Connecting the Dots in Multi-Class Classification: From Nearest Subspace to Collaborative Representation

Yuejie Chi
Princeton University
ychi1@princeton.edu

Fatih Porikli
Mitsubishi Electric Research Lab
fatih@merl.com

Abstract

We present a novel multi-class classifier that strikes a balance between the nearest-subspace classifier, which assigns a test sample to the class that minimizes the distance between the test sample and its principal projection in the selected class, and a collaborative representation based classifier, which classifies a sample to the class that minimizes the distance between the collaborative components of the test sample by using all training samples from all classes as the dictionary and its projection in the selected class. In our formulation, the sparse representation based classifier [1] and nearest subspace classifier become special cases under different regularization parameters. We show that the classification performance can be improved by optimally tuning the regularization parameter, which can be done at almost no extra computational cost. We give extensive numerical examples for digit identification and face recognition with performance comparisons of different choices of collaborative representations, in particular when only a partial observation of the test sample is available via compressive sensing measurements.

1. Introduction

Multi-class classification, where the goal is to assign one of several class labels to a test sample, is an important problem encountered in many applications and has attracted significant research interests in decades. It is widely used for protein function identification, text classification, face recognition, etc.

With the recent advances in Compressive Sensing (CS) [2, 3], which aims to reconstruct an image from only a small number of linear measurements given it can be sparsely represented in a predefined basis or dictionary, such as wavelets or DCT, it is of increasing interests to develop multi-class classification algorithms that can achieve high classification accuracy without acquiring the full image. There is also a trend to explore sparsity in the feature domain to increase recognition performance for face recognition [1, 4] in particular. Assume that the test sample can be represented by the training samples of the same class, it then admits a Sparse Representation (SR) in the dictionary spanned by all training samples from all classes, where most nonzero components are expected to be found in the correct class. By reconstructing this SR using sparse recovery algorithms such as Basis Pursuit [5] or Orthogonal Matching Pursuit [6], and combining it into a Sparse Representation based Classifier (SRC), Wright et al. showed both accuracy and robustness can be greatly improved for face recognition. However, one main drawback of this approach is the complexity of acquiring SRs. Even using sparsity inducing ℓ_1 minimization, which is the convex approximation of solving the ℓ_0 NP-hard problem, the computational load is prohibitively high if the training set is large. Many works have been steered in this direction including using Gabor frame based SRs [7], using a learned dictionary instead of the whole training set for the dictionary [8], using random hashing to increase speed [9], etc.

Despite the initial success, several studies have raised the question as whether SRs are really necessary. In fact, the test sample has an infinite number of possible representations in the dictionary spanned by all training samples. They are referred to as Collaborative Representations (CRs), since all training samples collaboratively form a representation for the test sample, and SR is only one example. It is argued in [10, 4] that not the SR but the adoption of CRs in general is more crucial in the success of the SRC. For instance, using a different CR for the SRC, such as a regularized least-norm representation [4], similar performance can be achieved with much lower complexity.

In this paper, we effectively decompose the multi-class recognition problem (not restricting only to face recognition) into two parts, namely finding the CR and imposing it to a classifier that computes the residual toward each class in order to properly harness the CR of the test sample. Using the CR, the test sample is decomposed into a sum of components that each coming from a different subspace spanned by a separate class. We propose a novel multi-class clas-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

sifier, dubbed as Collaborative Representation Optimized Classifier (CROC), that achieves the most optimal combination of the Nearest-Subspace Classifier (NSC), which classifies a sample to the class with the minimal distance between the test sample and its principal projection, and the Collaborative Representation based Classifier (CRC), which assigns a sample to the class with the minimal distance between the collaborative components and the projection within the class. Moreover, we show that the well known SRC, the NSC, and the CRC become special cases under different regularization parameters. This enables us to further improve the classification performance by optimally tuning the regularization parameter, which is done at almost no extra computational cost. We also provide numerical examples to compare the classification performance for sparse and non-sparse CRs, and show in some cases the gain of using SRs can be achieved by using a non-sparse CR with an optimally tuned regularization parameter.

The paper is organized as follows. The multi-class classification problem is described in Section 2. The new collaborative representation optimized classifier is presented in Section 3. Numerical examples are given for digit recognition and face recognition in Section 4.

2. Multi-Class Classification

Assume there are K classes, where there are n_i training samples from the ith class stacked in a matrix as $\mathbf{A}_i = [\mathbf{a}_{i,1}, \cdots, \mathbf{a}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$, where $\mathbf{a}_{i,j} \in \mathbb{R}^m$ is the jth training sample of dimension m from the ith class. By concatenating all training samples we get the training dictionary $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_K] \in \mathbb{R}^{m \times n}$, where $n = \sum_{i=1}^{K} n_i$ is the total number of training samples. We are interested in classifying the test sample $\mathbf{y} \in \mathbb{R}^m$, given the labeled training samples in $\mathbf{A}$.

In this paper, the multi-class classification problem is explicitly decomposed into two parts, namely representing the test sample using the training dictionary, and inputting this CR to a classifier to estimate the label. We will discuss these two parts respectively below.

2.1. Representing Test Samples

We assume that samples within a class lie in the same low-dimensional linear subspace, for example, it is well-established that the face images of the same individual under various illuminations and expressions will approximately span a low-dimensional linear subspace in $\mathbb{R}^m$ [11, 12]. If the test sample $\mathbf{y}$ can be represented as a superposition of training samples in the dictionary $\mathbf{A}$, given a linear model as

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a collaborative representation of the test sample by exploring all training samples as a dictionary. When $\mathbf{A}$ is over-determined, i.e. the dimension of the samples is much larger than the number of training samples, the Least-Squares (LS) solution of (1) is given as

$$\mathbf{x}_{LS}^{FD} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 = \mathbf{A}^{\dagger}\mathbf{y}, \tag{2}$$

where $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, and FD refers to the fact that this CR is computed using the full test sample.

In many cases the LS solution (2) might lead to overfitting, therefore the test sample is mapped into a low-dimensional feature domain via dimensional reduction, examples including Eigenfaces [13], Fisherfaces [11], Randomface [1] for face recognition. Another important argument is motivated by the theory of CS, when it is impossible to acquire the full samples, only a partial observation is available via linear measurements and one is interested in classification upon the incomplete information. This can be viewed equivalently as linear feature extraction. In this paper, we focus on linear features, i.e. the extracted features can be written in terms of linear transformation:

$$\widetilde{\mathbf{y}} = \mathbf{R}\mathbf{y}; \qquad \widetilde{\mathbf{A}} = \mathbf{R}\mathbf{A}, \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{d \times m}$ is the linear transformation, and d is the feature dimension. For face recognition, both Eigenface and Randomface are linear features while Fisherface is not.

Now finding the CR of the test sample can be viewed as solution to the under-determined equation:

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{A}}\mathbf{x}. \tag{4}$$

Two popular choices for CRs are given below, where RD denotes the CR is computed using the extracted features (reduced dimensionality).

1) The sparse representation by minimizing the ℓ_1 norm of $\mathbf{x}$:

$$\mathbf{x}_{L1}^{RD} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \widetilde{\mathbf{y}} = \widetilde{\mathbf{A}}\mathbf{x}. \tag{5}$$

or the relaxed version

$$\mathbf{x}_{L1}^{RD} = \arg\min_{\mathbf{x}} \|\widetilde{\mathbf{y}} - \widetilde{\mathbf{A}}\mathbf{x}\|_2^2 + \epsilon\|\mathbf{x}\|_1. \tag{6}$$

The ℓ_1 constraint aims to use a minimal number of training samples, as it is beneficial in certain cases where most of the nonzero entries will come from the correct class, but the complexity is greatly increased.

2) The least-norm representation by minimizing the ℓ_2 norm of $\mathbf{x}$:

$$\mathbf{x}_{L2}^{RD} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_2 \quad \text{s.t.} \quad \widetilde{\mathbf{y}} = \widetilde{\mathbf{A}}\mathbf{x}, \tag{7}$$

which gives $\mathbf{x}_{L2}^{RD} = \widetilde{\mathbf{A}}^{\dagger}\widetilde{\mathbf{y}}$, where $\widetilde{\mathbf{A}}^{\dagger} = \widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T)^{-1}$; or the relaxed version

$$\mathbf{x}_{L2}^{RD} = \arg\min_{\mathbf{x}} \|\widetilde{\mathbf{y}} - \widetilde{\mathbf{A}}\mathbf{x}\|_2^2 + \epsilon\|\mathbf{x}\|_2^2, \tag{8}$$

which gives $\mathbf{x}_{L2}^{RD} = (\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} + \epsilon\mathbf{I})^{-1}\widetilde{\mathbf{A}}^T\widetilde{\mathbf{y}}$.

These solutions for the relaxed version can also be computed for the full test sample (1) without dimensionality reduction, given as $\mathbf{x}_{L1}^{FD}$ and $\mathbf{x}_{L2}^{FD}$.

Above definitions represent the test image $\mathbf{y}$ using all examples from all classes. Since different classes "collaborate" in the process of forming the representation, they are considered as "Collaborative Representations". In particular, $\mathbf{x}_{L1}^{RD}$, $\mathbf{x}_{LS}^{FD}$ and $\mathbf{x}_{L2}^{RD}$ are adopted respectively in [1], [10], and [4] for face recognition. However, the computational cost of $\mathbf{x}_{LS}^{FD}$ and $\mathbf{x}_{L2}^{RD}$ is much smaller than that of $\mathbf{x}_{L1}^{RD}$.

2.2. Sparse Representation based Classifier

We now examine the Sparse Representation based Classifier (SRC) that was first proposed in [1] and then quickly adopted in many follow-up work including [10, 4]. Although its original name indicates this method is reserved for SRs, in principle it can use any CR as an input. For consistency, we keep the sparse representation name here.

The SRC uses the CR of the test sample $\mathbf{y}$ as an input, denoted by $\mathbf{x} = [\mathbf{x}_1, \cdots, \mathbf{x}_K]$, where $\mathbf{x}_i$ is the part of coefficients corresponding to the ith class in $\mathbf{x}$. The test sample can be rewritten as a sum of components from different classes, namely

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{y}_i, \tag{9}$$

where the ith CR component is $\mathbf{y}_i = \mathbf{A}_i\mathbf{x}_i$, $1 \leq i \leq K$. The SRC will identify the test image with the ith class if the residual

$$r_i^{SR} = \|\mathbf{y} - \mathbf{y}_i\|_2^2 = \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i\|_2^2,\ 1 \leq i \leq K \tag{10}$$

is the smallest for the ith class.

In the supplementary material of [1] the authors discussed the benefits of the SRC from a sparse representation viewpoint. If the test image can be sparsely represented by all training images as $\mathbf{x} = [0, \cdots, \mathbf{x}_i, \cdots, 0]$, such that it can be represented by using only training samples within the correct class, given the abundance availability of training, then the residual for the correct class will be zero while the residual from other classes is the norm of the test image, resulting in maximal discriminative power for classification. In [4] the author shows that the SRC checks not only the angle between the test image and the the partial signal represented by the coefficient on the correct class (which should be small); but also the angle between the partial signal represented by the coefficient on the correct class and that on the rest classes (which should be large).

3. Regularizing the Classifier From NS to CR

In this section we will first dissect the NSC, which classifies a sample to the class with the minimal distance between the test sample and its principal projection. We then present a generic CRC, which classifies a sample to the class with the minimal distance between the sample reconstruction using the collaborative representation component and its projection within the class. Finally we give the new optimized classifier, which is a regularized path of classifiers that connects the NSC and the CRC, and the well-known SRC can be viewed as a particular dot on the path.

3.1. Nearest Subspace Classifier

Nearest Subspace Classifier (NSC) [14] assigns the test image $\mathbf{y}$ to the ith class if the distance, or the projection residual r_i^{NS} from $\mathbf{y}$ to the subspace spanned by the ith training set $\mathbf{A}_i = [\mathbf{a}_{i,1}, \cdots, \mathbf{a}_{i,n_i}]$ is the smallest among all classes, i.e.

$$i = \arg\min_i r_i^{NS}.$$

where r_i^{NS} is given as

$$\begin{aligned} r_i^{NS} &= \min_{\mathbf{x}_i} \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i\|_2^2 & (11)\\ &= \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i^{LS}\|_2^2 \\ &= \|(\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\dagger})\mathbf{y}\|_2^2. \quad i = 1, \ldots, K. & (12) \end{aligned}$$

where $\mathbf{x}_i^{LS} = \mathbf{A}_i^{\dagger}\mathbf{y}$.

Note that the above formulation of the NSC is used when the training samples per class is small so that they do span a subspace, which for face recognition is usually the case. When the number of training samples is large, such as in digit recognition, a principal subspace $\mathbf{B}_i$ for each class is extracted using Principle Component Analysis (PCA), then the projection residual $\tilde{r}_i^{NS}$ is computed as

$$\tilde{r}_i^{NS} = \min_{\mathbf{x}_i} \|\mathbf{y} - \mathbf{B}_i\mathbf{x}_i\|_2^2, \quad i = 1, \ldots, K. \tag{13}$$

The NSC does not require the CR of the test sample, and simply measures the similarity between the test sample and each class without considering the similarities between classes.

3.2. Collaborative Representation based Classifier

In our formulation, the Collaborative Representation based Classifier (CRC) assigns a test sample to the class with the minimal distance r_i^{CR} between the ith collaborative representation component, and its projection within that class, as

$$\begin{aligned} r_i^{CR} &= \|\mathbf{A}_i(\mathbf{x}_i - \mathbf{x}_i^{LS})\|_2^2 & (14)\\ &= \begin{cases} \|\mathbf{A}_i(\mathbf{x}_i - \mathbf{A}_i^{\dagger}\mathbf{y})\|_2^2 & \text{if } \mathbf{A}_i \text{ is over-determined.} \\ \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i\|_2^2 & \text{if } \mathbf{A}_i \text{ is under-determined.} \end{cases} \end{aligned}$$

Note that, the residual measures the difference between signal representation obtained from using only the intra-class information and the one using the inter-class information obtained from the collaborative representation. If the test sample can be sparse represented by all training samples, the residual for the correct class will be zero while the residual from other classes is the projection of the test sample, with similar discriminative power as the SRC in this scenario. The CRC is different from the SRC only when $\mathbf{A}_i$'s are over-determined, the test sample is replaced by its projection in each class.

3.3. Balancing Between NSC and CRC

Given the NSC and the CRC, which look at intra-class residual and inter-class residual respectively, we introduce the Collaborative Representation Optimized Classifier (CROC), which computes a regularized path to study the trade-off between these two classifiers, where the residual for each class is calculated as follows

$$r_i(\lambda) = r_i^{NS} + \lambda r_i^{CR}, \tag{15}$$

where $\lambda \geq 0$. The test sample is then assigned to the class that has the minimal residual. When $\lambda = 0$, it is equivalent to the NSC; and when $\lambda = +\infty$, it is equivalent to the CRC. In practice, cross-validation may be used to determine the optimal λ, as shown in the Section 4.1.

When $\mathbf{A}_i$ is over-complete and training is abundant, the NS residual error shall be computed using (13) and the CROC is equivalent to the SRC/CRC only when $\lambda = +\infty$. We will focus on the case when $\mathbf{A}_i$ is over-determined. First we show the SRC is equivalent to the CROC when $\lambda = 1$ using the same collaborative representation. The residual of each class for SRC (10) can be rewritten in the following way:

$$\begin{aligned} r_i^{SR} &= \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i\|_2^2 \\ &= \|\mathbf{y} - \mathbf{A}_i\mathbf{A}_i^{\dagger}\mathbf{y} + \mathbf{A}_i(\mathbf{A}_i^{\dagger}\mathbf{y} - \mathbf{x}_i)\|_2^2 \\ &= \|(\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\dagger})\mathbf{y}\|_2^2 + \|\mathbf{A}_i(\mathbf{A}_i^{\dagger}\mathbf{y} - \mathbf{x}_i)\|_2^2 && (16) \\ &= r_i^{NS} + r_i^{CR}, && (17) \end{aligned}$$

where (16) follows from

$$(\mathbf{I} - \mathbf{A}_i\mathbf{A}_i^{\dagger})\mathbf{A}_i = \mathbf{0}. \tag{18}$$

Alternatively, we can represent the CROC as a regularization between NSC and SRC:

$$r_i(\lambda) = (1-\lambda) r_i^{NS} + \lambda r_i^{SR} \tag{19}$$

Clearly, the widely used SRC only considers one possible trade-off between NSC and CRC by weighting two residual terms equally. As we will further show in the numerical examples, better regularization parameter exists to outperform the SRC regardless of the choice of collaborative representations for the test sample.

We could also rewrite the residual error for the CROC by plugging in (11) and (14), and apply (18), as

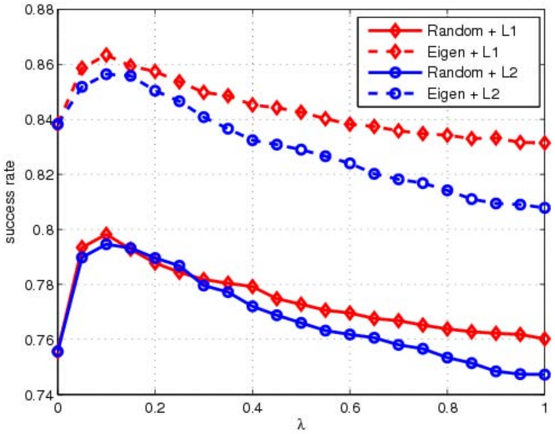


Figure 1. Classification results of CROC shown as a regularization path using partial measurements from random projection and eigen projections for the MNIST digits database.

$$\begin{aligned} r_i(\lambda) &= \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i^{LS}\|_2^2 + \lambda\|\mathbf{A}_i(\mathbf{x}_i^{LS} - \mathbf{x}_i)\|_2^2 && (20) \\ &= \|\mathbf{y} - \mathbf{A}_i\mathbf{x}_i^{LS} + \sqrt{\lambda}\mathbf{A}_i(\mathbf{x}_i^{LS} - \mathbf{x}_i)\|_2^2 && (21) \\ &= \|\mathbf{y} - \mathbf{A}_i[(1-\sqrt{\lambda})\mathbf{x}_i^{LS} + \sqrt{\lambda}\mathbf{x}_i]\|_2^2 \\ &= \|\mathbf{y} - \mathbf{A}_i\tilde{\mathbf{x}}_i\|_2^2 \end{aligned}$$

where $\tilde{\mathbf{x}}_i = (1-\sqrt{\lambda})\mathbf{x}_i^{LS} + \sqrt{\lambda}\mathbf{x}_i$. If we write

$$\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_K] = (1-\sqrt{\lambda})\mathbf{x}^{LS} + \sqrt{\lambda}\mathbf{x},$$

where $\mathbf{x}$ is the input CR, and $\mathbf{x}^{LS} = [\mathbf{x}_1^{LS}, \cdots, \mathbf{x}_K^{LS}]$ is "combined representation" by the least-square solution within each class, then $\tilde{\mathbf{x}}$ can be viewed as a different collaborative representation induced by $\mathbf{x}$ and the CROC will be equivalent to the SRC with a different collaborative representation as the input. However, $\tilde{\mathbf{x}}$ is not a solution to any of the optimization problem in Section 2.1, and this interpretation is only valid when $\mathbf{A}_i$'s are over-determined.

4. Numerical Results

In this section, we present numerical results on digit recognition and face recognition to show the classification accuracy gain by optimally choosing the regularization parameter. For digit recognition, the number of training images per class is very high, corresponding to the case $\mathbf{A}_i$ is under-determined; for face recognition, the number of training images per class is usually small, corresponding to the case $\mathbf{A}_i$ is over-determined.

4.1. Digit Recognition

The MNIST Handwritten Digits database [15] is used to test the proposed multi-class classification algorithm. There are about 6000 training examples and 1000 test examples of

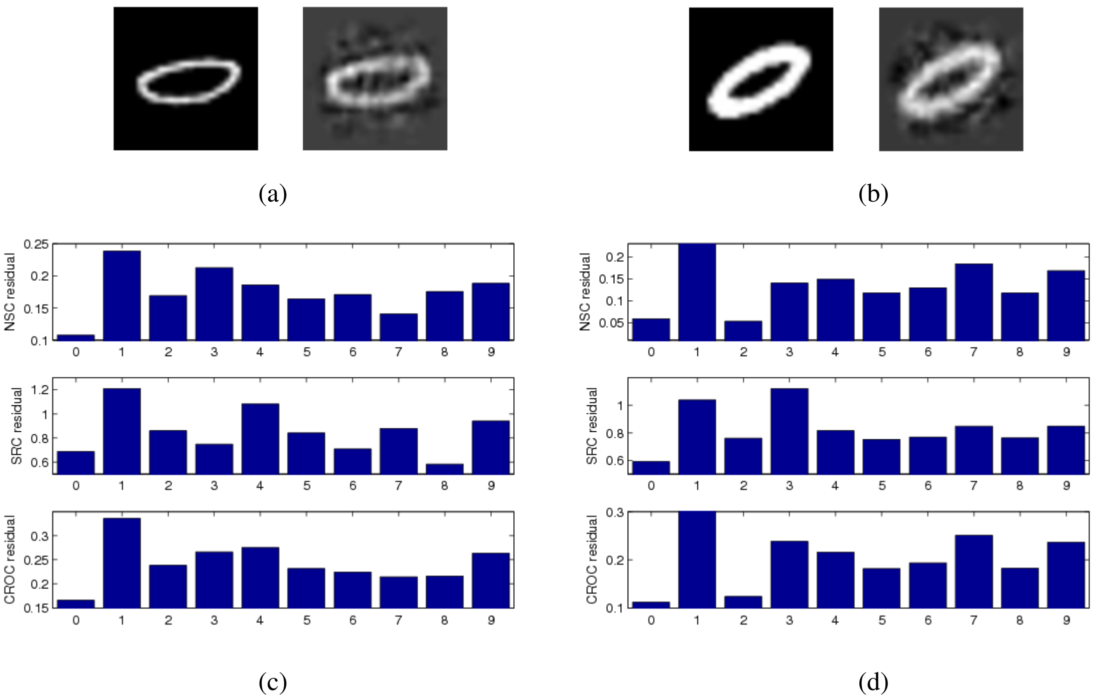


Figure 2. Classifier residual for two examples of digit “0”: (a) and (b) show the corresponding original digit and its reconstruction from the eigen projection of $k = 80$; (c) shows the classifier residual for digit in (a), which is correctly classified by the NSC, but misclassified as “8” by the SRC; (d) shows the classifier residual for digit in (b), which is correctly classified by the SRC, but misclassified as “2” by the NSC. Both are correctly classified by CROC with $\lambda = 0.1$.

each class in the data set. Each image is an 8-bit gray-scale image of “0” through“9” of dimension $d = 28 \times 28$.

We consider a toy example where only $n_i = 50$ training examples is provided per class, and the number of test examples per class is $n = 500$. We make $K = 80$ measurements of each test sample, and the whole test image is assumed unknown. We test the CROC against different regularization parameters, with $\lambda \in [0, 1]$. When $\lambda > 1$, the residual is mainly dominated by the CRC and the result is no longer interesting.

In the case where the full sample is not known, we could make partial observations using either random projections or projection along the eigenvector directions. Fig. 1 shows the classification accuracy for both scenarios using sparse (L1) and least-norm (L2) CRs. Projections using eigenvectors achieve better result than random projections in terms of accuracy. When $\lambda = 1$, the sparse CRs achieves slightly better result than the least-norm CRs using random projections, and this gain is even larger using eigen projections. However, a better classification can be achieved with λ around 0.1 for both CRs with very small performance gap between sparse and least-norm CRs. Table 1 further summarized the classification results for comparison. The optimal λ can be obtained by performing cross-validation on randomly selected training examples and testing examples for a few times, and compute the average classification accuracy for different λ and choose the optimal one. Fig. 3 shows the average classification accuracy over 5 times for the least-norm CR using eigen projections, showing the optimal $\lambda = 0.1$ in this case.

Scenario	NSC	SRC	CROC ($\lambda = 0.1$)
Random+L1[%]	75.56	76.02	**79.82**
Random+L2[%]	75.56	74.72	**79.46**
Eigen+L1[%]	83.82	83.14	**86.34**
Eigen+L2[%]	83.82	80.78	**85.64**

Table 1. Classification results of NSC, SRC and CROC using partial measurements from random projection and eigen projections.

Fig. 2 exemplifies how the CROC outperforms both the NSC and the SRC by using the least-norm CR. Each row shows the classifier residual using the NSC, the SRC and the CROC when $\lambda = 0.1$ respectively. For two test examples of digit “0”: in (a) it is correctly classified by the NSC, but the SRC misclassifies it as digit “8”; while in (b) it is correctly classified by the SRC, but the NSC misclassifies it as digit “2”. However, both can be correctly identified as “0” using a properly regularized CROC.

If we increase the number of training samples per class to $n_i = 500$, the training dictionary per class is now overcomplete and we will use a principal subspace $\mathbf{B}_i$ of dimen-

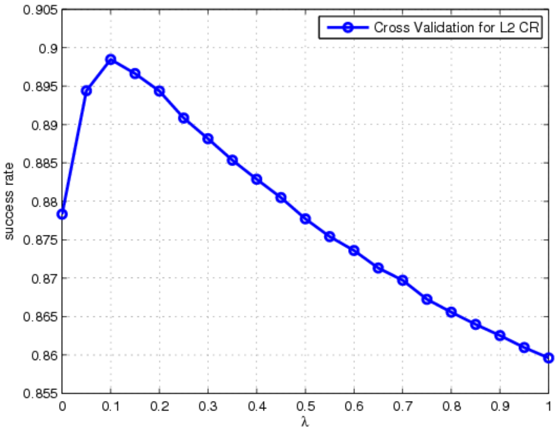


Figure 3. Cross-validation for choice of λ using least-norm CR from eigen projections for the MNIST digits database.

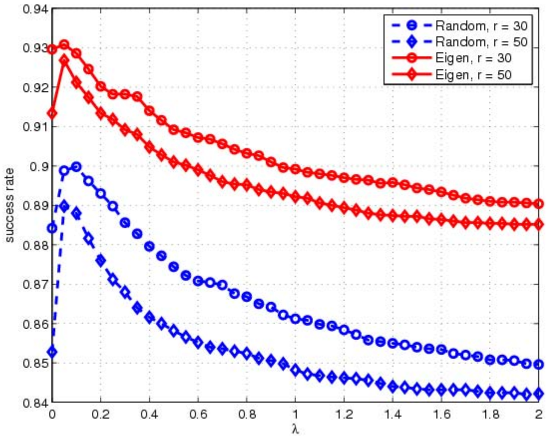


Figure 4. Classification results for the regularization path for different methods using partial measurements for the MNIST digits database.

sion r for the NSC. We use the least-norm CRs to redo the experiment for both random projection and eigen projection when $r = 30$ and $r = 50$. Notice that now the SRC is equivalent to $\lambda = +\infty$, however we only plot up to $\lambda = 2$ to show the trends, as shown in Fig. 4. We see there is a jump in the performance when λ is around 0.05; and adopting the sparse CR does not give particular gain compared with optimizing the regularization parameter λ.

4.2. Face Recognition

We test the proposed CROC against the Extended Yale-B database [16, 17] and the AR database [18]. Since our main goal is to show the benefit of the extra freedom by considering the regularization path, we do not test the robustness of face recognition with disguise (sunglasses, scarves, etc) in this work, yet such an extension is straightforward.

Scenario	Dim.	NSC	SRC	CROC (λ)
Full+LS	32256	97.46	**99.73**	**99.73** (0.8)
Full+L1	100	97.46	96.73	**97.82** (0.3)
	300	97.46	97.82	**98.28** (0.2)
Full+L2	100	97.46	91.20	**97.64** (0.2)
	300	97.46	97.82	**98.19** (0.2)
Reduced+L1	100	96.10	96.55	**97.19** (0.1)
	300	97.01	97.55	**98.19** (0.6)
Reduced+L2	100	96.10	89.11	**96.55** (0.1)
	300	97.01	97.19	**97.73** (0.2)

Table 2. Face recognition results for the NSC, CRC and CROC (with optimal λ): Full image with LS, L1 and L2 representations, partial images of various dimensions using Randomface with L1 and L2 representations for the Extended Yale-B database.

4.2.1 The Extended Yale-B Database

The Extended Yale-B database contains 2414 frontal-face images of 38 individuals [16]. We use the cropped and unnormalized face images of size 192×168 which are captured under different illuminations [17] for our experiments. For each individual, we randomly select $n_i = 30$ training samples and the rest are for testing. We consider random features of dimensions $d = 100$ and 300 and test the variations below depending on if the full test image is available:

- With the full image: three CRs, namely the least-squares representation $\mathbf{x}_{LS}^{FD}$ (2), the sparse representation $\mathbf{x}_{L1}^{FD}$ (5) and the least-norm representation $\mathbf{x}_{L2}^{FD}$ (7), are computed and tested (sparse representations are computed without dimensionality reduction by $\widetilde{\mathbf{A}}$).
- Without the full image: two CRs, namely the sparse representation $\mathbf{x}_{L1}^{RD}$ (5) and the least-norm representation $\mathbf{x}_{L2}^{RD}$ (7), are computed and tested.

In Fig. 5 it is obvious to see that when the full image is available, the $\mathbf{x}_{LS}^{FD}$ representation achieves the best classification accuracy with low complexity. When the full image is not available, for SRC corresponding to $\lambda = 1$, the sparse representation $\mathbf{x}_{L1}^{RD}$ achieves better accuracy than the least-norm representation $\mathbf{x}_{L2}^{RD}$ in terms of accuracy, in line with the previous work showing sparsity helps classification, in particular for smaller $d = 100$. However, this gain of using sparse representation [1] can be achieve by the least-norm representation with a properly tuned regularization parameter, at around $\lambda = 0.1$, at much lower computational cost. The classification accuracy for NSC, SRC and CROC with optimal λ are summarized in Table. 2.

4.2.2 The AR Database

Same as [1], we use a subset of 50 male subjects and 50 female subjects with only changes of illumination and expressions. For each subject, the seven images from Session

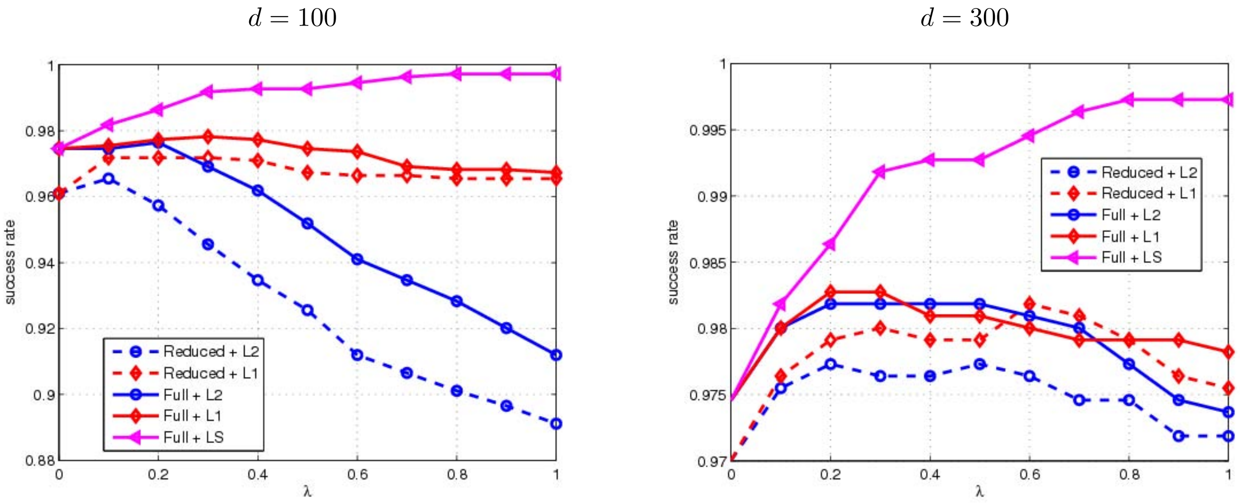


Figure 5. Face recognition results on the regularization path for different scenarios: using full image with three different CRs, Randomface of dimension $d = 100$ and $d = 300$ with ℓ_1 and ℓ_2 CRs for the Extended Yale-B database.

1 are used for training, and the other seven images from Session 2 are used for testing. The images are cropped to size 60×43.

Fig. 6 shows the regularization path of face recognition results for CROC of different scenarios: full image with least-squares CR $\mathbf{x}_{LS}^{FD}$, random projection of full image with least-norm CR, eigen-projection of full image with least-norm CR, and partial image using random pixel selection with least-norm CR at feature dimension $d = 100$ and $d = 300$. In the full image case, we show that better accuracy can be achieved at $\lambda = 0.3$, about 1.5% improvement than at $\lambda = 1$, corresponding the result in [10] using least-square CR. In almost all curves shown, some gain can be obtained by optimizing the regularization parameter λ. Fig. 7 shows two face examples and corresponding random pixel selection features: (a) face "1" is correctly classified by NSC, but misclassified as face "58" by SRC; (b) face "2" is correctly classified by SRC, but misclassified as face "25" by NSC. Both are correctly classified by CROC with $\lambda = 0.1$.

Fig. 8 compares classification result for NSC, SRC and CROC with optimal λ using random pixel selection (partial), Randomface and Eigenface and least-norm CR with different feature dimensions $d = 30, 50, 100, 300$. The gain of CROC with random pixel selection and Randomface is more significant than the gain with Eigenface.

5. Conclusions

In this paper we explicitly decompose the multi-class classification problem into two steps, namely finding the collaborative representation and inputting it to the multi-class classifier. We focus on the second step and propose a novel regularized collaborative representation based classifier where the NSC and the SRC are special cases on the whole regularization path. We show that classification performance can be further improved by optimally tuning the regularization parameter at no extra computational cost, in particular when only a partial test image is available via CS measurements. Numerical examples for digit recognition and face recognition demonstrate the benefit of our algorithm.

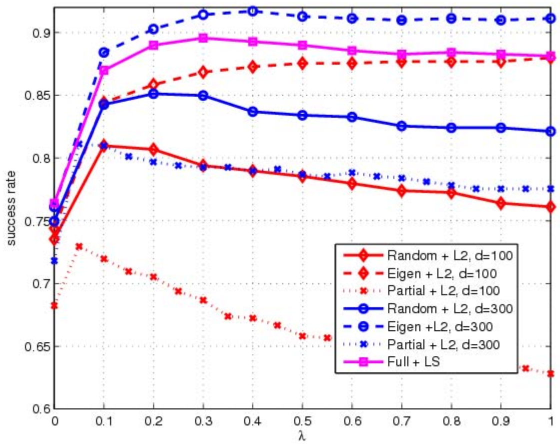
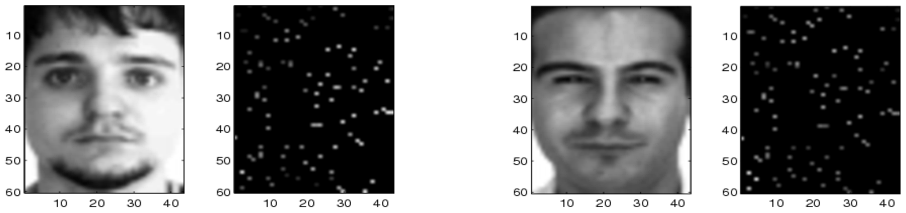


Figure 6. Face recognition results on the regularization path for different projection and CR combinations: Full image with LS representation, random pixel selection (partial), random projection and eigen-projection of full image with ℓ_2 CR for the AR database.

References

[1] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on*



(a)

(b)

Figure 7. Two face examples and corresponding random pixel selection features: (a) face “1” is correctly classified by NSC, but misclassified as face “58” by SRC; (b) face “2” is correctly classified by SRC, but misclassified as face “25” by NSC. Both are correctly classified by CROC with $\lambda = 0.1$.

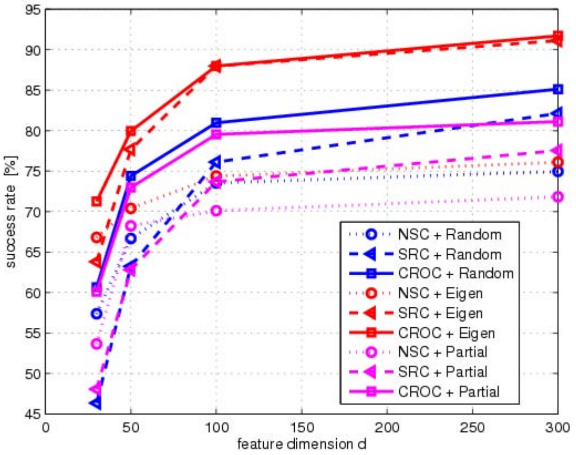


Figure 8. Face recognition results for NSC, SRC and CROC versus different feature dimensions with features of random pixel selection (partial), Randomface and Eigenface with least-norm CR for the AR database.

Pattern Analysis and Machine Intell., vol. 31, no. 2, 2009.

[2] D. Donoho, “For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution,” *Comm. Pure and Applied Math.*, vol. 59, no. 6, pp. 797–829, 2006.

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

[4] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: which helps face recognition?,” in *International Conference on Computer Vision (ICCV)*, 2011.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *East*, vol. 43, no. 1, pp. 129–159, 2001.

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

[7] M. Yang and L. Zhang, “Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary,” *European Conference on Computer Vision (ECCV)*, 2010.

[8] Q. Zhang and B. Li, “Discriminative K-SVD for dictionary learning in face recognition,” in *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] Q. Shi, H. Li, and C. Shen, “Rapid face recognition using hashing,” in *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] Q. Shi, A. Eriksson, A. Hengel, and C. Shen, “Is face recognition really a compressive sensing problem?,” in *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[11] P. Belhumeur, J. Hespanda, and D. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[12] R. Basri and D. Jacobs, “Lambertian reflectance and linear subspaces,” *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 25, no. 2, pp. 218–233, 2003.

[13] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

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

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

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

[18] A. Martinez and R. Benavente, “The AR face database,” *CVC Technical Report 24*, 1998.