Robust Face Recognition via Sparse Representation

John Wright, *Student Member, IEEE,* Arvind Ganesh, *Student Member, IEEE,*
Allen Yang, *Member, IEEE,* and Yi Ma, *Senior Member, IEEE*

John Wright, Arvind Ganesh, and Yi Ma are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. Allen Yang is with Electrical Engineering and Computer Science, University of California Berkeley. Corresponding author: Yi Ma, yima@uiuc.edu.

Abstract

In this paper, we consider the problem of automatically recognizing human faces from partially occluded frontal views. We argue that new mathematical theory from sparse signal representation offers the key to address this difficult problem. The desired representation is sparse, since the test face image should only be represented in terms of training face images of the same object. The occlusion of the face is also sparse, typically affecting only a fraction of the image pixels. We propose a simple, novel algorithm which uses ℓ^1 minimization to express the test image as a sparse linear combination of the training images plus a sparse error due to occlusion. This algorithm achieves state-of-the-art performance using raw imagery data, with no need for dimension reduction, feature selection, synthetic training examples or domain-specific information. We investigate the implications of this new mathematical framework for the engineering of recognition systems, showing how to predict how much occlusion the algorithm can handle and how to choose the training data to maximize robustness to occlusion. Extensive experiments on publicly available databases verify the efficacy of the proposed method.

Index Terms

Face Recognition, Occlusion, Random Corruption, Sparse Representation, ℓ^1-Minimization, Outlier Rejection.

I. Introduction

Occlusion poses a significant obstacle to robust, real-world object recognition [1], [2]. Consider, for example, a camera capturing an image of a human face, which we wish to recognize automatically. The image can be viewed as a measurement sampled from the fairly restricted set of possible images of the same face. Now suppose that the face is partially occluded or corrupted, as the examples shown in Figure 1. Occlusion corrupts the measured image, introducing errors that are

- large in magnitude in the value of pixels affected (*gross* errors),
- unpredictable in location of the pixels affected (*randomly supported* errors),
- concentrated only on relatively small portion of pixels of the image (*sparse* errors).

Ensuring robust recognition performance despite such errors incurred by occlusion is undoubtably a challenging task. Nevertheless, several aspects of the problem work in our favor.

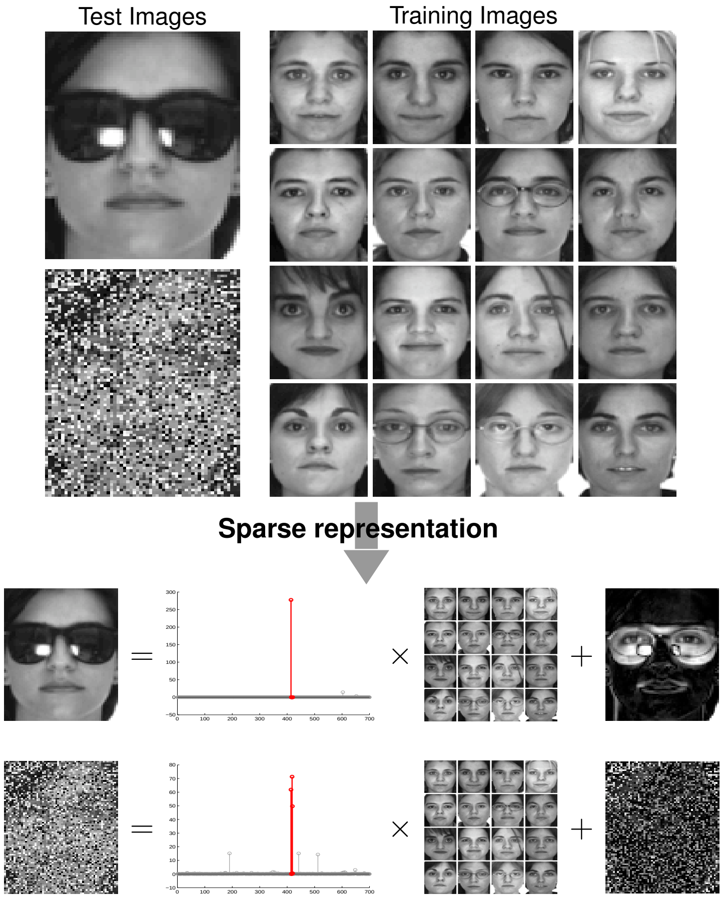


Fig. 1. **Top:** Two typical examples of test images to be considered in this paper: one is with contiguous occlusion and one with random corruption (50% pixels corrupted in the example shown). Both test images at left belong to one of the sixteen individuals on the right. Can you tell which one? **Bottom:** Our method represents a test image (left) as a *sparse* linear combination of the training images (middle) plus *sparse* errors (right) due to occlusion or corruption. Red (darker) coefficients correspond to training images of occluded individual. Algorithm 1 determines the true identity (second row, third from left) from amongst 100 individuals in the database.

a) Robustness from redundancy: A fundamental principle of coding theory [3] is that *redundancy* in the measurement is essential to detecting and correcting gross errors. Redundancy arises in object recognition because the number of image pixels is typically far greater than the number of subjects that have generated the images. In this case, even if a fraction of the pixels are completely corrupted by occlusion, recognition may still be possible based on the remaining pixels. On the other hand, schemes based on dimension reduction or feature extraction (*e.g.,* PCA [4], ICA [5], LDA [6]) discard redundant information that could compensate for the occlusion.

In this sense, no representation is more redundant, robust, or informative than the original image.

b) Robustness from locality: Of course, redundancy would be of no use without efficient computational tools for exploiting the information encoded in the redundant data. The difficulty of directly harnessing the redundancy of raw images has led researchers to instead focus on *spatial locality* as a guiding principle for robust recognition. Local features computed from only a small fraction of the image pixels are clearly less likely to be corrupted by occlusion than holistic features. In face recognition, methods such as ICA [5] and LNMF [7] exploit this observation by adaptively choosing filter bases that are locally concentrated. A related approach partitions the image into fixed regions and computes features for each region [8], [9]. Notice, though, that projecting onto locally concentrated bases transforms the domain of the occlusion problem, rather than eliminating the occlusion. Errors on the original pixels become errors in the transformed domain, and may even become less local. The role of feature extraction in achieving spatial locality is therefore questionable, since *no bases or features are more spatially localized than the original image pixels.* In fact, the most popular approach to improving the robustness of feature-based methods is based on randomly sampling individual pixels [10], sometimes in conjunction with statistical techniques such as multivariate trimming [11].

c) Robustness from sparsity: In this paper, we contend that the *sparsity* that arises in the context of recognizing partially occluded objects provides the key to achieving robust and accurate recognition. This sparsity arises from two sources: the identity of the test image and the nature of the occlusion. Ideally, the test image can be represented in terms of just the training images of the same object, a small portion of the entire training set. The corruption incurred by occlusion is also typically sparse, affecting only a relatively small fraction of the image pixels. To the best of our knowledge, there has been very little work in computer vision on predicting how much occlusion a robust recognition algorithm can handle before it starts to break down.

Sparse representations have attracted a great deal of attention in signal processing and information theory [12]–[15]. Recent progress has focused on the surprising effectiveness of the ℓ^1 norm for recovering sparse representations[1]. One significant implication is that under quite general conditions, the combinatorial problem of finding sparse solutions to systems of linear

[1] In face recognition, the ℓ^1 norm has been proposed as a distance measure for nearest neighbor classifiers (see, *e.g.,* [16]). This use of the ℓ^1 norm is not directly related to our work here, and hence it does not convey the same advantages in terms of sparsity or robustness.

equations can be efficiently and exactly solved via convex optimization, by minimizing the ℓ^1 norm [15].

Early work on sparse representation has been applied by [17] to detect translated face templates from a small library. Whereas our use of the ℓ^1 norm leads to tractable, convex problems, [17] utilizes the (non-convex) ℓ^p norm ($p < 1$) and so must resort to a greedy matching pursuit strategy. Sparsity induced by the identity of the test image is not identified as a critical factor in that work, due to the use of much smaller image libraries.

In this paper, we show for the first time how ℓ^1 minimization provides a principled framework for exploiting the two types of sparsity inherent in the robust recognition problem: sparsity in interpreting the test image and sparsity in the measurement error incurred by occlusion. We propose a simple, novel algorithm for recognition in the presence of occlusion. The algorithm uses ℓ^1 minimization to express the test image as a sparse linear combination of the given training images plus a sparse error due to occlusion. Directly exploiting the sparse structure of the problem enables our method to achieve performance exceeding the state of the art, using raw imagery data, with no need for dimension reduction, feature selection, synthetic training examples or domain-specific information (such as illumination models [18], [19]). We also investigate the implications of this framework for the engineering of recognition systems, showing how to predict how much occlusion the algorithm can handle and how to choose the training data to maximize robustness to occlusion. Extensive experiments on publicly available databases verify the efficacy of the proposed method.

d) What we do not do: While the proposed method is of broad interest to object recognition in general, the studies and experimental results in this paper are confined to human frontal face recognition. We will deal with illumination and expressions but we *do not* explicitly account for object pose, nor rely on any 3-D model of the face. The ℓ^1-minimization based framework *is* robust to small variations in pose and displacement, for example, due to registration errors. However, we do assume that detection, cropping, and normalization of the face have been performed prior to applying our algorithm.

e) Relation to the companion paper: In this paper, we will represent the test image as a sparse linear combination of training images of all the subjects in the database. This global approach has significant advantages over one-subject-at-a-time methods such as nearest neighbor and nearest subspace. However, the focus of this paper is on robustness and we will simply employ this new framework here. One can find a more thorough justification for this choice in the companion paper [20]. Also, one should be aware of the distinction between the random corruption considered in this paper and the random measurements studied in the companion paper: with random corruption, we do not know which measurements are valid.

II. Robust Recognition via Sparsity

A. Problem Formulation

The basic problem in object recognition is to use labeled training images from k distinct object classes to correctly determine the class of a test image whose identity is initially unknown. Throughout, we identify a $w \times h$ grayscale image with the vector $\boldsymbol{v} \in \mathbb{R}^m$ ($m = wh$) given by stacking its columns. We arrange the given n training images of as columns of a single matrix $A = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{m \times n}$, and let $A_i \in \mathbb{R}^{m \times n_i}$ denote the (submatrix of) n_i training images of the i-th subject. An immense variety of statistical, generative and discriminative models have been proposed for exploiting the structure of the A_i for recognition. One particularly simple and effective approach models the images from a single class as lying on a linear subspace. Subspace models are flexible enough to capture much of the variety in real datasets, and are especially well-motivated in the context of face recognition, where it has been observed that the images of faces under varying lighting conditions lie on a special 9-dimensional subspace [18] (see Figure 2(b)).

Thus, given sufficient training images, a test image $\boldsymbol{y}$ of the i-th object will approximately lie in the linear span of the training images from the same class: $\boldsymbol{y} = A_i \boldsymbol{x}_i$, where $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$ is a vector of coefficients. $\boldsymbol{y}$ can also be expressed in terms of the entire training set $A = [A_1 \ldots A_k]$:

$$\boldsymbol{y} = A\,\boldsymbol{x}_0, \tag{1}$$

where $\boldsymbol{x}_0 \doteq [\mathbf{0} \ldots \mathbf{0}\ \boldsymbol{x}_i^T\ \mathbf{0} \ldots \mathbf{0}]^T \in \mathbb{R}^n$ is a vector with zero entries everywhere except those associated with the corresponding subject i.[2]

[2] In the presence of noise, the linear subspace model does not hold exactly, and it may be more realistic to write $\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{z}$, where $\boldsymbol{z}$ is a vector of small-magnitude (*e.g.,* Gaussian) errors. For simplicity of the exposition, we will neglect the effect of $\boldsymbol{z}$. However, the geometry and algorithms described are provably stable under noise [14]. $\boldsymbol{z}$ can be explicitly accounted for by replacing the linear program in Algorithm 1 with a second-order cone program [13].

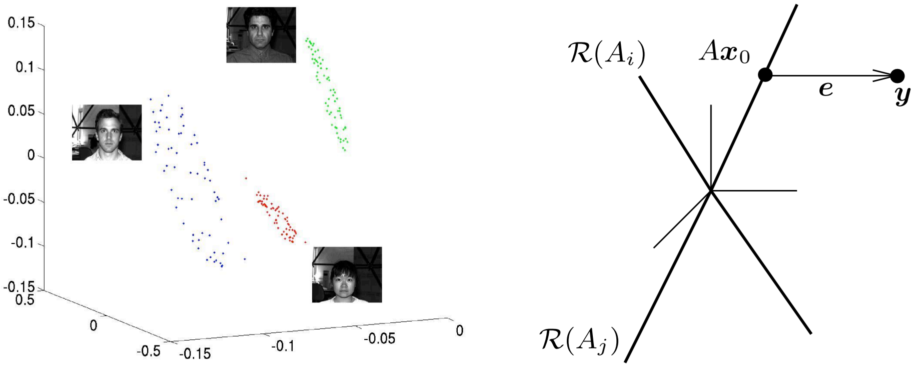


Fig. 2. **Our model.** Left: First three principal components of images of three individuals in the Yale Face Database B. The distribution suggests that training images A_i of each individual approximately lie on a subspace, denoted as $\mathcal{R}(A_i)$. Right: An occluded test image $\boldsymbol{y}$ is generated by sampling a point $A\boldsymbol{x}_0$ from one of the subspaces, say $\mathcal{R}(A_j)$, and perturbing it by an error vector, $\boldsymbol{e}$.

Now suppose that the observed image $\boldsymbol{y}$ is also partially occluded. Let ρ be the fraction of pixels in the occluded region. Then a fraction $(1-\rho)$ of the pixels belong to an image $\boldsymbol{y}_0$ from one of the K object classes:

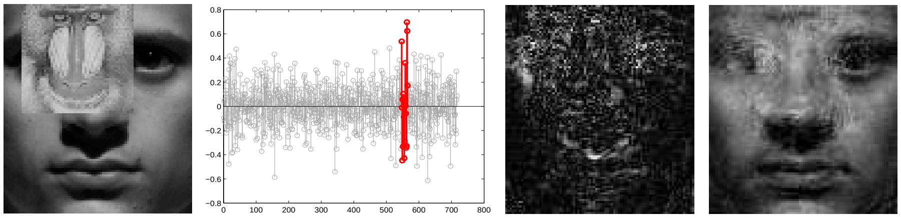
$$\boldsymbol{y} \;=\; \boldsymbol{y}_0 + \boldsymbol{e} \;=\; A\,\boldsymbol{x}_0 + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{e} \in \mathbb{R}^m$ is a vector of errors – a fraction ρ of its entries are nonzero. The errors may have arbitrary magnitude and therefore cannot be ignored or treated with techniques designed for small-magnitude noise. Notice, however, that the true $\boldsymbol{e}$ is *sparse*; its non-zero entries lie on a set of size ρm corresponding to the occluded pixels (Figure 2).

Given this formulation, consider the following problem:

> *Given labeled sets of training images $A_1, \ldots, A_k$ from k classes and a test image $\boldsymbol{y}$ generated by sampling an image from the ith class and then perturbing the values of a fraction ρ of its pixels arbitrarily, identify the correct class i.*

Notice that in the problem statement, we did not make any specific assumption on the location or value of the perturbation: the corrupted pixels could be a contiguous region or totally random in location; the change in pixel values does not have to obey any prior probabilistic distribution. In a sense, we are looking for a solution that works even when the corruption occurs at the



(a) (b) (c) (d)

Fig. 3. **Brittleness of the ℓ^2 minimizer.** (a) occluded test image, $\boldsymbol{y}$. (b) coefficients against the training set, $\hat{\boldsymbol{x}}_2$, estimated by minimizing the ℓ^2 norm. (c) error, $\hat{e}_2$. (d) reconstruction from the estimated coefficients. The estimated coefficient vector is *dense*, and the reconstruction is significantly corrupted.

worst possible locations and with the worst possible change of pixel values.[3]

As a first step toward solving this problem, notice that due to its special, sparse structure the vector $\boldsymbol{x}_0$ is extremely informative for classification. Since $\boldsymbol{x}_0$ is initially unknown, we seek (approximate) solutions to the system of linear equations, $A\boldsymbol{x} = \boldsymbol{y}$. Solving a system of linear equations involving the entire training set represents a significant departure from “one-subject-at-a-time” methods such as nearest neighbor and subspace methods. In related work [20], we argue that such a global representation is naturally discriminative, and demonstrate its superiority over local methods for identifying subjects represented in the training set and rejecting outlying images of subjects not included in the training set.

When the number of image pixels, m, is greater than the number of training images, n, the system is overdetermined, and may not have an exact solution. In this case, conventional wisdom[4] overwhelmingly favors a solution in the least-squares sense, by minimizing the ℓ^2 norm of the residual:

$$\hat{\boldsymbol{x}}_2 \doteq \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - A\boldsymbol{x}\|_2. \tag{3}$$

[3] The rationale for considering the worst-case scenario is to have a guaranteed performance for the resulting algorithm, even against the worst possible malicious disguise or corruption of the face image. However, towards the end of this paper, we will show how to incorporate additional knowledge of the occlusion to further enhance the performance of the algorithm under specific conditions.

[4] including common practice in computer vision. Popular techniques (*e.g.,* [4] amongst others) represent $\boldsymbol{y}$ as a linear superposition of basis vectors whose coefficients are computed by minimizing the ℓ^2 norm. For orthonormal bases, these coefficients are just the projection of $\boldsymbol{y}$ onto the basis vectors.

In the presence of isotropic Gaussian noise, $\hat{\boldsymbol{x}}_2$ is the maximum likelihood estimate of $\boldsymbol{x}$. However, the error $\boldsymbol{e}$ introduced by occlusion is highly non-Gaussian, consisting of gross errors concentrated on a subset of the image pixels. In this situation, $\hat{\boldsymbol{x}}_2$ can be arbitrarily bad: since $\|\boldsymbol{e}\|_2$ is unbounded, $\|\hat{\boldsymbol{x}}_2 - \boldsymbol{x}_0\|_2$ is also unbounded. Figure 3 gives a visual demonstration of this phenomenon. Notice that both the coefficients $\hat{\boldsymbol{x}}_2$ and the error $\hat{\mathbf{e}}_2$ estimated by ℓ^2 minimization are *densely supported*. Consequently, the reconstruction in Figure 3 (d) is quite poor, comparing to our results in Figure 7. In the remainder of this section, we will show how to exploit sparsity by minimizing the ℓ^1 norm, rather than the ℓ^2 norm, leading to a simple, robust and efficient algorithm for recognition in the presence of occlusion.

B. Imposing Sparsity of Occlusion

Since the error $\boldsymbol{e}$ is known to be sparse, but of arbitrary magnitude, a reasonable alternative to minimizing the ℓ^2 norm of the residual is to instead seek the $\boldsymbol{x}$ which gives the sparsest residual:

$$\hat{\boldsymbol{x}}_0 = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - A\boldsymbol{x}\|_0. \tag{4}$$

Here, the ℓ^0 "norm" $\|\boldsymbol{x}\|_0$ counts the number of nonzero entries[5] of the vector $\boldsymbol{x}$. Computing (4) gives the vector $A\hat{\boldsymbol{x}}_0$ in the range of A such that the error, $\boldsymbol{e} = \boldsymbol{y} - A\hat{\boldsymbol{x}}_0$, has the fewest nonzero entries.

In fact, if A is in general position and the error $\boldsymbol{e}$ has support less than $m/2$ (less than half the image is occluded), then $\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$. That is, the true solution $\boldsymbol{x}_0$ gives the sparsest residual and can be found exactly by solving (4). However, in the general case, problem (4) is NP-hard and also difficult to approximate [21]. Unless P = NP, there is no procedure significantly more efficient than exhaustive search over all supports of $\boldsymbol{e}$. It may therefore seem that computing the true $\boldsymbol{x}_0$ is prohibitively costly. Fortunately, and somewhat surprisingly, this is not the case for our problem of interest. Recently, a series of papers [12]–[15] have shown that if the error $\boldsymbol{e}$ is *sufficiently sparse*, then the ℓ^0 minimizer $\hat{\boldsymbol{x}}_0$ is equal to the ℓ^1 minimizer:

$$\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_1, \tag{5}$$

where $\hat{\boldsymbol{x}}_1 \doteq \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - A\boldsymbol{x}\|_1$. This is a convex optimization problem, whose solution is unique and can be efficiently computed by linear programming.

[5] $\|\boldsymbol{x}\|_0$ is not a true norm, since $\|\alpha\boldsymbol{x}\|_0 = \|\boldsymbol{x}\|_0$ for $\alpha \neq 0$.

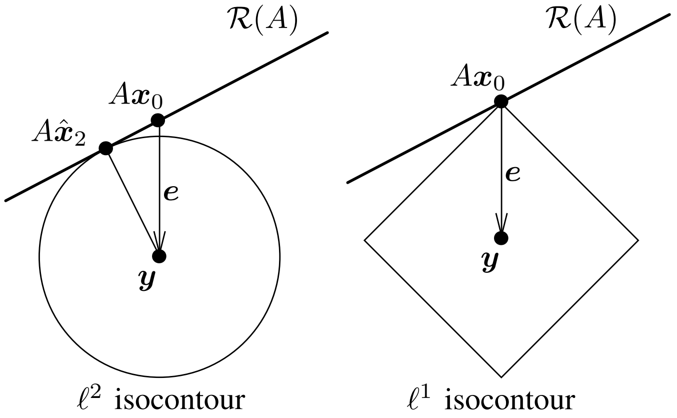


Fig. 4. **Robustness of the ℓ^1 minimizer to sparse errors.** Left: the ℓ^2 minimizer is obtained by placing a sphere (ℓ^2 ball) centered at $\boldsymbol{y}$ and expanding it to touch $\mathcal{R}(A)$. Right: the ℓ^1 minimizer is obtained by placing an ℓ^1 ball (a polytope) centered at $\boldsymbol{y}$ and expanding it to touch $\mathcal{R}(A)$.

Thus, if $\boldsymbol{y} = A\boldsymbol{x}_0 + \boldsymbol{e}$ and the error $\boldsymbol{e}$ is sufficiently sparse, then the generating coefficients, $\boldsymbol{x}_0$ can be recovered efficiently and exactly by minimizing the ℓ^1 norm of the residual. A proof of this equivalence and a detailed exposition of the technical conditions under which it holds are beyond the scope of this paper. However, Figure 4 gives the basic intuition for why the ℓ^1 minimizer conveys such dramatic advantages in terms of robustness. In Figure 4, the observation $\boldsymbol{y}$ is generated by sampling a point $A\boldsymbol{x}_0$ in the range of A and then perturbing it by $\boldsymbol{e}$ along one of the coordinate axes. Minimizing the ℓ^2 norm is equivalent to finding the smallest radius sphere around $\boldsymbol{y}$ which touches the range, $\mathcal{R}(A)$, of A. The point where it touches is $A\hat{\boldsymbol{x}}_2$, and is not equal to $A\boldsymbol{x}_0$ unless $\boldsymbol{e} \perp \mathcal{R}(A)$. On the other hand, the level sets of $\|\cdot\|_1$ are octahedral, aligned with the coordinate axes. Minimizing the ℓ^1 norm is equivalent to centering an ℓ^1 ball (octahedron) about $\boldsymbol{y}$, and expanding it until it touches $\mathcal{R}(A)$. This point of intersection is $A\hat{\boldsymbol{x}}_1$, which in this case is equal to $A\boldsymbol{x}_0$. Notice that the right picture in Figure 4 is scale invariant. This implies that the ability of ℓ^1 minimization to recover the true solution $\boldsymbol{x}_0$ is independent of the magnitude of the $\boldsymbol{e}$, and depends only on the signs of $\boldsymbol{e}$ and the relative geometry of the subspace $\mathcal{R}(A)$ and the unit ℓ^1 ball.

C. Imposing Simultaneous Sparsity of Identity and Occlusion

The previous section showed how, in overdetermined systems ($m \gg n$), minimizing the ℓ^1 norm of the residual could recover $\boldsymbol{x}_0$, even in the presence of arbitrarily large errors, provided those errors are sparsely supported. This approach does not explicitly enforce the sparseness of the estimated coefficient vector $\hat{\boldsymbol{x}}_1$. As the number of training images n increases, the equivalence $\hat{\boldsymbol{x}}_1 = \hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$ begins to break down. The reason for the this breakdown is clear from examining the linear system $A\boldsymbol{x} = \boldsymbol{y}$. As n increases, this system becomes square, and then underdetermined. $\mathcal{R}(A)$ will then span all of $\mathbb{R}^m$, and even the occluded test image $\boldsymbol{x}$ will be expressible as a linear combination of the columns of A.

From an error correction perspective, increasing n (*e.g.,* by expanding the number of object classes K in the database) seems to decrease the *redundancy* of the representation: the same number of image pixels must express a greater number of degrees of freedom. Notice, however, that this is not the case. Each test image still has an expression, $\boldsymbol{y}_0 = A\boldsymbol{x}_0$, which is highly redundant – it depends only on a few ($\leq n_i$) nonzero entries of $\boldsymbol{x}_0$. However, in order to exploit the robustness inherent in such a redundant representation, we need to enforce that the estimated coefficients $\hat{\boldsymbol{x}}$ have such sparse structure.

We therefore seek a *simultaneously sparse solution* for $\boldsymbol{x}$ and $\boldsymbol{e}$. Rewriting (2),

$$\boldsymbol{y} = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{e} \end{bmatrix} \doteq B\,\boldsymbol{w}. \tag{6}$$

Here, $B = [A\ I] \in \mathbb{R}^{m\times(n+m)}$, so the system $B\boldsymbol{w} = \boldsymbol{y}$ is underdetermined and does not have a unique solution. However, from the above discussion, the generating $\boldsymbol{w}$ has at most $n_i + \rho m$ nonzeros. This motivates us to seek the sparsest solution to this system of equations:

$$\hat{\boldsymbol{w}}_0 = \arg\min \|\boldsymbol{w}\|_0 \ \text{ subject to } \ B\boldsymbol{w} = \boldsymbol{y}. \tag{7}$$

In fact, if the matrix B is in general position, then as long as $\boldsymbol{y} = B\tilde{\boldsymbol{w}}$ for some $\tilde{\boldsymbol{w}}$ with less than $m/2$ nonzeros, $\tilde{\boldsymbol{w}}$ is the unique sparsest solution: $\hat{\boldsymbol{w}}_0 = \tilde{\boldsymbol{w}}$. Thus, if the occlusion $\boldsymbol{e}$ covers less than $\frac{m-n_i}{2}$ pixels, $\approx 50\%$ of the image, the solution to (7) is the true generator, $[\boldsymbol{x}_0\ \boldsymbol{e}]$.

Like the overdetermined problem (4), finding the sparsest solution to a system of linear equations is NP-hard in general. In fact, the problem of solving overdetermined systems $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ with sparse errors and the problem of solving underdetermined systems $\boldsymbol{y} = B\boldsymbol{w}$ with a spare solution are equivalent, and can be converted from one to the other [13]. A corresponding equivalence result states that if the sparsest solution $\boldsymbol{w}_0$ is sufficiently sparse, then it is equal to the minimum ℓ^1 norm solution [15],

$$\hat{\boldsymbol{w}}_1 = \arg\min \|\boldsymbol{w}\|_1 \ \text{ subject to } \ B\boldsymbol{w} = \boldsymbol{y}. \tag{8}$$

This implies that as long as the occlusion $\boldsymbol{e}$ and the coefficients $\boldsymbol{x}_0$ together are sufficiently sparse, they can be efficiently and exactly computed by ℓ^1 minimization.

D. When Does the ℓ^1-ℓ^0 Equivalence Hold?

Thus far, we have sketched several results in the theory of ℓ^1-ℓ^0 equivalence, and shown how these provide an efficient and tractable means of exploiting the two types of sparsity inherent in the robust recognition problem. However, determining whether these results are practically relevant requires a more precise notion of when $\boldsymbol{w}_0$ is "sufficiently sparse."

In an effort to quantify when ℓ^1-ℓ^0 equivalence holds (*i.e.,* when $\hat{\boldsymbol{w}}_1 = \boldsymbol{w}_0$), Donoho [15] defines the *equivalence breakdown point* (EBP) of a matrix B as the maximum number, k, such that if $\boldsymbol{y} = B\boldsymbol{w}_0$ for some $\boldsymbol{w}_0$ with less than k nonzero entries, then the minimal ℓ^1 norm solution $\hat{\boldsymbol{w}}_1$ to the system $B\boldsymbol{w} = \boldsymbol{y}$ is equal to that sparse generator $\boldsymbol{w}_0$. A number of sufficient conditions have been given in the literature for the existence of a constant, ρ_0, such that EBP$(B) > \rho_0 m$ (i.e. for solutions with a non-vanishing fraction of nonzeros to be recoverable by ℓ^1 minimization). For example, [12] shows that even for a random matrix B drawn from a Gaussian ensemble, EBP$(B) > \rho_0 m$ with overwhelming probability, as $m \to \infty$. An important upper bound on EBP(B) comes from the theory of centrally neighborly polytopes [22]:

$$\text{EBP}(B) \leq \lfloor (m+1)/3 \rfloor . \tag{9}$$

This result indicates that we should not expect to perfectly recover $[\boldsymbol{x}_0 \; \boldsymbol{e}]$ if $n_i + |\text{support}(\boldsymbol{e})| > m/3$. Generally, $m \gg n_i$, so (9) implies that the largest fraction of occlusion under which we can hope to still achieve perfect reconstruction is 33%. This bound is corroborated by our experimental results; see Figure 7.

In our context, we often need to have more accurate information about the breakdown point than a loose upper bound. For instance, we would like to know for a given set of training images, what is the largest amount of occlusion it can handle. While the best known algorithms for exactly computing EBP$(\cdot)$ are combinatorial in nature (see [23] for details), tight upper bounds can be obtained by restricting the search for intersections between $\mathcal{R}(B^{\perp})$ and the ℓ^1 ball to a random subset of the k-skeletons of the ℓ^1 ball. This will be the technique that we use to estimate all the EBPs of the training datasets considered.

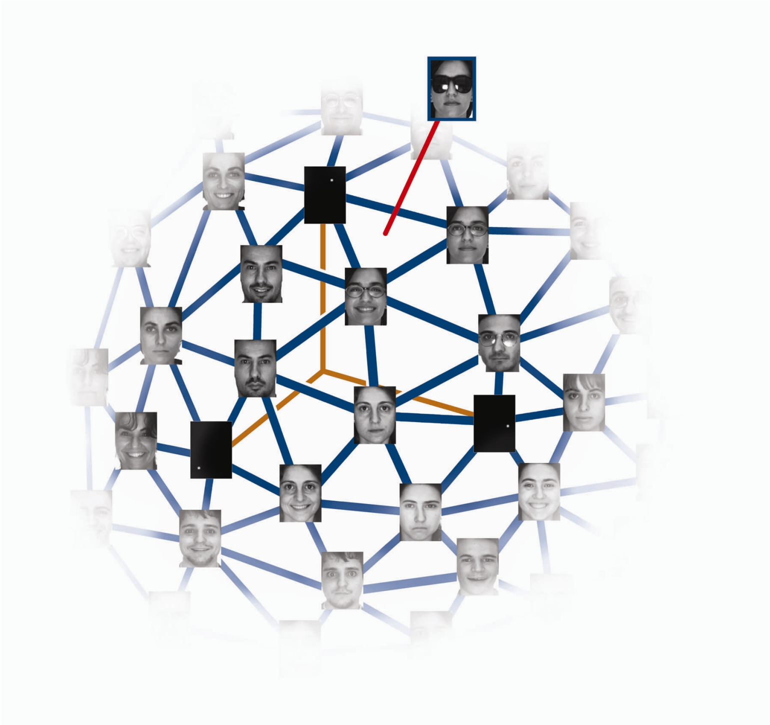


Fig. 5. **Geometry of Classification via ℓ^1 Minimization.** The convex hull of the columns of $B = [A\, I]$ is a high-dimensional polytope in $\mathbb{R}^m$. Each vertex of this polytope is either a training image or an image with just a single pixel illuminated (corresponding to the identity submatrix of B). Given a test image, solving the ℓ^1 minimization problem essentially locates which facet of the polytope the test image falls on. The algorithm finds the facet with the fewest possible vertices. Only vertices of that facet contribute to the representation; all other vertices have no contribution.

E. Face Recognition Algorithm

Suppose now that we are given a test image $\boldsymbol{y}$ generated according to (6), and further suppose that support$(\hat{\boldsymbol{w}}_0) <$ EBP(B), so that (7) and (8) are equivalent. Then, by minimizing $\|\boldsymbol{w}\|_1$ subject to $[A\ \ I]\boldsymbol{w} = \boldsymbol{y}$, we can recover the coefficient vector $\boldsymbol{x}_0$ and the error $\boldsymbol{e}$ induced by occlusion. Figure 5 gives a geometric interpretation of this recovery, in the context of face recognition with occlusion. The training images and the standard basis of single-pixel images together span a convex polytope in $\mathbb{R}^m$. Solving the ℓ^1 minimization problem determines which facet of this polytope the test image lies on. All training images not incident on this facet do not contribute to the representation; their coefficients in $\hat{\boldsymbol{x}}_1$ are zero.

There are many potential ways that the estimates, $\hat{\boldsymbol{x}}_1$ and $\hat{\boldsymbol{e}}_1$ can be used for classification. For example, setting $\boldsymbol{y}_r \doteq \boldsymbol{y} - \hat{\boldsymbol{e}}_1$ compensates for the effect of occlusion. One could then classify the reconstructed image $\boldsymbol{y}_r$ based on which of the face subspaces $\mathcal{R}(A_1), \ldots, \mathcal{R}(A_k)$ it is closest to.

Notice, however, that the sparse coefficients, $\hat{\boldsymbol{x}}_1$ *already encode the identity of the test subject.* We therefore directly use the $\hat{\boldsymbol{x}}_1$, assigning $\boldsymbol{y}$ to the class whose coefficients best approximate it. More precisely, we define k functions $\delta_i : \mathbb{R}^n \to \mathbb{R}^n$ the i-th of which preserves the coefficients corresponding to the i-th group and sets the rest to zero: $\delta_i(\boldsymbol{x}) = [\boldsymbol{0} \dots \boldsymbol{0}\, \boldsymbol{x}_i^T\, \boldsymbol{0} \dots \boldsymbol{0}]^T \in \mathbb{R}^n$. The approximation in terms of the coefficients associated with the i-th group is then $\hat{\boldsymbol{y}} = A\delta_i(\hat{\boldsymbol{x}}_1) + \hat{\boldsymbol{e}}_1$, and classification can be achieved by assigning $\boldsymbol{y}$ to the group that minimizes $\|\boldsymbol{y} - A\delta_i(\hat{\boldsymbol{x}}_1) - \hat{\boldsymbol{e}}_1\|_2$. The entire process is summarized as Algorithm 1. Our implementation minimizes the ℓ^1 norm via a primal-dual algorithm for linear programming based on [24], [25].

Algorithm 1 (Robust Recognition via Sparse Representation).

```
Input: $n$ training samples partitioned into $k$ classes, $A_1, \dots, A_k$ and a test sample $\boldsymbol{y}$.
Normalize the training samples to have unit $\ell^2$ norm and set $B = [A_1 \dots A_k \; I]$.
Solve the $\ell^1$ minimization problem:
    $$\hat{\boldsymbol{w}}_1 = \arg \min_{\boldsymbol{w}=[\boldsymbol{x}\, \boldsymbol{e}]} \|\boldsymbol{w}\|_1 \text{ s.t. } B\boldsymbol{w} = \boldsymbol{y}, \qquad (10)$$
    by linear programming.
Compute the residuals $r_i(\boldsymbol{y}) = \|\boldsymbol{y} - A\,\delta_i(\hat{\boldsymbol{x}}_1) - \hat{\boldsymbol{e}}_1\|_2$, for $i = 1, \dots, k$.
Output: $\mathrm{id}(\boldsymbol{y}) = \arg\min_{i=1,\dots,k} r_i(\boldsymbol{y})$.
```

In the above algorithm, the equation $\boldsymbol{y} = B\boldsymbol{w}$ is considered as a hard constraint in the optimization. Although in reality the equality would never be satisfied exactly, most numerical implementations of linear program is stable and can tolerate small amount of error in the constraints. For all the experiments in this paper, we find this already gives sufficiently good performance. If the user has good reason to believe that the test image has certain amount of noise, one can revise the above optimization problem as $\min_{\boldsymbol{w}=[\boldsymbol{x}\,\boldsymbol{e}]} \|\boldsymbol{w}\|_1$ s.t. $\|B\boldsymbol{w} - \boldsymbol{y}\|_2 \leq \epsilon$ so as to tolerate up to the error ϵ. This optimization problem is convex and can still be solved very efficiently.[6] The solution gives a *provably stable* estimate of the desired sparse solution [14].

As discussed above, Algorithm 1 perfectly compensates for arbitrary occlusions covering upto $(\mathrm{EBP}(B) - n_i)$ pixels. Beyond this range, theory no longer guarantees exact recovery with the

[6] This is the method of choice in the companion paper [20] in which we deal with images or features of much lower dimension, where the signal-to-noise ratio may become considerably lower.

worst possible occlusion, and so recognition performance may start to suffer. Nevertheless, a body of experience suggests that minimizing the ℓ^1 norm still encourages sparsity far beyond the breakdown point. The worse-case analysis may be too conservative for the average performance with random corruption and occlusion. In the next section, we will see that this is indeed the case.

III. Experiments

A. Recognition despite Random Pixel Corruption

For this experiment, we use the Extended Yale B Face Database [26] (cropped and normalized [19]). This dataset contains frontal images of 38 subjects under various illumination conditions. We choose Subsets 1 and 2 (717 images, normal-to-moderate lighting conditions) for training, and Subset 3 (453 images, more extreme lighting conditions) for testing. Without occlusion, this is a relatively easy recognition problem. This choice is deliberate, in order to isolate the effect of occlusion. The images are resized to 96×84 pixels,[7] so in this case B is an $8,064\times 8,761$ matrix. Each of the training images $\boldsymbol{v}_i$ is scaled to have unit ℓ^2 norm. For this dataset, we have estimated EBP$(B)\approx 1,185$ (using the method given in [23]), suggesting that perfect reconstruction can be achieved upto 13.3% percent occlusion.

We corrupt certain percentage of randomly chosen pixels from each of the test images, replacing their values with iid samples from a uniform distribution[8]. The corrupted pixels are randomly chosen for each test image and the locations are unknown to the computer. We vary the percentage of corrupted pixels from 0% to 90%. Figure 6 (left) visualizes several representative examples of the test images. To the human eye, beyond 50% corruption, the corrupted images (Figure 6(a) second row) are barely recognizable as face images; determining their identity by humans seems out of the question. Yet even in this extreme circumstance, Algorithm 1 correctly recovers the identity of the subject. While such random patterns of corruption are not typical of real-world occlusions, this robustness might be useful for recognition over extremely unreliable lossy communication channels, or even for recognizing faces in outdoor environments (*e.g.,* in the presence of rain, or partially occluding fence or foliage).

[7] The only reason for resizing the images is to be able to run all the experiments within the memory size of MATLAB on a typical PC. The algorithm relies on linear programming and is scalable in the image size.

[8] Uniform over $[0, y_{max}]$, where y_{max} is the largest possible pixel value.

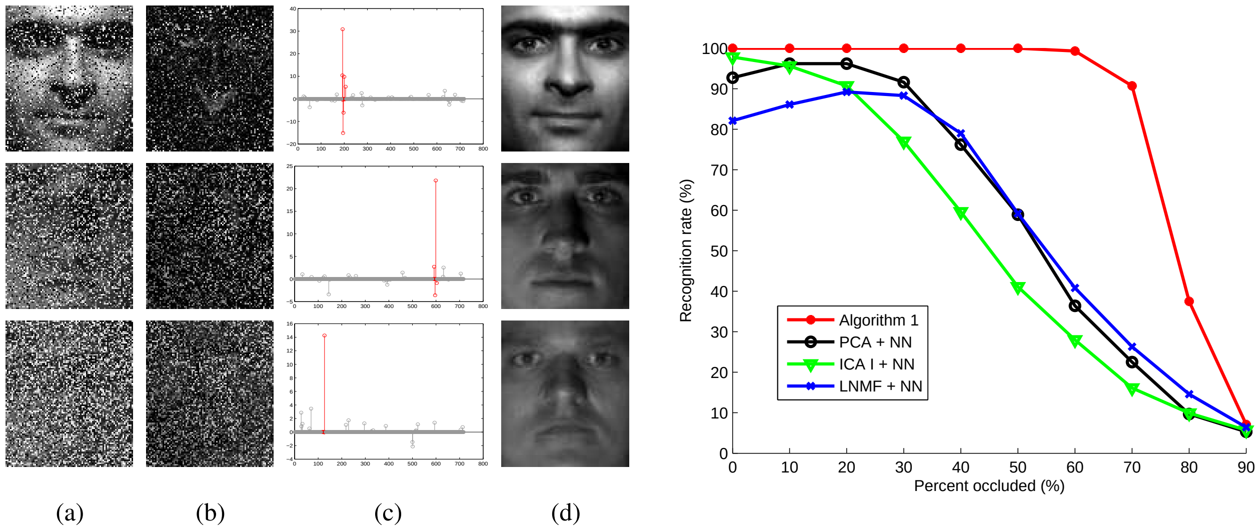


Fig. 6. **Recognition under random corruption.** Left: (a) Test images $\boldsymbol{y}$ from Extended Yale B, with random corruption. Top row: 30% of pixels are corrupted, Middle row: 50% corrupted, Bottom row: 70% corrupted. (b) Estimated errors $\hat{\boldsymbol{e}}_1$. (c) Estimated sparse coefficients $\hat{\boldsymbol{x}}_1$. (d) Reconstructed images $\boldsymbol{y}_r$. Algorithm 1 correctly identifies all three corrupted face images. Right: The recognition rate across the entire range of corruption for four algorithms. Algorithm 1 (red curve) significantly outperforms others, performing almost perfectly upto 60% random corruption.

TABLE I

RECOGNITION RATE OF ALGORITHM 1 ON EXTENDED YALE B WITH VARYING LEVEL OF RANDOM CORRUPTION.

Corruption (%)	0	10	20	30	40	50	60	70	80	90
Rec. rate (%)	100	100	100	100	100	100	99.3	90.7	37.5	7.1

We quantitatively compare our method to three popular techniques for face recognition in the vision literature. The Principal Component Analysis (PCA) approach of [4] is not robust to occlusion. There are many variations to make PCA robust to corruption or incomplete data, and some have been applied to robust face recognition, e.g., [11]. We will later discuss their performance against ours on more realistic conditions. But here we use the basic PCA to provide a standard baseline for comparison[9]. The remaining two techniques are designed to be more robust to occlusion. Independent Component Analysis (ICA) architecture I [5] attempts to express the training set as a linear combination of statistically independent basis images. Local Non-negative Matrix Factorization (LNMF) [7] approximates the training set as an additive combination of

[9] Following [26] we normalize the image pixels to have zero mean and unit variance before applying PCA.

basis images, computed with a bias toward sparse bases.[10]

Figure 6 (right) plots the recognition performance of Algorithm 1 and the other three methods, as a function of the level of corruption. We see that the algorithm dramatically outperforms others. From 0% upto 50% occlusion, Algorithm 1 correctly classifies all subjects. At 50% corruption, none of the others achieves higher than 60% recognition rate, while the proposed algorithm achieves 100%. Even at 70% occlusion, the recognition rate is still 90.7% (see Table I). This greatly surpasses the theoretical bound of worst-case corruption (13.3%) that the algorithm is ensured to tolerate. Clearly, the worst-case analysis is too conservative for random corruption.

Notice that, interestingly the recognition rates for PCA and LNMF actually increase with 10% and 20% corruption. This phenomenon is due to the differences in the statistics of the training and test image: the test images are taken from more extreme lighting conditions and hence are darker in certain areas (see Figure 14). The uniform noise contains more bright pixels than the test images, and when this corruption is filtered through the PCA and LNMF bases, its main effect is to increase the magnitude of the coefficients, essentially compensating for this overall illumination difference. Notice, however, that as the number of pixels corrupted grows larger than 30%, the performance of these algorithms degrades quickly and significantly, while Algorithm 1's performance remains stable and superior.

B. Recognition despite Random Block Occlusion

We next simulate various levels of contiguous occlusion, from 0% to 50%, by replacing a *randomly located* square block of each test image with an unrelated image, as in Figure 7(a). Again, the location of occlusion is randomly chosen for each image and is unknown to the computer. Methods that select fixed facial features or blocks of the image (*e.g.,* [8], [9]) are less likely to succeed here, due to the unpredictable location of the occlusion. Figure 7 left shows two representative results of Algorithm 1 with 30% occlusion. Figure 7(a) is the occluded image. In the second row, the entire center of the face is occluded; this is a difficult recognition task even for humans. Figure 7(b) shows the magnitude of the estimated error $\hat{e}_1$. Notice that $\hat{e}_1$ compensates not only for occlusion due to the baboon, but also for the violation of the

[10] For PCA, ICA and LNMF, the number of basis components is chosen to give the optimal test performance over the range $\{100, 200, 300, 400, 500, 600\}$.

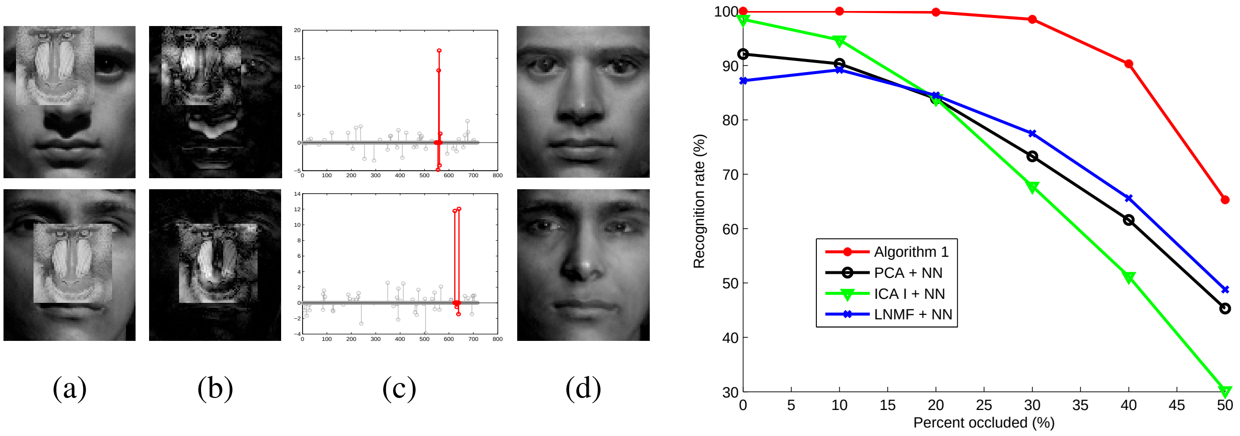


Fig. 7. **Recognition under varying level of random contiguous occlusion.** Left: (a) 30% occluded test face images $\boldsymbol{y}$ from Extended Yale B. (b) Estimated sparse errors, $\hat{\boldsymbol{e}}_1$. (c) Estimated sparse coefficients, $\hat{\boldsymbol{x}}_1$, red (darker) entries correspond to training images of the same person. (d) Reconstructed images, $\boldsymbol{y}_r$. Algorithm 1 correctly identifies both occluded faces. Right: The recognition rate across the entire range of corruption for four algorithms. Algorithm 1 (red curve) significantly outperforms others, performing almost perfectly upto 30% contiguous occlusion.

TABLE II

RECOGNITION RATE OF ALGORITHM 1 ON EXTENDED YALE B WITH VARYING LEVELS OF SYNTHETIC OCCLUSION.

Occlusion	0%	10%	20%	30%	40%	50%
Rec. rate	100%	100%	99.8%	98.5%	90.3%	65.3%

linear subspace model caused by the shadow under the nose. Figure 7(c) plots the estimated coefficient vector $\hat{\boldsymbol{x}}_1$. The red entries are coefficients corresponding to test image's true class. In both examples, the estimated coefficients are indeed sparse, and have large magnitude only for training images of the same person. In both cases, Algorithm 1 correctly classifies the occluded image. For this dataset, our Matlab implementation requires 90 seconds per test image on a PowerMac G5.

The graph in Figure 7 (right) shows the recognition rates of all four algorithms. Algorithm 1 again significantly outperforms the other three methods, for all levels of occlusion. Upto 30% occlusion, Algorithm 1 performs almost perfectly, correctly identifying over 98% of test subjects. Even at 40% occlusion, only 9.7% of subjects are misclassified (Table II). Compared to random pixel corruption, contiguous occlusion is certainly a worse type of errors for the algorithm. Notice that though, the algorithm does not assume any knowledge about the nature of corruption or occlusion. In Section IV-B, we will see if we know in advance the occlusion is contiguous, how that information can be used to customize the algorithm and greatly enhance the recognition performance.

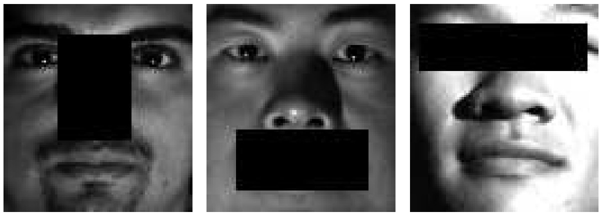
This result has interesting implications for the debate over the use of holistic versus local features in face recognition [1]. It has been suggested that both ICA I and LNMF are robust to occlusion: since their bases are locally concentrated, occlusion corrupts only a fraction of the coefficients. By contrast, if one uses ℓ^2 minimization (orthogonal projection) to express an occluded image in terms of a holistic basis such as the training images themselves, all of the coefficients may be corrupted (as in Figure 3). The implication here is that the problem is *not* the choice of representing the test image in terms of a holistic or local basis, but rather *how the representation is computed.* Properly harnessing redundancy and sparsity is the key to error correction and robustness. Extracting local or disjoint features can only reduce redundancy, resulting in inferior robustness.

C. Most Informative Face Regions for Recognition

Experiments in human vision indicate that the eye and eyebrow region is most important for face recognition by humans; if the eyebrows are removed, even familiar faces become quite difficult to recognize [27], [28]. We test the effect of occluding various face regions on our algorithm, using the Extended Yale B database. We generate three separate test sets in which a black patch occludes the eyes, nose and mouth, respectively. Figure 8 visualizes the three sets. In each, the black box occludes 20% of the image. With the nose area occluded, Algorithm 1 achieves 98.7% recognition, whereas occluding the eyes and eyebrows reduces the recognition rate to 95.6%. This corroborates the results on humans [27]; eyes appear to be the most important feature for our algorithm as well. However, the performance of our algorithm is arguably more stable than humans with respect to the choice of occluded region – the difference is merely 3% between eyes and nose.

D. Recognition Despite Disguise

Given the importance of the eyes and mouth for face recognition, it is not surprising that people trying to conceal their identity often obscure these regions (as in Figures 9 and 10). We test our



Region occluded	Rec. rate
Nose	**98.7%**
Mouth	97.1%
Eyes	95.6%

Fig. 8. **Effect of occluding different regions.** Left: test images with 20% occlusion covering different parts of the face, nose, mouth, and eyes respectively. Right: recognition rate with Algorithm 1.

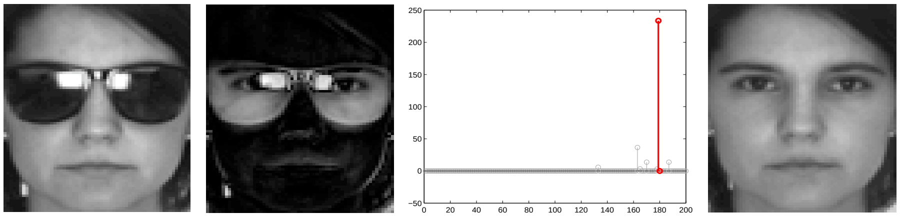
algorithm's ability to cope with such real and possibly malicious occlusions using a subset of the AR Face Database [29]. The chosen subset consists of $1,399$ images (14 each, except for a corrupted image `w-027-14.bmp`) of 100 subjects, 50 male and 50 female. All of the images are cropped, resized to 83×60 pixels, and normalized to have unit ℓ^2 norm. For training, we use 799 images (about 8 per subject) of unoccluded frontal views with varying facial expression, giving a matrix B of size $4,980 \times 5,779$. We estimate EBP$(B) \approx 577$, indicating that perfect reconstruction is possible upto 11.6% occlusion. On this dataset, our Matlab implementation requires about 75 seconds per test image on a PowerMac G5.

We consider two separate test sets of 200 images. The first test set contains images of the subjects wearing sunglasses, which occlude roughly 20% of the image. Figure 9 shows a successful example from this test set. Notice that $\hat{\boldsymbol{e}}_1$ compensates for misalignment of the image edges as well as occlusion due to sunglasses.[11]

The second test set considered contains images of the subjects wearing a scarf, which occludes roughly 40% of the image. Since the occlusion level is more than three times EBP(B), Algorithm 1 is unlikely to succeed in this domain. Figure 10 (a) shows an example that the algorithm fails to identify the correct subject. Notice that the image with the largest coefficient, Figure 10(d), is that of a bearded man whose mouth region most closely resembles the scarf.

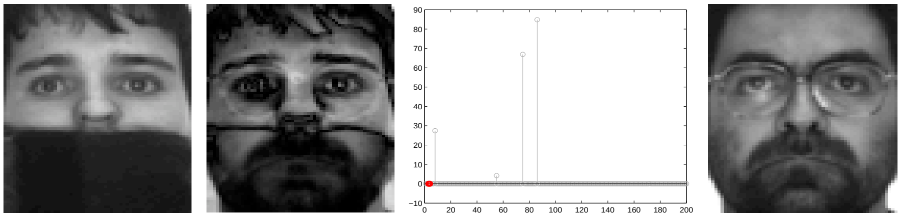
Table III compares Algorithm 1 to the other three algorithms described in the previous section. On faces occluded by sunglasses, Algorithm 1 achieves a recognition rate of 87%, more than 17% better than the nearest competitor. For occlusion by scarves, its recognition rate is 59.5%, more than double its nearest competitor but still quite poor. This confirms that although the algorithm

[11] Larger misalignments do cause problems, however. Most of the failures on this dataset seem to be due to registration errors.



(a) (b) (c) (d)

Fig. 9. **Recognition despite sunglasses.** (a) A test image from AR Database, occluded by sunglasses. (b) Magnitude of estimated sparse error, $\hat{\boldsymbol{e}}_1$. (c) Estimated sparse coefficient vector, $\hat{\boldsymbol{x}}_1$. Red (darker) coefficients correspond to training images from the same class. (d) Reconstructed image, $\boldsymbol{y}_r$.



(a) (b) (c) (d)

Fig. 10. **A failed example in the scarf test set.** (a) A test image from AR database, roughly 40% occluded by scarf. (b) Magnitude of estimated sparse error, $\hat{\boldsymbol{e}}_1$. (c) Estimated sparse coefficient vector, $\hat{\boldsymbol{x}}_1$. The largest do not belong to the correct class. (d) Training image with the largest coefficient, bearded man. When amount of occlusion exceeds EBP(B), the algorithm is susceptible to occlusions that resemble one of the training images.

is provably robust to *arbitrary* occlusions upto EBP(B), beyond that point it is sensitive to occlusions that resemble regions of a training image from a different individual. Because the amount of occlusion exceeds the breakdown point, additional assumptions are needed to achieve good recognition. In Section IV-B, we will describe how spatial continuity of the occlusion can be exploited within our framework, allowing successful recognition even in this difficult circumstance (see Table IV).

IV. Variations and Improvements of the Algorithm

A. *Outlier Rejection and Reciever Operator Characteristic*

Practical recognition systems are sometimes confronted with invalid test images: images of some person not in the gallery of training images, or even images of some completely different

TABLE III

Performance on the AR database. ALGORITHM 1 AND ITS PARTITIONED VERSION (DESCRIBED IN SECTION IV-B) ACHIEVE THE HIGHEST RECOGNITION RATE.

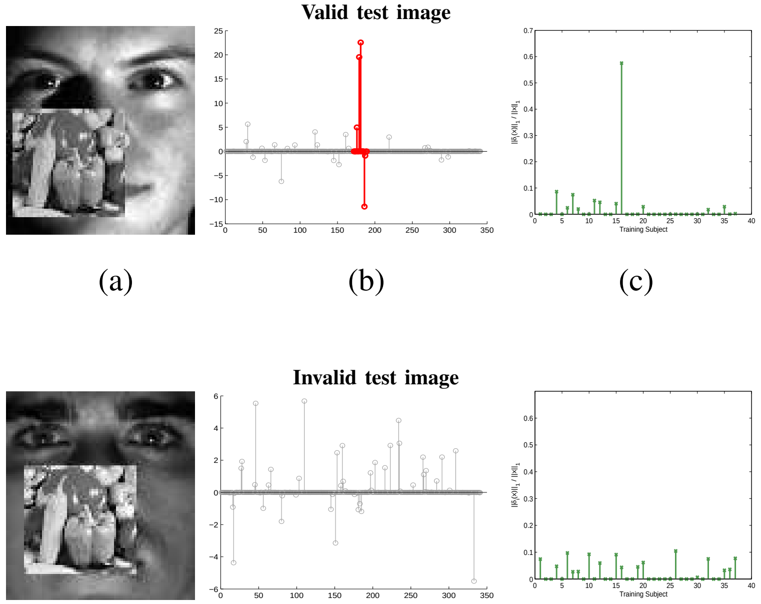
Algorithms	Recognition rate with sunglasses	Recognition rate with scarves
Algorithm 1	**87.0%**	**59.5%**
PCA + NN	70.0%	12.0%
ICA I + NN	53.5%	15.0%
LNMF + NN	33.5%	24.0%

object. Thus, in addition to returning a best match within the training set, an effective recognition system should also return a measure of confidence as to whether the test image represents any of the gallery subjects, or even reject invalid images outright. One simple and effective heuristic comes from the observation that the coefficients $\hat{\boldsymbol{x}}_1$ associated with invalid test images are generally not concentrated on any individual subject, but rather be spread across several training subjects, as shown in Figure 11. We can therefore reject images for which

$$\frac{k \max_i \|\delta_i(\hat{\boldsymbol{x}}_1)\|_1 / \|\hat{\boldsymbol{x}}_1\|_1 - 1}{k-1} < \tau, \tag{11}$$

where k is the number of classes and τ is a preselected threshold. The left hand side of (11) falls in $[0,1]$, with value 0 if the coefficients are evenly spread between the classes and 1 if all of the coefficients are associated with a single class. We dub this quantity the *sparsity concentration index (SCI)*.

We test this simple outlier rejection rule on the Extended Yale B database, using Subsets 1 and 2 for training and Subset 3 for testing as before. We again simulate varying levels of occlusion (10%, 30%, 50%) by replacing a randomly chosen block of each test image with an unrelated image. However, in this experiment, we include only half of the subjects in the training set. Thus, half of the subjects in the testing set are new to the algorithm. We test the system's ability to determine whether a given test subject is in the training database or not by sweeping the threshold τ through a range of values in $[0,1]$, generating the reciever operator characteristic (ROC) curves in Figure 12. For comparison, we also considered outlier rejection by thresholding



(a) (b) (c)

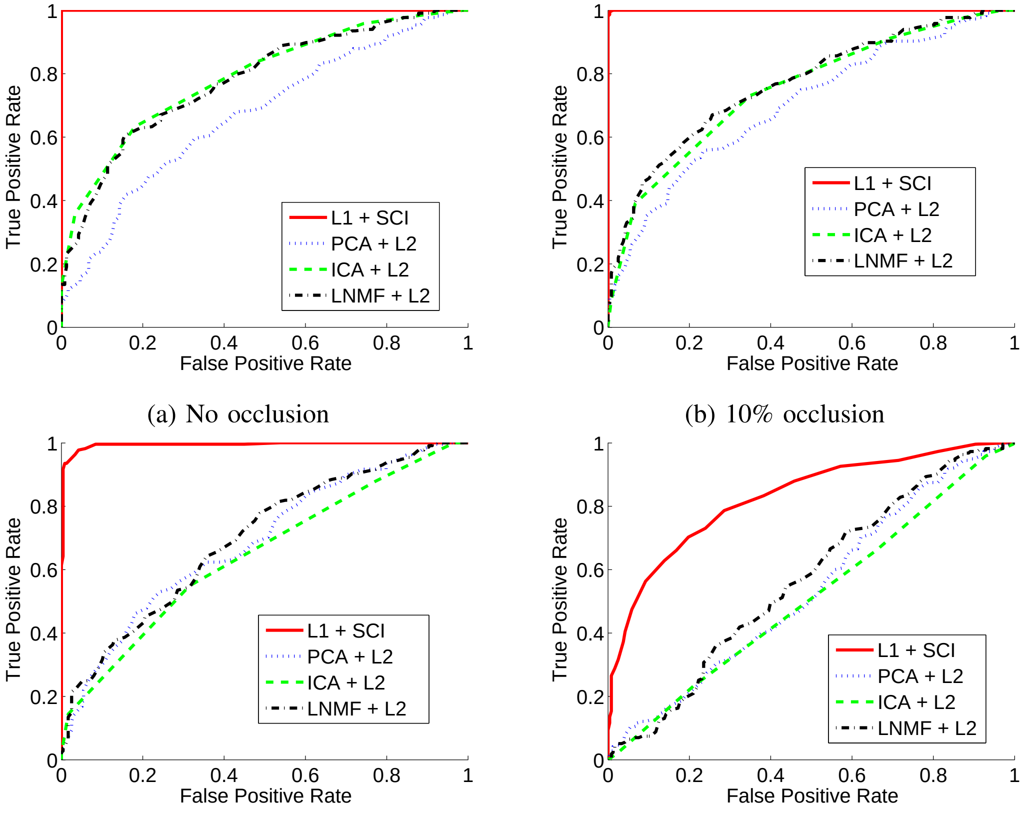
(a) (b) (c)

Fig. 11. **Rejecting invalid test images.** Top: valid test image (subject from the training set). Bottom: invalid test image (subject not from the training set). (a) Test image, 30% occluded. (b) Recovered coefficients. (c) Value of the ratio (11) for each class, i. Notice that the coefficients for the valid test image are much more sparse, and the ratio (11) is much larger.

the Euclidean distance between (features of) the test image and (features of) the nearest training images within the PCA, ICA and LNMF feature spaces. These curves are also displayed in Figure 12. Notice that the simple rejection rule (11) performs nearly perfectly at 10% and 30% occlusion. At 50% occlusion, it still significantly outperforms the other three algorithms, and is the only one of the four algorithms that performs significantly better than chance.

B. Aggregating Image Blocks

In Section III, we saw that with no assumptions on the support of the occlusion, e, Algorithm 1 was still able to tolerate arbitrary occlusions of upto 30% of the image pixels. Moreover, in cases where the occlusion is roughly orthogonal to all of the training images (*e.g.,* the corruption example in Figure 6), the algorithm tolerates upto 70% corruption. Nevertheless, thus far we have not exploited the fact that in many real recognition scenarios, the occlusion falls on some patch of image pixels which is a-priori unknown, but is known to be connected. A somewhat traditional approach (explored in [8], [9] amongst others) to exploiting this information in face recognition is to partition the image into blocks and process each block independently. The results for



(a) No occlusion

(b) 10% occlusion

(c) 30% occlusion

(d) 50% occlusion

Fig. 12. **ROC curves for outlier rejection.** Vertical axis: true positive rate. Horizontal axis: false positive rate. The solid red curve is generated by computing a sparse representation as in Algorithm 1 and then rejecting outliers via equation (11). This approach performs almost perfectly for upto 30% occlusion.

individual blocks are then aggregated, for example, by voting, while discarding blocks believed to be occluded (using, say, the outlier rejection rule introduced above). The major difficulty with this approach is that the occlusion cannot be expected to respect any fixed partition of the image; while only a few blocks are assumed to be completely occluded, some or all of the remaining blocks may be partially occluded. Thus, in such a scheme there is still a need for robust techniques *within each block*.

If the amount of occlusion is known to be less than $\mathrm{EBP}(B)$, Algorithm 1 can and should be directly applied, as it will achieve superior performance. If, however, this cannot be guaranteed, performing an ℓ^1 minimization within each block can increase the level of occlusion the algorithm tolerates, at the expense of generality. Moreover, such an approach strictly improves existing block techniques based on non-robust methods such as PCA.

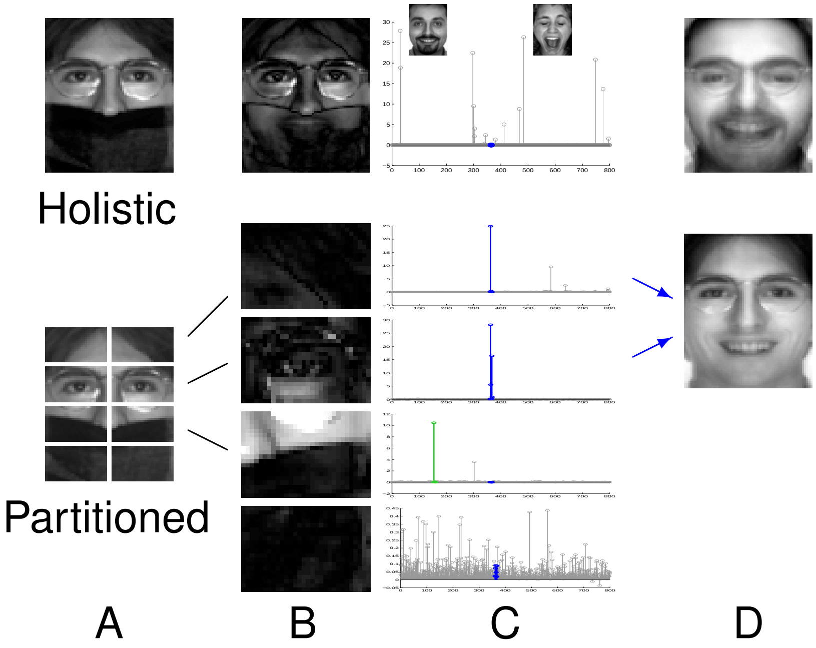


Fig. 13. **A partition scheme to tackle contiguous disguise.** If the occlusion is known to be contiguous, one can partition the image into multiple smaller blocks, apply Algorithm 1 to each of the blocks and then aggregate the results by voting. This process is illustrated by the figure to the left. (A) Test image, occluded by scarf. (B) Estimated sparse error $\hat{e}_1$. (C) Estimated sparse coefficients $\hat{\boldsymbol{x}}_1$. Small images pictured are the training images corresponding to large coefficients. (D) Reconstructed image. The top row visualizes the performance of Algorithm 1 on the whole image (holistic) notice that the two largest coefficients correspond to a bearded man and a screaming woman, two images whose mouth region resembles the occluding scarf. The second row visualizes the partition-based scheme described above. The table shows the performance of all the algorithms for both types of occlusion.

We partition each of the training images into L blocks of size $a \times b$, producing a set of matrices $A^{(1)}, \ldots, A^{(L)} \in \mathbb{R}^{p \times n}$, where $p \doteq ab$. We similarly partition the test image $\boldsymbol{y}$ into $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(L)} \in \mathbb{R}^p$. We write the l-th block of the test image as a sparse linear combination $A^{(l)}\boldsymbol{x}^{(l)}$ of l-th blocks of the training images, plus a sparse error $\boldsymbol{e}^{(l)} \in \mathbb{R}^p$: $\boldsymbol{y}^{(l)} = A^{(l)}\boldsymbol{x}^{(l)} + \boldsymbol{e}^{(l)}$. We can recover can again recover a sparse $\boldsymbol{w}^{(l)} = [\boldsymbol{x}^{(l)}\ \boldsymbol{e}^{(l)}] \in \mathbb{R}^{n+p}$ by ℓ^1 minimization:

$$\hat{\boldsymbol{w}}_1^{(l)} \doteq \underset{\boldsymbol{w} \in \mathbb{R}^{n+p}}{\operatorname{argmin}} \|\boldsymbol{w}\|_1 \text{ subject to } \left[A^{(l)}\ I\right] \boldsymbol{w} = \boldsymbol{y}^{(l)}. \qquad (12)$$

We apply the classifier from Algorithm 1 within each block[12] and then aggregate the results by voting.

[12] Totally occluded blocks can also be rejected via Equation (11). In practice, we find that this does not significantly increase the recognition rate.

TABLE IV

Performance on the AR database. GENDER/CONDITION BREAKDOWN OF RECOGNITION RATES USING ALGORITHM 1 WITH PARTITIONING.

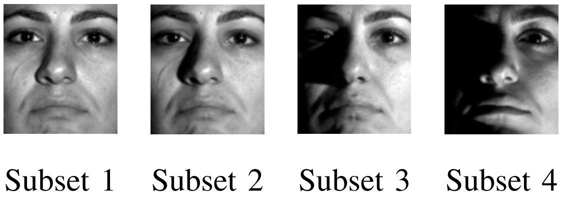
Cases	Rec. rate	Cases	Rec. rate
Sunglasses	97.5%	Scarves	93.5%
Men	97.5%	Women	93.5%
Men, sunglasses	100%	Women, sunglasses	95%
Men, scarves	95%	Women, scarves	92%

We verify the efficacy of this scheme on the AR database for faces disguised with sunglasses or scarves. We partition the images into eight (4×2) blocks, each of size 20×30 pixels. Doing so increases the overall recognition rate on scarf images from 59.5% to 93.5%, and also improves the recognition rate on sunglasses images from 87.0% to 97.5%. This performance exceeds the best known results on the AR dataset [11] to date. That work obtains 84% on the sunglasses and 93% on the scarfs, on only 50 subjects, using more sophisticated random sampling techniques. Table IV shows a more complete breakdown of the new recognition rate by gender and type of occlusion. Interestingly, females are consistently more difficult to recognize on this dataset. We conjecture that this may be due to the presence of additional distinguishing features (e.g. beard) on the male face, as well as increased within-class variability in female hairstyles.

C. Designing the Training Set for Robustness

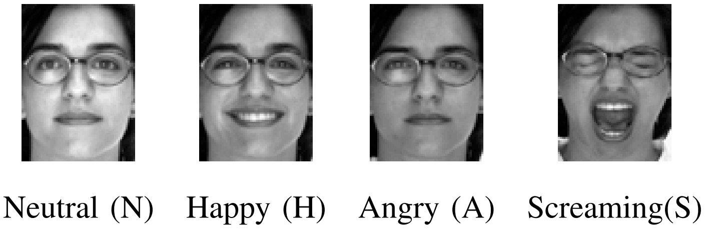
An important consideration in designing recognition systems is selecting the number of training images as well as the conditions (viewpoint, expression, lighting) under which they are to be taken. The set of training images should obviously be extensive enough to span the conditions that occur in the test set, *i.e.,* they should be "sufficient" from a pattern recognition standpoint. The notion of equivalence breakdown point discussed in Section 2 provides a different, quantitative measure of the goodness of the training set: higher EBP implies greater robustness to occlusion.

In fact, these two concerns, sufficiency and robustness, are complementary. Figure 14 shows the estimated breakdown point for the four subsets of the Extended Yale B database. Notice that the highest EBP, $\approx 1,330$, is achieved with Subset 4, the most extreme lighting conditions. Figure 15



Training	1	2	3	**4**
Est. EBP	1,124	1,122	1,190	**1,330**

Fig. 14. **Robust design: illumination** Top: four subsets of Extended Yale B, containing increasingly extreme lighting conditions. Bottom: equivalence breakdown point EBP(B) for each subset.



Training	**N+H**	N+A	N+S	H+A	H+S	A+S
Est. EBP	**585**	421	545	490	550	510

Fig. 15. **Robust design: facial expression.** Top: four facial expressions in the AR database. Bottom: equivalence breakdown point EBP(B) when taking the training set from different pairs of expressions.

shows the breakdown point for subsets of the AR database with different facial expressions. The dataset contains four facial expressions, Neutral, Happy, Angry, and Scream, pictured in Figure 15. We generate training sets from all pairs of expressions and compute the EBP of each. The highest breakdown points are achieved by the Neutral+Happy and Happy+Scream combinations, while the lowest comes from Neutral+Angry. Notice that the Neutral and Angry images are quite similar in appearance, while (for example) Happy and Scream are very dissimilar.

Thus, both for varying lighting (Figure 14) and expression (Figure 15), training sets with wider variation in the images allow greater robustness to occlusion. Designing a training set that allows recognition under widely varying conditions does not hinder our algorithm; in fact it helps it. However, the training set should not contain too many similar images, as in the Neutral+Angry example of Figure 15. In the language of signal representation, the training images should form an *incoherent dictionary* [15].

V. Discussion and Future Work

In this paper, we have shown how the theory of sparse representation and ℓ^1 minimization significantly improves the robustness of computational face recognition. Furthermore, results from compressed sensing[13] [30] suggest an important role for sparsity in the choice of features for face recogntion *without occlusion*. This issue is discussed in detail in the companion paper [20], which advocates that if the sparsity in the recognition problem is properly harnessed, details of feature selection become less important. In addition, in both papers, we witness that sparsity of the representation provides an effective cue for validating the test image. It advocates the importance of the global, or joint-class, representation in validation, as opposed to representation by individual classes.

This paper and its companion [20] have introduced a novel and comprehensive approach to face recognition which has provided new understandings on many fundamental issues such as feature selection, occlusion, and outlier rejection. The formulation of the problem via the perspective of sparse representation also leads to simple, efficient, and robust algorithms based on the mathematical theory of compressed sensing. We strongly believe that this new approach and framework may provide new solutions to many other problems in the general area of pattern analysis and object recognition.

An intriguing problem for future work is whether this framework can be useful for object detection, in addition to recognition. The usefulness of sparsity in detection has been noticed in the work of [17] and more recently explored in [31]. We believe that the full potential of sparsity in object detection and recognition together is yet to be uncovered. From a practical standpoint, it would also be useful to extend the algorithm to handle less constrained conditions, such as variations in object pose. Robustness to occlusion allows the algorithm to tolerate minor pose variation or misalignment. In the companion paper [20], we have discussed that sparse representation can potentially adapt to a nonlinear distribution that could be used to model face images with varying pose. However, it remains open problems how effective sparse representation will still be in this case and how many more training images it may require.

[13] Especially the broad equivalence properties of random matrices.

References

[1] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, pp. 399–458, 2003.

[2] "Electronic privacy information center, face recognition," *http://www.epic.org/privacy/facerecognition/*, 2006.

[3] F. Macwilliams and N. Sloane, *The Theory of Error-Correcting Codes*. North-Holland, 1981.

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

[5] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, 2005.

[6] P. Belhumeur, J. Hesanha, and D. Kreigman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[7] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 1–6.

[8] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[9] A. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[10] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 99–118, 2000.

[11] F. Sanja, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, 2006.

[12] E. Candes, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," in *IEEE Symposium on FOCS*, 2005, pp. 295–308.

[13] E. Candes and P. A. Randall, "Highly robust error correction by convex programming," *preprint, http://arxiv.org/abs/cs.IT/0612124*, 2006.

[14] D. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm near solution approximates the sparsest solution," *preprint, http://www-stat.stanford.edu/ donoho/Reports/*, 2004.

[15] ——, "For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution," *Comm. Pure and Applied Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[16] I. Ciocoiu, "Occluded face recognition using parts-based representation methods," in *Proceedings of European Conference on Circuit Theory and Design*, 2005.

[17] D. Geiger, T. Liu, and M. Donahue, "Sparse representations for image decompositions," *International Journal of Computer Vision*, vol. 33, no. 2, 1999.

[18] R. Basri and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 218–233, 2003.

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

[20] A. Yang, J. Wright, S. Sastry, and Y. Ma, "Feature selection in face recognition: A sparse representation perspective,"

Technical Report, University of Illinois, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.

[21] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237–260, 1998.

[22] D. Donoho, "Neighborly polytopes and sparse solution of underdetermined linear equations," *IEEE Transactions on Information Theory*, 2006.

[23] Y. Sharon, J. Wright, and Y. Ma, "Computation and relaxation of conditions for equivalence between ℓ^1 and ℓ^0 minimization," *submitted to IEEE Transactions on Information Theory*, 2007.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[25] E. Candes and J. Romberg, "ℓ^1-magic: Recovery of sparse signals via convex programming," *http://www.acm.caltech.edu/l1magic/*, 2005.

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

[27] J. Sadr, I. Jarudi, and P. Sinha, "The role of eyebrows in face recognition," *Perception*, vol. 32, pp. 285–293, 2003.

[28] H. Ellise, J. Shepherd, and G. Davies, "Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition," *Perception*, vol. 8, no. 4, pp. 431–439, 1979.

[29] A. Martinez and R. Benavente, "The AR face database," *CVC Tech. Report No. 24*, 1998.

[30] E. Candes, "Compressive sampling," in *Proc. International Congress of Mathematicians*, 2006.

[31] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *Proc. Neural Information and Processing Systems*, 2006.