# Towards a robust face recognition system using compressive sensing

*Allen Y. Yang*[1], *Zihan Zhou*[2], *Yi Ma*[2], *and S. Shankar Sastry*[1]

[1]Department of EECS, University of California, Berkeley, USA
[2]Coordinated Science Lab, University of Illinois, Urbana, USA

yang@eecs.berkeley.edu, zzhou7@uiuc.edu, yima@uiuc.edu, sastry@eecs.berkeley.edu

## Abstract

An application of compressive sensing (CS) theory in image-based robust face recognition is considered. Most contemporary face recognition systems suffer from limited abilities to handle image nuisances such as illumination, facial disguise, and pose misalignment. Motivated by CS, the problem has been recently cast in a sparse representation framework: The sparsest linear combination of a query image is sought using all prior training images as an overcomplete dictionary, and the dominant sparse coefficients reveal the identity of the query image. The ability to perform dense error correction directly in the image space also provides an intriguing solution to compensate pixel corruption and improve the recognition accuracy exceeding most existing solutions. Furthermore, a local iterative process can be applied to solve for an image transformation applied to the face region when the query image is misaligned. Finally, we discuss the state of the art in fast $\ell_1$-minimization to improve the speed of the robust face recognition system. The paper also provides useful guidelines to practitioners working in similar fields, such as acoustic/speech recognition.

**Index Terms**: face recognition, compressive sensing, $\ell_1$-minimization

## 1. Introduction

Face recognition has been a classical problem in pattern recognition. The main research can be categorized in two closely related areas. First, due to the concern of high dimensionality in the facial image space, investigators are interested in searching for effective dimensionality reduction methods to extract useful image features, either holistic [1, 2] or local [3, 4], to concisely represent the appearance of facial images. These low-dimensional vector representations are often called *face features*. Second, most existing classifiers have been applied to face recognition using the extracted face feature space, including nearest neighbor (NN) and support vector machines (SVM).

Although human perception is known to be very effective in identifying human subjects from facial images, most contemporary face recognition systems have failed to achieve equally good recognition accuracy. It is well understood that the performance is affected by image nuisances, including illumination, pixel corruption, facial disguise, and 3-D pose variation (as shown in Figure 1). Traditional methods would have difficulty in compensating these nuisances in the image space or extracting robust face features that are not sensitive to the nuisances.

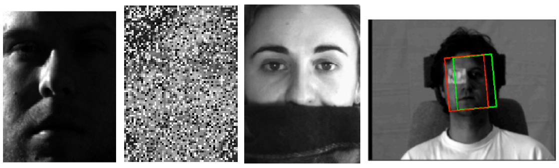

Figure 1: Examples of image nuisances in face recognition. **Left**: Illumination. **Middle Left**: Pixel corruption. **Middle Right**: Disguise. **Right:** Occlusion and misalignment, where an irrelevant image is superimposed to occlude the right-eye region. The green bounding box indicates the initial alignment; the red bounding box is the final position after misalignment correction (see Section 4).

In this paper, we provide an overview about a recent solution to robust face recognition [5, 6], which has been motivated by the emerging theory of compressive sensing (CS) [7, 8]. The method reformulates the face recognition problem as a sparse representation problem. In this framework, the distribution of multiple classes is modeled as a mixture of subspaces, one for each class. Given $C$ classes and a query image $\boldsymbol{b}$ (stacked in vector form), the method seeks the sparsest linear representation of the sample with respect to all training examples:

$$\boldsymbol{b} = [A_1, A_2, \cdots, A_C]\boldsymbol{x} = A\boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

where the column vectors of each $A_i$ represent training examples from the $i$th class. The method stipulates that if $\boldsymbol{b}$ is a valid sample from the true class $i$, it satisfies a linear model $\boldsymbol{b} = A_i\boldsymbol{x}_i$. Therefore, the corresponding representation in (1) admits a sparse representation $\boldsymbol{x} = [\cdots, \mathbf{0}^T, \boldsymbol{x}_i^T, \mathbf{0}^T, \cdots]^T \in \mathbb{R}^n$: on average only a fraction of $\frac{1}{C}$ coefficients are nonzero. Furthermore, the dominant nonzero coefficients in $\boldsymbol{x}$ reveal the true class of sample $\boldsymbol{b}$, as shown in Figure 2.

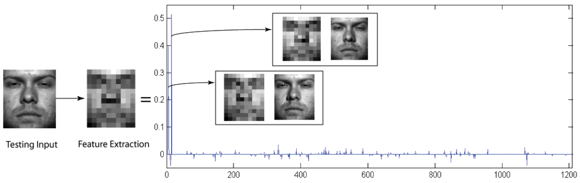

Figure 2: Sparse representation of a 12×10 downsampled query image, which belongs to Class 1 [5].

In addition, extensions of the basic formulation (1) also provide quite surprising results to address a wide arrange of problems in face recognition, such as dimensionality reduction, image corruption, and face alignment. In certain conditions, the performance of the new algorithm may even exceed that of human perception.

This work was supported in part by ARO MURI W911NF-06-1-0076, NSF IIS 08-49292, NSF ECCS 07-01676, and ONR N00014-09-1-0230.

Firstly, for dimensionality reduction, if a linear projection $R$ is chosen to reduce the dimension of the linear model:

$$\tilde{\boldsymbol{b}} \doteq R\boldsymbol{b} = RA\boldsymbol{x} = \tilde{A}\boldsymbol{x}, \qquad (2)$$

one can show that the recognition rates of most classical feature spaces (e.g., principal component analysis (PCA) and locality preserving projection (LPP) [2]) all converge to high accuracy, if the dimension of the feature space is sufficiently high. Furthermore, random projections [7, 8, 5] as an unconventional linear operator can achieve equally high accuracy compared to the traditional operators. To this end, the choice of face features becomes *insignificant*, as long as $\boldsymbol{x}$ is properly sought.

Secondly, for error corruption to compensate image corruption and disguise, a slightly modified linear constraint can be considered:

$$\boldsymbol{b} = A\boldsymbol{x} + \boldsymbol{e}, \qquad (3)$$

where $\boldsymbol{e}$ is another unknown vector whose nonzero entries correspond to the corrupted pixels in the observation $\boldsymbol{b}$. One can show that simultaneously minimizing the sparsity of $\boldsymbol{x}$ and $\boldsymbol{e}$ can effectively compensate the corrupted pixel values in $\boldsymbol{b}$, and at the same time correctly classify the identity of the underlying face with very high accuracy. To this end, the percentage of corrupted pixels that the algorithm can correct approaches $100\%$ asymptotically [9].

Thirdly, in the general setup (1), we often assume both the training images and the query image are properly aligned to the frontal position. In other words, salient facial features such as the eyes, nose, or mouth are assumed to share the same image coordinates, respectively. In the presence of large misalignment in the query image $\boldsymbol{b}$, the solution $\boldsymbol{x}$ may not be sparse. Nevertheless, a local iterative process can be applied as a preprocessing step to correct the misalignment up to some finite-dimensional group of transformations $T$ on the image plane [6]:

$$\boldsymbol{b} \circ \tau = A\boldsymbol{x} + \boldsymbol{e}, \qquad (4)$$

where $\tau \in T$ represents a 2-D transformation applied to the query image $\boldsymbol{b}$. The experiment has shown that the algorithm that primarily applies to the 2-D image plane can effectively compensate 3-D pose variation up to $\pm 45^\circ$ away from the frontal position.

Finally, practitioners should be concerned about the computational complexity of recovering a sparse signal from systems of linear equations (3) or (4). Traditionally, they have been formulated as a linear or quadratic programming problem in convex optimization, which is called *basis pursuit* (BP). The complexity of the standard steepest descent interior-point methods for BP is bounded by $O(n^3)$, where $n$ denotes the number of training examples. More recently, several first-order approximations of the linear programming problem have been proposed [10]. These algorithms can be efficiently implemented to process very high-dimensional data, and they are much faster than the interior-point methods.

## 2. Sparsity-based Classification

For a sparse signal $\boldsymbol{x}_0$, denote $k = \|\boldsymbol{x}_0\|_0$ as its sparsity (i.e., the number of nonzero coefficients). In a system of linear equations (1), if the dictionary $A$ is overdetermined, the solution can be uniquely determined by taking the pseudo-inverse: $\boldsymbol{x}^* = A^\dagger \boldsymbol{b}$, which is a linear least squares problem. In sparsity-based classification (SBC), we are interested in the case where $A$ is underdetermined. Since the number of observations in $\boldsymbol{b}$ is usually much smaller than the number of unknowns in $\boldsymbol{x}$, clearly there exist infinitely many solutions of $\boldsymbol{x}$.

The results in CS [11, 7, 8, 12] reveal that if $\boldsymbol{x}_0$ is sufficiently sparse and $A$ is *incoherent* to the basis in which $\boldsymbol{x}_0$ is sparse, the solution can be uniquely recovered by solving the following $\ell_1$-minimization ($\ell_1$-min) program:

$$(P_1): \quad \boldsymbol{x}^* = \arg\min \|\boldsymbol{x}\|_1 \quad \text{subj. to} \quad \boldsymbol{b} = A\boldsymbol{x}. \qquad (5)$$

The literature of convex optimization has provided several $\ell_1$-min solvers that actually predate CS theory, including *orthogonal matching pursuit* (OMP)[13], *basis pursuit* (BP)[14], and the *LASSO*[15]. In Section 5, we will discuss how to improve the speed of $\ell_1$-min solvers by contemporary first-order methods.

One particular problem in solving $(P_1)$ is that the dimension of a face image $\boldsymbol{b}$ may be very high.[1] High dimensionality not only affects the complexity of the $(P_1)$ algorithm, but also violates the fundamental assumption in CS that the dictionary $A$ shall be underdetermined. For face recognition, it means the dimension of the images in vector form is much larger than the number of available training examples, i.e., $d \gg n$.

Since the number of training examples $n$ is often determined by the application, in order to maintain an underdetermined dictionary $A$, dimensionality reduction methods can be employed. Many well-known operators such as PCA, LDA, and LPP can be treated as a linear projection $R \in \mathbb{R}^{d' \times d}$, as shown in (2). After the projection, the dimension of the system becomes smaller than the dimension of $\boldsymbol{x}$, i.e., $d' < n$.

In CS, *random projections* have been considered as a universal dimensionality reduction technique. In particular, $R$ is a Gaussian random matrix if its entries are drawn independently from a Gaussian distribution. One can show that, in general, random projections are incoherent to most classical orthonormal basis. A short insight to this result is that with high probability, randomly generated column vectors of $A$ are linearly independent. We have compared the performance of the SBC algorithm under different dimensionality reduction methods in [5]. The results corroborate the theoretical findings in CS that random projections perform equally well or even better than many traditional methods when the feature space dimension is sufficiently high (e.g., $d' > 500$). In addition to good performance, it is worth noting that random projections are data independent and extremely efficient to generate at any dimension compared to other methods such as PCA and LPP.

## 3. Corruption and Disguise Compensation

In this section, we consider the situation where the query image $\boldsymbol{b}$ may be severely corrupted or occluded. The problem is modeled by a linear system (3) with an additional error term $\boldsymbol{e}$. In [5], the authors have proposed to simultaneously recover the sparse signals $\boldsymbol{x}$ and $\boldsymbol{e}$ in the following $\ell_1$-min problem:

$$\min \|\boldsymbol{w}\|_1 \quad \text{subj. to} \quad \boldsymbol{b} = [A, I]\boldsymbol{w}, \qquad (6)$$

where $I \in \mathbb{R}^{d \times d}$ is an identity matrix and $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{e}^T]^T \in \mathbb{R}^{n+d}$ is also assumed sparse.

In (6), the new dictionary $[A, I]$ has been dubbed as a cross-and-bouquet (CAB) model in the following sense. The columns of A are highly correlated, as the convex hull spanned by all

[1] For example, a grayscale 640 × 480 image contains more than 300,000 pixels, i.e., the dimension of the linear system $d$ >300,000.

face images only occupies an extremely tiny portion of the image space $\mathbb{R}^d$. These training vectors are tightly bundled together as a "bouquet"; whereas the vectors in the identity matrix and their negative counterparts $\pm I$ form a $d$-dim "cross", as shown in Figure 3. A quite surprising result was shown in [9] that accurate recovery of sparse signals $\boldsymbol{x}$ is still possible and computationally feasible even when the fraction of corruption approaches $100\%$ as the dimension $d$ goes to infinity.

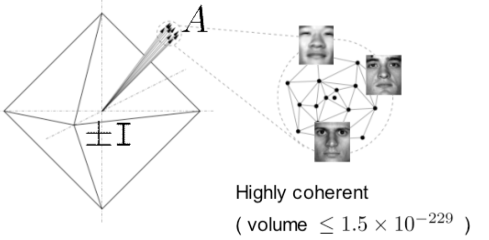

Figure 3: The CAB model for face recognition. The raw images of human faces expressed as columns of $A$ are clustered with very small variance. (Courtesy of John Wright [9])

## 4. Image Alignment

In addition to possible pixel corruption from the previous section, suppose $\boldsymbol{b}$ is also subject to some misalignment (as shown in Figure 1 Right). Therefore, the observation can be approximated as a warped image $\boldsymbol{b} = \boldsymbol{b}_0 \circ \tau^{-1}$ for some 2-D transformation $\tau$. In this case, directly seeking a sparse representation of $\boldsymbol{b}$ against properly aligned training images is no longer appropriate. Nevertheless, if the true deformation $\tau$ can be efficiently found, then we can still recover $\boldsymbol{b}_0$ and it becomes possible to find a sparse representation $\boldsymbol{x}$ in (4).

Naturally, one would like to use the sparsity as a strong cue for finding the correct deformation $\tau$, such as simultaneous minimization of $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{e}\|_1$ as in (6). However, after adding the unknown transformation $\tau$ on the left-hand side, it becomes a difficult nonconvex optimization problem. Furthermore, due to the concern of local minima, the query image $\boldsymbol{b}$ may be aligned to multiple subjects in the database. Hence, it is more desirable to locally seek the best alignment w.r.t. each subject $i$ as:

$$\tau_i^* = \arg\min_{\boldsymbol{x},\boldsymbol{e},\tau_i} \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{b} \circ \tau_i = A_i \boldsymbol{x} + \boldsymbol{e}. \qquad (7)$$

In (7), $\|\boldsymbol{x}\|_1$ is no longer penalized, since $A_i$ only contains images of the same subject. In practice, the initial values of the transformation parameters $\tau^{(0)}$ are usually obtained by applying a face detector to the test image. Then the estimate of $\tau$ can be iteratively refined by repeatedly linearizing (7), which leads to a convex program:

$$\min_{\boldsymbol{x},\boldsymbol{e},\Delta\tau_i} \|\boldsymbol{e}\|_1 \text{ subj. to } \boldsymbol{b}\circ\tau_i + \nabla_\tau(\boldsymbol{b}\circ\tau_i)\cdot\Delta\tau_i \approx A_i\boldsymbol{x}+\boldsymbol{e}. \quad (8)$$

During each iteration $k$, the current alignment parameters $\tau_i^{(k)}$ correct the observation as $\boldsymbol{b}_i^{(k)} = \boldsymbol{b} \circ \tau_i^{(k)}$. Denote $J_i^{(k)} = \nabla_{\tau_i}(\boldsymbol{b} \circ \tau_i^{(k)})$, then the update $\Delta\tau_i$ of the transformation estimate can be computed by solving the following problem:

$$\min_{\boldsymbol{w},\boldsymbol{e}} \|\boldsymbol{e}\|_1 \quad \text{subj. to} \quad \boldsymbol{b}_i^{(k)} = [A_i, -J_i^{(k)}]\boldsymbol{w} + \boldsymbol{e}, \qquad (9)$$

where $\boldsymbol{w} \doteq [\boldsymbol{x}^T, \Delta\tau_i^T]^T$. The convex program (9) can then be solved by $\ell_1$-min algorithms (with necessary modifications).

## 5. Fast $\ell_1$-Min Algorithms

Finally, we briefly discuss the state of the art in solving the convex program $(P_1)$ via accelerated $\ell_1$-min techniques. A comprehensive review of existing fast $\ell_1$-min algorithms can be found in [10].

The convex program $(P_1)$ has traditionally been formulated as a linear programming problem called *basis pursuit* (BP), which has several well-known solutions via interior-point methods. However, the computational complexity of these interior-point methods is often too high for many real-world, large-scale applications. The main reason is that they all involve expensive operations such as matrix factorization and solving linear least squares.

Recently, *iteraive shrinkage-thresholding* (IST) methods have been proposed as a good approximation to the exact BP solutions. The approach is also appealing to large-scale applications because its implementation mainly involves lightweight operations such as vector operations and matrix-vector multiplications, in contrast to other past $\ell_1$-min algorithms.

In a nutshell, IST considers a variation of $(P_1)$ that takes into account the existence of measurement errors in the sensing process:

$$(P_{1,2}): \quad \min \|\boldsymbol{x}\|_1 \quad \text{subj. to} \quad \|\boldsymbol{b} - A\boldsymbol{x}\|_2 \leq \epsilon, \qquad (10)$$

where $\epsilon$ is a bound on the additive white noise in $\boldsymbol{b}$. By the Lagrangian method, $(P_{1,2})$ is rewritten as an unconstrained *composite objective function*:

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) \doteq \frac{1}{2}\|\boldsymbol{b} - A\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}), \quad (11)$$

where $\lambda > 0$ is a scalar parameter.

We can immediately see that the main issue in optimizing such a composite function $F(\boldsymbol{x})$ is that its second term $\|\boldsymbol{x}\|_1$ is not a smooth function and therefore is not differentiable everywhere. Nevertheless, one can always approximate the objective function in an iterative fashion as [16, 17]:

$$\begin{aligned} \boldsymbol{x}^{(k+1)} \quad \approx \quad & \arg\min_{\boldsymbol{x}}\{(\boldsymbol{x} - \boldsymbol{x}^{(k)})^T \nabla f(\boldsymbol{x}^{(k)}) \\ & + \tfrac{\alpha^{(k)}}{2}\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\|_2^2 + \lambda g(\boldsymbol{x})\}, \end{aligned} \qquad (12)$$

where the hessian $\nabla^2 f(\boldsymbol{x}^{(k)})$ is approximated by a diagonal matrix $\alpha^{(k)} I$ to further reduce the computational cost.

Then one can show that the objective function (12) has a closed-form solution called the *soft-thresholding* function [16, 17]. Furthermore, the speed of convergence from an initial guess $\boldsymbol{x}^{(0)}$ to the ground-truth sparse signal can be *accelerated* by a numerical technique called the *augmented Lagrange multiplier* (ALM) [18]. For $\ell_1$-min, ALM iteratively optimizes both the sparse signal $\boldsymbol{x}$ and the Lagrange multiplier $\boldsymbol{y}$:

$$\min_{\boldsymbol{x},\boldsymbol{e},\boldsymbol{y}} \{\|\boldsymbol{x}\|_1 + \frac{1}{2\mu}\|\boldsymbol{e}\|^2 + \frac{1}{2\lambda}\|A\boldsymbol{x}+\boldsymbol{e}-\boldsymbol{b}\|^2 - \boldsymbol{y}^T(A\boldsymbol{x}+\boldsymbol{e}-\boldsymbol{b})\}, \qquad (13)$$

where $\mu > 0$ is an additional scalar variable. It is easy to see that when $\boldsymbol{y}$ and $\boldsymbol{e}$ are fixed, (13) can be converted to the standard IST problem for $\boldsymbol{x}$ in (12); when $\boldsymbol{x}$ is fixed, since the $\ell_1$-norm $\|\boldsymbol{x}\|_1$ becomes a constant, the objective function becomes smooth and its optimum is trivial to compute.

## 6. Experiment

We measure the performance of the SBC algorithm for face recognition, which is capable of correcting image misalignment in (9) and pixel corruption in (6). In the previous works, it has been demonstrated that if a query image can be properly aligned with the training images, the recognition rate for SBC is high (i.e., above 99% in most normal conditions) [5, 6]. In this paper, we demonstrate the ability of the SBC algorithm in alignment correction.

The experiment uses a public face database called CMU Multi-PIE [19], where a subset of 50 subjects from the database are chosen, each of which is caputred in 20 frontal images under a fixed set of illumination settings. Out of the 20 images for each subject, images $\{0, 1, 7, 13, 14, 16, 18\}$ with extreme illumination conditions are chosen as the training images. We randomly choose one image from the remaining images as the query image for each subject. All images are cropped and down-sampled to $40 \times 30$ pixels.

During the testing stage, a rescaled Baboon image is randomly superimposed in the query image to create an occlusion of about 10% of the pixels. Then the bounding box of the face region is manually perturbed from its ground-truth location by either a translation w.r.t. the $x-y$ axes or an in-plane rotation $\theta$, as shown in Figure 1 Right and Figure 4 respectively. An ALM algorithm [10] is then applied to solve for the transformation parameters $\tau$ and the sparse error $\boldsymbol{e}$ in (9).

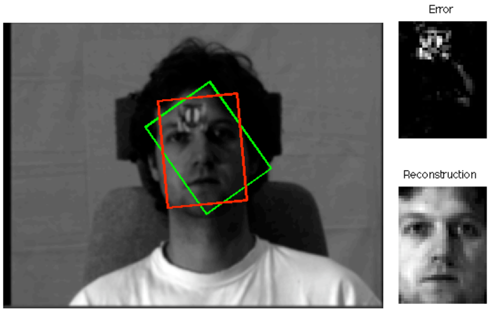

Figure 4: **Left:** Misalignment correction with 10% pixel occlusion (on the forehead) and a 30° in-plane rotation. The green bounding box indicates the simulated alignment perturbation; the red bounding box indicates the alignment result. **Upper Right**: Estimated error $\boldsymbol{e}$ that indicates the location of the corruption. **Lower Right**: Reconstruction result based on $\boldsymbol{x}_i$.

We measure the accuracy of the algorithm in terms of the average error of the pixel coordinates of the eye corners between the ground truth and the estimates. Figure 5 Left shows the estimation error when the test alignment undergoes $x - y$ translations up to ±8 pixels in the canonical frame (with size $40 \times 30$), and Figure 5 Right shows the estimation error when the test alignment undergoes in-plane rotation up to 45°. In general, the algorithm works well with translation within 4 pixels and rotation within 30°. Note that without pixel corruption, the accuracy of the algorithm shall be even higher. Compared to interior-point methods used in [6], ALM also improves the speed for face alignment by 25% on average.

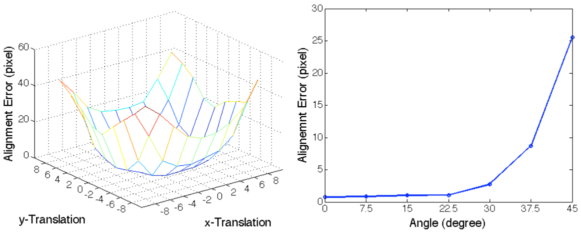

Figure 5: Alignment error w.r.t. $x - y$ plane and $\theta$ rotation.

## 7. Conclusion and Discussion

Based on compressive sensing theory, we have proposed a comprehensive framework/system to tackle the classical problem of face recognition. The success of our solution relies on careful analysis of the special data structure in high-dimensional face images. In addition, to maintain high recognition accuracy in the presence of large illumination change, a novel training image acquisition system has been proposed and patented [6], which uses four projectors to illuminate the subject from all directions. For future topics, we believe one open problem is how to perform small-scale face validation on portable mobile devices (e.g., iPhones and gPhones). Another open problem is how to perform large-scale face detection and recognition in dense urban environments. Parallel implementations of the current algorithms may be needed to support real-time performance of these functions.

## 8. References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neuro.*, vol. 3, no. 1, pp. 71–86, 1991.

[2] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *PAMI*, vol. 27, no. 3, pp. 328–340, 2005.

[3] A. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *PAMI*, vol. 24, no. 6, pp. 748–763, 2002.

[4] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition," *Pat. Ana. App.*, vol. 9, no. 2–3, pp. 273–292, 2006.

[5] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *PAMI*, vol. 31, no. 2, pp. 210 – 227, 2009.

[6] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Toward a practical face recognition: Robust pose and illumination via sparse representation," in *CVPR*, 2009.

[7] E. Candès, "Compressive sampling," in *Pro. Int. Con. Math.*, 2006.

[8] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE TIT*, vol. 52, no. 1, pp. 6–18, 2006.

[9] J. Wright and Y. Ma, "Dense error correction via $\ell^1$-minimization," *IEEE TIT*, (accepted) 2010.

[10] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma, "Fast $\ell_1$-minimization algorithms and an application in robust face recognition: a review," UC Berkeley, Tech. Rep. UCB/EECS-2010-13, 2010.

[11] D. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *PNAS*, vol. 102, no. 27, pp. 9452–9457, 2005.

[12] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *(in press) SIAM Review*, 2007.

[13] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE TSP*, vol. 41, no. 12, pp. 3397–3415, 1993.

[14] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[15] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Stat. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[16] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," in *ICASSP*, 2008.

[17] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[18] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2003.

[19] R. Gross, I. Mathews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *FGR*, 2006.