589	0.734	0.843
5	0.399	0.519	0.658	0.749
2 [‡]	0.264	0.313	0.378	0.436

[†] NCSC-CSR with $\kappa = 1216$ is literally Yang-COMP
[‡] NCSC-CSR with $\kappa = 2$ is a variant of NFL based on Yang-COMP
^{*} The highest accuracy in this table

B. Comparison of Robustness

After the evaluation of classification accuracy, we evaluate the robustness of the classifiers on the corrupted *NN-challenging* samples of the Extended Yale B database.

In this experiment, with a sampling ratio of 1/8, all samples from the Extended Yale B database are downsampled to vectors of 504 dimensions. Then, using the same selection setting mentioned in section V for training samples and query samples, in each round the *NN-challenging* query samples are obtained by sifting the samples classified by NN.

There are 3 classification rounds employed for this evaluation. The total number of the *NN-challenging* query samples is 1,172. Note that the classification accuracies of NN in these three rounds are respectively 0.682, 0.665 and 0.674, thus, $1172 = 1198 \times (3 - 0.682 - 0.665 - 0.674)$. On the other hand, the accuracy in Table I is the average for 10 random classification rounds. The training samples and the *NN-challenging* samples are kept unchanged for evaluating all classifiers. Then, the 1172 query samples are corrupted by random noise with noise level r ($r = 0.1, 0.2, \dots, 0.9$). More specifically, given noise level r , the random corrupted pixels are uniformly distributed in a query image. The number of corrupted pixels is equal to the round number of $504 \times r$ and the corrupted pixel intensity is uniformly distributed in $\{0, 1, \dots, 255\}$.

In such a noisy environment, Yang-L1, SRC-OMP, Yang-COMP and NCSC-CSR are evaluated. Figure 2 shows the accuracy curves of SRC-L1, Yang-L1, SRC-OMP, Yang-COMP, and NCSC-CSR respectively with $\kappa = 50$ and 100 on corrupted *NN-challenging* query samples. When $r = 0$, corresponding to the classification on uncorrupted query samples, the accuracies are extracted from last columns of Tables II and III.

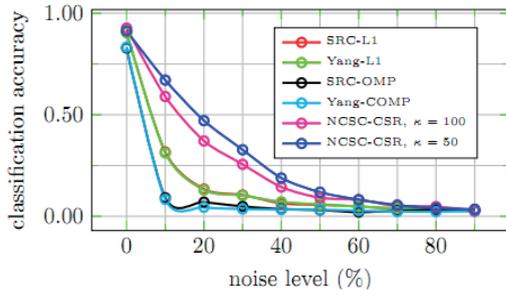


FIGURE II. ACCURACY CURVES OF SRC-L1, YANG-L1, SRC-OMP, YANG-COMP AND NCSC-CSR RESPECTIVELY WITH $\kappa = 50$ AND 100 ON CORRUPTED NN-CHALLENGING QUERY SAMPLES

It can be seen from Figure 3 that NCSC-CSR with $\kappa = 50$ or 100 yields higher accuracies than SRC-L1, Yang-L1, SRC-OMP and Yang-COMP. For example, when noise level $r = 0.1$, the accuracies of SRC-L1, Yang-L1, SRC-OMP and

Yang-COMP dramatically decrease to less than 0.312, while, on the other hand, NCSC-CSR respectively with $\kappa = 50$ and 100 yields a classification accuracy larger than 0.590.

Table IV gives more details of the accuracies of SRC-L1, Yang-L1 and SRC-OMP, Yang-COMP and NCSC-CSR respectively with $\kappa = 400, 200, 100, 50, 10, 5$ and 2 in the noisy environment. It can be seen from Table IV that, for different $r \in \{0, 0.1, \dots, 0.9\}$, the highest classification accuracies are all yielded by NCSC-CSR with tuned κ . The highest accuracies are respectively 0.927, 0.730, 0.590, 0.454, 0.282, 0.170, 0.113, 0.06, 0.058 and 0.034. The corresponding tuned values of κ are respectively 100, 15, 10, 10, 5, 2, 2, 15 and 50.

TABLE IV. CLASSIFICATION ACCURACIES OF SRC-L1, YANG-L1, SRC-OMP, YANG-COMP AND NCSC-CSR WITH A VARIETY OF VALUES OF κ ON THE NN-CHALLENGING QUERY SAMPLES OF THE EXTENDED YALE B IMAGE DATABASE. NCSC=NCSC-CSR(ALGORITHM 2.)

method		noise level (%)									
		0	10	20	30	40	50	60	70	80	90
SRC-L1		0.905	0.317	0.134	0.105	0.064	0.056	0.049	0.037	0.035	0.035
Yang-L1		0.906	0.312	0.130	0.102	0.071	0.058	0.049	0.039	0.044	0.026
SRC-OMP		0.829	0.091	0.070	0.049	0.036	0.031	0.020	0.027	0.030	0.026
Yang-COMP [†]		0.833	0.084	0.044	0.036	0.033	0.032	0.027	0.023	0.023	0.026
NCSC [‡]	$\kappa = 400$	0.846	0.109	0.049	0.038	0.042	0.042	0.026	0.026	0.019	0.029
	$\kappa = 200$	0.907	0.410	0.200	0.146	0.087	0.054	0.058	0.034	0.032	0.039
	$\kappa = 100$	0.927	0.590	0.371	0.256	0.144	0.091	0.081	0.050	0.048	0.028
	$\kappa = 50$	0.913	0.671	0.471	0.328	0.189	0.119	0.082	0.055	0.040	0.034
	$\kappa = 15$	0.869	0.730	0.575	0.425	0.259	0.165	0.103	0.051	0.058	0.032
	$\kappa = 10$	0.843	0.723	0.590	0.454	0.282	0.160	0.104	0.055	0.055	0.028
	$\kappa = 5$	0.749	0.665	0.586	0.434	0.269	0.170	0.112	0.061	0.044	0.027
	$\kappa = 2$	0.436	0.470	0.480	0.368	0.236	0.141	0.113	0.061	0.049	0.031

[†]Yang-COMP is equivalent to NCSC-CSR with $\kappa = 216$

[‡]The classifier with the highest classification accuracies in this table (when κ is tuned)

In Table IV, when $r = 0$ (for uncorrupted samples), the tuned κ is equal to 100. If r is not too large ($r \leq 0.7$), the tuned κ is a decreasing function of r . Thus, we contend that the sparsity imposed by a decreased κ compensates for the loss of classification accuracy when the image is corrupted. Also interesting is an observation that when $r \geq 0.8$, with the increase of r , the tuned κ increases from 2 to 50, toward 100 which is the tuned value of κ for uncorrupted samples. We contend this is primarily because the sparsity imposed by reducing κ has only a limited effect if the image is corrupted. Thus, when r is larger than a certain threshold, the tuned κ goes toward the value (i.e., 100, in this experiment) for uncorrupted data.

So far, with the experiments presented in Section V, we have demonstrated the possibility of a classifier NCSC-CSR, with high classification accuracy and robustness, via a constrained sparse representation with an intrinsic dimension parameter κ to be tuned.

VI. CONCLUSIONS AND FUTURE WORK

A model called constrained subspace is proposed for a classifier. In this model, the constrained subspace is defined as a union of a series of affine hulls. Each affine hull is spanned by training points via sparse representation and serves as a local linear approximation of the corresponding manifold.

This model establishes a generalized framework for some classical classifiers including NM (Nearest Manifold), NN (Nearest Neighbor), NFL (Nearest Feature Line) and NS (Nearest Subspace). In this model, NN, NFL and NS are all approximations to NM, which in principle enjoys a high classification accuracy. The constrained subspaces of NN and NFL have respectively intrinsic dimensions 0 and 1.

To make the constrained subspace a more accurate approximation to the corresponding manifold, we contend that the intrinsic dimension of the constrained subspace should be equal to that of manifold. Based on this assumption, we contend that the sparsity parameter κ of the sparse

representation, employed with the training samples to span the constrained subspace, should be carefully tuned. Thus, the searching of the nearest constrained subspace point to a query data point is formulated as a constrained least-squares problem. For the constrained least-squares problem, we propose a greedy optimizer called COMP (Constrained Orthogonal Matching Pursuit) and then a COMP-based classifier called NCSC-CSR (Nearest Constrained Subspace Classifier via Constrained Sparse Representation). NCSC-CSR can be viewed as a COMP-based variant to the state-of-the-art classifier SRC (Sparse Representation Classifier). NCSR-CSR seeks the projected point on a constrained subspace rather than an unconstrained subspace employed by SRC.

The proposed constrained subspace model connects the dots of NN, NFL, NS, SRC and NM and offers an intuitive explanation to the mechanism of the Yang's method (also a constrained variant to SRC) which reportedly outperforms SRC in terms of classification accuracy, although not always significantly. We demonstrate the COMP-based implementation of Yang's method is a special case of NCSC-CSR with a high dimensional constrained subspace.

Our experiments on the Extended Yale B database for supervised image classification show the possibility of a stronger classifier NCSC-CSR with a tuned intrinsic dimension parameter κ , which outperforms the relevant classifiers in terms of classification accuracy and robustness.

Since intrinsic dimension plays an important role in our constrained subspace model, we suggest some possible future work for the proper intrinsic dimension estimation from a limited number of samples.

Although there are some interesting estimators in this field, they are usually designed to offer a global estimation for all the points in the observed dataset rather than a local estimation for a single point in the dataset. The local estimators for a single point, especial for those points near to the borders of several classes, is still an open problem. For this reason, we are interested in local intrinsic dimension estimators on the points employed to span a linear manifold, which has the nearest point to a query data point. We believe these estimators can contribute to the design of a high-performance classifier via the proposed constrained subspace model.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. U1404607) and the Cultivation Program of the Backbone Teachers of Henan Colleges and Universities (No. 2014GGJS-91).

REFERENCES

- [1] S. Z. Li and A. K. Jain, "Handbook of Face Recognition," Springer, 2005, ch. Introduction, pp. 141–142.
- [2] G. Peyre, "Manifold models for signals and images," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 249–260, 2009.
- [3] 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.
- [4] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and

- illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.
- [5] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.
- [6] X. Huang, D. P. Dione, C. B. Compas, X. Papademetris, B. A. Lin, A. Bregasi, A. J. Sinusas, L. H. Staib, and J. S. Duncan, "Contour tracking in echocardiographic sequences via sparse representation and dictionary learning," *Medical image analysis*, vol. 18, no. 2, pp. 253–271, 2014.
- [7] L. Wang, F. Shi, Y. Gao, G. Li, J. H. Gilmore, W. Lin, and D. Shen, "Integration of sparse multi-modality representation and anatomical constraint for iso-intense infant brain mr image segmentation," *NeuroImage*, vol. 89, pp. 152–164, 2014.
- [8] J. Yang, L. Zhang, Y. Xu, and J.-y. Yang, "Beyond sparsity: The role of ℓ_1 -optimizer in pattern classification," *Pattern Recognition*, vol. 45, no. 3, pp. 1104–1118, 2012.
- [9] M. Elad, M. A. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," in *Proceedings of the IEEE*, vol. 98, IEEE, 2010, pp. 972–982.
- [10] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.
- [11] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp.553–560.
- [12] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp. 439–443, 1999.
- [13] R. Bennett, "The intrinsic dimensionality of signal collections," *IEEE Transactions on Information Theory*, vol. 15, no. 5, pp. 517–525, 1969.
- [14] L. Liao, Y. Zhang, S. J. Maybank, and Z. Liu, "Intrinsic dimension estimation via nearest constrained subspace classifier," *Pattern Recognition*, vol. 47, no. 3, p. 1485, 1485–1493 2014.
- [15] L. Liao, Y. Zhang, S. J. Maybank, Z. Liu, and X. Liu, "Image recognition via two-dimensional random projection and nearest constrained subspace," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1187–1198, 2014.
- [16] M. Fan, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of manifolds by incising balls," *Pattern Recognition*, vol. 42, no. 5, pp. 780–787, 2009.
- [17] A. massoud Farahmand, C. Szepesvári, and J. Audibert, "Manifold-adaptive dimension estimation," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 265–272.
- [18] F. Camastra and A. Vinciarelli, "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1404–1407, 2002.
- [19] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and I. Bottou, Eds., Cambridge, MA, December 2005, pp. 777–784.
- [20] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 471–478.
- [21] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.
- [22] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [23] UCSD, "Extended Yale face database B," <http://vision.ucsd.edu/extyaleb/CroppedYaleBZip/CroppedYale.zip>.