LEARNING TO SENSE SPARSE SIGNALS: SIMULTANEOUS SENSING MATRIX AND SPARSIFYING DICTIONARY OPTIMIZATION

By

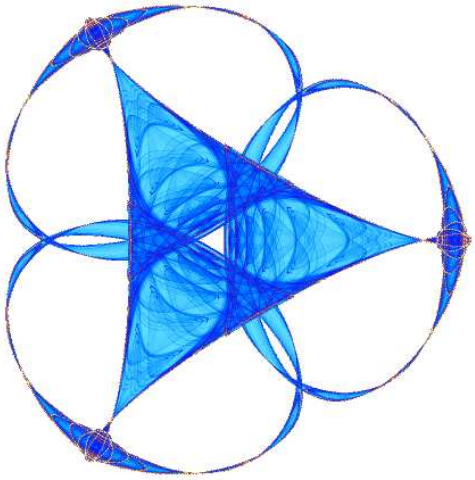
Julio Martin Duarte-Carvajalino

and

Guillermo Sapiro

IMA Preprint Series # 2211

(May 2008)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370
URL: http://www.ima.umn.edu

Learning to Sense Sparse Signals: Simultaneous Sensing Matrix and Sparsifying Dictionary Optimization

Julio Martin Duarte-Carvajalino and **Guillermo Sapiro**
Department of Electrical and Computer Engineering, University of Minnesota

***Abstract*- Sparse signals representation, analysis, and sensing, has received a lot of attention in recent years from the signal processing, optimization, and learning communities. On one hand, the learning of overcomplete dictionaries that facilitate a sparse representation of the image as a liner combination of a few atoms from such dictionary, leads to state-of-the-art results in image and video restoration and image classification. On the other hand, the framework of compressed sensing (CS) has shown that sparse signals can be recovered from far less samples than those required by the classical Shannon-Nyquist Theorem. The goal of this paper is to present a framework that unifies the learning of overcomplete dictionaries for sparse image representation with the concepts of signal recovery from very few samples put forward by the CS theory. The samples used in CS correspond to linear projections defined by a sampling projection matrix. It has been shown that, for example, a non-adaptive random sampling matrix satisfies the fundamental theoretical requirements of CS, enjoying the additional benefit of universality. On the other hand, a projection sensing matrix that is optimally designed for a certain signal class can further improve the reconstruction accuracy or further reduce the necessary number of samples. In this work we introduce a framework for the joint design and optimization, from a set of training images, of the overcomplete non-parametric dictionary and the sensing matrix. We show that this joint optimization outperforms both the use of random sensing matrices and those matrices that are optimized independently of the learning of the dictionary. The presentation of the framework and its efficient numerical optimization is complemented with numerous examples on classical image datasets.**

***Index Terms*— Compressed Sensing, Image Patches, Overcomplete Dictionary, Sensing Projection Matrix, Sparse Representation, Learning.**

I. INTRODUCTION

IMAGE compression algorithms have been successfully employed in the past to transform a high resolution image into a relatively small set of (quantized) coefficients that efficiently represent the image on an appropriate, often orthonormal, basis such as DCT or wavelets. This representation is designed to

preserve the essential content of the image while at the same time reducing costs in storage, processing, and transmission. Since natural images can be compressed on an appropriate basis, sampling the scene into millions of pixels to obtain high resolution images that are then to be compressed before processing, seems often to be wasteful [1]-[11]. The main reason why signals in general and images in particular have been traditionally sensed using a large number of samples is the Shannon-Nyquist Theorem: the sampling rate must be at least twice the bandwidth of the signal. Images are not naturally band limited, however, acquisition systems use anti-aliasing low pass filters before sampling, hence, Shannon-Nyquist Theorem plays an implicit role in images and signals in general.

Compressive sensing (CS) is an emerging framework stating that sparse signals, i.e., signals that have a concise (sparse) representation on an appropriate basis, can be exactly recovered from a number of linear projections of dimension considerably lower than the number of samples required by the Shannon-Nyquist Theorem (in the order of 2-3 times the sparsity of the signal, regardless of the actual signal bandwidth) [1]-[5],[7]. In addition, signals that are well approximated by sparse representations (i.e., *compressible*), such as natural images [12]-[18], can be also sensed by linear measurements at a much lower rate than double their actual bandwidth, as required by the Shannon-Nyquist Theorem, with minimum loss of information [1]-[3]. This means that instead of sensing an image using millions of pixels to obtain high resolution, the image can be sensed directly in compressed form, by sampling a relatively small number of projections that depends on the actual sparsity (and not bandwidth) of the image [1]-[11].[1]

Compressive sensing relies on two fundamental principles, e.g., see the recent review [4]:

• *Sparsity*: Let $\mathbf{x} \in \mathbb{R}^N$ be an N-pixels image and $\mathbf{\Psi} = [\boldsymbol{\psi}_1 \quad ... \quad \boldsymbol{\psi}_N]$ an orthonormal basis (dictionary),[2]

[1] There might be many reasons for still sampling following the traditional Shannon-Nyquist requirements, even for very sparse signals, e.g., the simplicity of the reconstruction of band limited signals from their samples as well as the existence of very efficient sampling and reconstruction hardware and software. We consider CS and the concepts of it exploited in this paper as an alternative and addition to the classical Shannon-Nyquist framework, viable and very useful for many signal and image processing scenarios, but not at all as a replacement for it.

[2] While following the basic theory of CS, we consider for the moment an orthonormal basis, the concept of sparsity is much more general, and best applied in image processing when including overcomplete dictionaries, as exploited in this paper.

with elements also in $\mathbb{R}^N$, such that

$$\mathbf{x} = \sum_{i=1}^{N} \theta_i \boldsymbol{\psi}_i = \boldsymbol{\Psi}\boldsymbol{\theta}, \quad (1)$$

where $\boldsymbol{\theta} = [\theta_1 \quad ... \quad \theta_N]^{\mathrm{T}}$ is the vector of coefficients that represents $\mathbf{x}$ on the basis $\boldsymbol{\Psi}$. A signal or image is said to be sparse if most of the coefficients of $\boldsymbol{\theta}$ are zero or they can be discarded without much loss of "information." Let $\mathbf{x}_S$ be the image where only the *S* largest coefficients of $\boldsymbol{\theta}$ are kept and the rest are set to zero (obtaining $\boldsymbol{\theta}_S$), i.e., $\mathbf{x}_S = \boldsymbol{\Psi}\boldsymbol{\theta}_S$. If the coefficients, sorted in decreasing order of magnitude, decrease quickly, then $\mathbf{x}$ is very well approximated by $\mathbf{x}_S$, when properly selecting both *S* and the basis/dictionary $\boldsymbol{\Psi}$. Such signal is said to be (approximately) *S*-sparse. Natural images are known to be sparse, with *S* significantly lower than the actual image dimension, when represented on an appropriated basis such as wavelets, sinusoids, or a learned (overcomplete) dictionary. Sparse representations form the basis of many successful image processing and analysis algorithms, from JPEG and JPEG2000 compression [19]-[21], to image enhancement and classification, e.g., [12], [22]-[25].

- *Incoherent Sampling*: Let $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \quad ... \quad \boldsymbol{\phi}_m]^{\mathrm{T}}$ be an *m*×*N* sampling matrix, with $m \ll N$, such that $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}$ is an *m*×1 vector of linear measurements (meaning we no longer observe the image $\mathbf{x}$ but an undercomplete linear projection of it). Compressive Sensing theory requires that $\boldsymbol{\Phi}$**,** the sensing matrix, and $\boldsymbol{\Psi}$**,** the sparse representation matrix, be as incoherent (orthogonal) as possible. A measure of coherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ is given by

$$\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) := \sqrt{N} \cdot max_{\substack{1 \le i \le m, \\ 1 \le j \le N}} \left| \langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_j \rangle \right|. \quad (2)$$

  $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) \in \left[1, \sqrt{N}\right]$ measures the maximal correlation between both matrix elements (see also [26], [27] for the related definition of *mutual coherence* of a dictionary, which plays an important role in the success of *basis pursuit* and the greedy sparsifying *orthogonal matching pursuit* algorithm as well [26]-[31]; see below).

CS deals with the case of low coherence between the sensing and sparsifying matrices. Intuitively, one can see that $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ must be uncorrelated (incoherent), such that the samples add new information that is not already represented by the known basis $\boldsymbol{\Psi}$. It turns out that random matrices, e.g., Gaussians or ± 1 random matrices, are largely incoherent with any fixed sparsifying basis $\boldsymbol{\Psi}$ with overwhelming probability. This has lead CS, at least at the theoretical level as well as early applications in image

processing [10], to strongly rely on random sensing matrices, since they provide universally incoherent sensing-sparsifying pairs and are well conditioned for reconstruction.

Compressed sensing combines both concepts of sparsity and incoherence between the sensing and sparsifying matrix by reconstructing the sparsest possible signal that agrees with the undercomplete ($m \ll N$) set of measurements. Let $\mathbf{y}$ be the vector of m linear measurements of the sparse signal $\mathbf{x}$ using the sampling/sensing matrix $\mathbf{\Phi}$. The retrieval of $\mathbf{x}$ from $\mathbf{y}$ can be done by ℓ_0-"norm" minimization (note how both the sensing and the sparsifying matrix appear in the formulation),

$$\widehat{\boldsymbol{\theta}} := min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\ell_0} \text{ subject to } \mathbf{y} = \mathbf{\Phi\Psi\theta}, \qquad \mathbf{x} = \mathbf{\Psi}\widehat{\boldsymbol{\theta}}, \tag{3}$$

where the ℓ_0-"norm" simply counts the number of non-zeros in $\boldsymbol{\theta}$.[3] Compressive Sensing theory ensures (in one of the many recent fundamental results), that if the number of measurements m satisfies

$$m \geq C \cdot \mu^2(\mathbf{\Phi}, \mathbf{\Psi}) \cdot S \cdot \log N, \tag{4}$$

for some positive constant C, then, with overwhelming probability, the reconstruction is exact (even actually using ℓ_1 for sparse promotion instead of ℓ_0, making the problem (3) convex).

As mentioned above, related to $\mu(\mathbf{\Phi}, \mathbf{\Psi})$ is the notion of *mutual coherence* $\mu(\mathbf{\Psi})$ of the sparsifying dictionary (or the equivalent dictionary $\mathbf{D} := \mathbf{\Phi\Psi}$), which is the largest absolute normalized inner product between the atoms of the dictionary (see Equation (7) in the next section for the exact definition). If the following inequality holds [26], [27], [30],

$$\|\boldsymbol{\theta}\|_{\ell_0} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{\Psi})}\right). \tag{5}$$

then, $\boldsymbol{\theta}$ is necessarily the sparsest solution ($min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\ell_0}$) such that $\mathbf{x} = \mathbf{\Psi\theta}$ (or $\mathbf{y} = \mathbf{\Phi\Psi\theta}$ when considering sensing and the equivalent dictionary $\mathbf{D}$), and Orthogonal Matching Pursuit (OMP), a fast greedy algorithm used to solve sparse representation problems, is guaranteed to succeed in finding the correct $\boldsymbol{\theta}$ (same for the basis pursuit, which again replaces the ℓ_0 by ℓ_1[4]), see e.g., [26], [27], [29]-[32]. Note of course that this property on the dictionary, and in contrast with (5), does not explicitly consider the number of samples m.

Since, in general, signals of interest are not exactly sparse, but nearly sparse, and also contain noise added by the measurement system, it is imperative for CS to be robust under such non-idealities [4]. A key

[3] This is not an actual norm, but it is often referred to as such.

[4] The use of the ℓ_1 norm brings interesting connections with robust statistics and regression.

notion in CS theory that comes to the rescue in this scenario is the *Restricted Isometry Property* (RIP) [1]-[4] (and references therein, see also [34] for some results for overcomplete dictionaries). The *S*-restricted isometry constant is the smallest $0 < \delta_S < 1$ such that

$$\forall \, \mathrm{T} \leq \mathrm{S}: \; (1-\delta_S)\|\boldsymbol{\theta}_\mathrm{T}\|_{\ell_2}^2 \leq \|\mathbf{D}_\mathrm{T}\boldsymbol{\theta}_\mathrm{T}\|_{\ell_2}^2 \leq (1+\delta_S)\|\boldsymbol{\theta}_\mathrm{T}\|_{\ell_2}^2,$$

where $\mathbf{D}_\mathrm{T}$ is a subset of T columns extracted from the equivalent dictionary $\mathbf{D} \coloneqq \boldsymbol{\Phi}\boldsymbol{\Psi}$, and $\boldsymbol{\theta}_\mathrm{T}$ are the (sparse) coefficients corresponding to the T selected columns. In words, for proper values of δ_S, the RIP ensures that any subset, with cardinality less than *S*, of columns of the equivalent dictionary $\boldsymbol{\Phi}\boldsymbol{\Psi}$ are nearly orthogonal (the columns cannot be exactly orthogonal since we have more columns that rows), i.e., incoherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ is ensured. If the RIP holds (see the above mentioned references for the exact needed values of the isometry constant), greedy algorithms such as *regularized* OMP, [35], and ℓ_1 convex optimization are guaranteed to succeed (the results for OMP are weaker). This holds as well, again with the possibility to optimize with the ℓ_1 norm instead of ℓ_0, in the presence of noise and for signals with non-exact sparsity [1]-[8]. Hence, Equation (3) becomes in practice,

$$min_{\widehat{\boldsymbol{\theta}}}\|\boldsymbol{\theta}\|_{\ell_0} \text{ subject to } \left\|\mathbf{y}-\boldsymbol{\Phi}\boldsymbol{\Psi}\widehat{\boldsymbol{\theta}}\right\|_{\ell_2} \leq \epsilon, \tag{6}$$

where $\epsilon > 0$ takes into account the possibility of noise in the linear measurements and of non-exact sparsity.

While the sampling projection matrix $\boldsymbol{\Phi}$ should, in theory, be independent of the signal, the sparsifying basis $\boldsymbol{\Psi}$ should adapt as much as possible to the image at hand, e.g., to make the representation as sparse as possible. A key result in image processing is that images can be coded and sparsely represented more efficiently using (often learned) overcomplete dictionaries rather than fixed bases, e.g., [18], [33], [36]. Let $\mathbf{x}_p \in \mathbb{R}^n$, $n \ll N$, be a patch, i.e., a square portion of the image $\mathbf{x}$ of size $B \times B = n$ pixels. An overcomplete dictionary is an $n \times K$ matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \quad ... \quad \boldsymbol{\psi}_K]$, $n < K$, such that $\mathbf{x}_p = \boldsymbol{\Psi}\boldsymbol{\theta}_p$, and for the *K*-dimensional vector of coefficients $\boldsymbol{\theta}_p$, we have $\left\|\boldsymbol{\theta}_p\right\|_{\ell_0} \ll n$. Such dictionaries are learned from patches extracted from large datasets of images, and thereby are optimized for the data (they can also be optimized to a particular image class, e.g., [16], [22]). From the learned non-parametric overcomplete dictionaries, state-of-the-art results in image denoising, inpainting, and demosaicing, have been obtained in the past [12]. This learning framework is also exploited in this work.

Overcomplete learned dictionaries are not orthonormal bases and hence, the full extent of the CS theory

does not entirely apply here (see for example [7], [34] for some results). Recently, Elad [37] (see also [38], [39]), showed experimentally that a well-designed sampling matrix can significantly improve the performance of CS when compared to random matrices, in terms of improving the incoherence for a given dictionary and the reconstruction accuracy from the corresponding linear samples. This means that for a specific pre-defined or pre-learned overcomplete dictionary, a sampling projection matrix specifically designed for the dictionary can indeed improve CS over a generic random sampling matrix. Note that the sampling is not adaptive, is just optimized for the signal class. Theoretical studies, with practical implications, regarding the construction of deterministic sampling matrices and their RIP, are starting to appear, e.g., see [40], [41] and references therein.

In this paper, we introduce a framework for simultaneously learning the overcomplete non-parametric dictionary $\mathbf{\Psi}$ and the sensing matrix $\mathbf{\Phi}$. That is, in contrast with Elad's work and those briefly discussed in the next section, we do not consider a pre-learned or pre-defined dictionary; we learn it together with the sensing matrix. In contrast with the more standard CS framework, we do not assume the sparsifying basis (or dictionary) is given and consider universal sampling strategies, but simultaneously optimize for both these critical components exploiting image datasets. Contrasting also with earlier work on the learning of overcomplete non-parametric dictionaries, we consider linear projections of the image as the available measurements for reconstruction, and not the image itself (or a noisy version of it). We experimentally show that the proposed framework of simultaneous optimization of the sensing matrix and sparsifying dictionary leads to improvements in the image reconstruction results. We also show that the learned sensing matrices have larger incoherence, as requested by the RIP, for a given dictionary, than random matrices and the ones obtained with the algorithm proposed by Elad, on top of leading to lower image reconstruction errors. Computational improvements, when compared with [37], are obtained as well with our proposed framework when considering a given dictionary, and these form the basis of our proposed simultaneous optimization framework.

The remainder of this paper is organized as follows. In Section II, we review Elad's approach [37] (as well as briefly [38] and other recent related works), and introduce our proposed new algorithm to learn deterministic sensing matrices when the sparsifying dictionary is given. We show that the algorithm is significantly faster than the one in [37] and leads to improved performance in terms of sensing-dictionary

incoherence and accuracy of the reconstructed images. In Section III, we review the KSVD algorithm for learning overcomplete non-parametric dictionaries from image datasets, [12], [18], [42], and introduce the novel *coupled-KSVD* as a necessary modification to include the simultaneous learning of both the dictionary and the corresponding sampling projection matrix. In Section IV, we present detailed experimental results indicating the superiority of our new framework to construct deterministic sensing matrices, and the results of simultaneously learning generic sensing matrices and overcomplete dictionaries using datasets of image patches. Finally, concluding remarks and directions for future research are presented in Section V.

While the framework here introduced is applicable to signals in general, from now on, we consider only natural images, and the sparsifying basis will always be overcomplete dictionaries learned from the images.

II. Optimized Projections For Compressive Sampling

In this section we show how to optimize the sensing matrix given a sparsifying dictionary. We start by reviewing prior related work, followed by the presentation of our proposed algorithm and the first results showing the computational and reconstruction advantages of this new approach.

A. Previous Related Work

To the best of our knowledge, only very few recent publications, e.g., [37]-[41], explicitly address the idea that non-random matrices are important and could be more effective than random projection matrices for sensing sparse signals. In particular, [39] shows that chirp-based sampling matrices can be used instead of random sampling matrices, retaining the same reconstruction accuracy, but with the advantage that the retrieval of the original signal becomes computationally much cheaper than using, for example, orthogonal matching pursuit (OMP). On the other hand, [37], [38] address a different idea: random sampling matrices are not necessarily optimal, in the reconstruction error sense, to sample specific classes of sparse signals, and in particular, natural images.

Considering well-known characteristics and models for the second order statistics of natural images, see e.g., [43]-[46], Weiss *et al*., [38], first showed that the Signal to Noise Ratio (SNR) of images projected using (almost any) random sampling matrices goes to zero as the number of pixels increase, while it

remains constant for their proposed deterministic sampling matrices. They then introduced the concept of Uncertain Component Analysis (UCA), that leads to maximize the posterior probability of the data $\mathbf{x}$, for a given projection matrix $\mathbf{\Phi}$ and training projection data $\mathbf{y}$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_P$ be a set of P training patches and $\mathbf{y}_1, \ldots, \mathbf{y}_P$ their respective projections, $\mathbf{y}_i = \mathbf{\Phi}\mathbf{x}_i$, $1 \le i \le P$. Then, the optimal projection matrix, maximizing the probability of retrieving the original patches, is given by $\widehat{\mathbf{\Phi}} := \text{argmax}_{\mathbf{\Phi}, \|\mathbf{\Phi}\|=1} \prod_i \text{P}(\mathbf{x}_i | \mathbf{y}_i; \mathbf{\Phi})$. The prior probability is assumed to be i.i.d. Gaussian, $\text{P}(\mathbf{y}_i | \mathbf{x}; \mathbf{\Phi}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|\mathbf{y}_i - \mathbf{\Phi}\mathbf{x}\|_{\ell_2}^2}{2\sigma^2}}$. Experimentally, the authors found that UCA works only slightly better than random projection on synthetic signals of fixed sparsity, although it works significantly better than random projections on patches of natural images.

Elad [37] introduced a new algorithm that does not make any assumptions on the statistics of the data set, but attempts to improve directly the incoherence between the $m \times n$ sampling matrix $\mathbf{\Phi}$ (m samples of the n dimensional signal),[5] and the $n \times K$ sparsifying dictionary $\mathbf{\Psi}$, as required by the CS theory, assuming a given overcomplete dictionary $\mathbf{\Psi}$ has been already provided (note that [38] is not explicitly based on a dictionary, which in our proposed work below we directly learn from the data). Experimentally, Elad showed, using synthetic random signals of fixed and exact sparsity, that a well-designed projection matrix that depends on the sparsifying basis $\mathbf{\Psi}$, can reduce the mutual coherence of the equivalent dictionary $\mathbf{D} = \mathbf{\Phi}\mathbf{\Psi}$, and hence, reduce the reconstruction error from the projections.

Our work follows in part Elad's idea, since we do not assume any prior knowledge or statistics on the data set, while in contrast with [37], we do not assume a given sparsifying dictionary. In fact, both the overcomplete dictionary, $\mathbf{\Psi}$, and projection matrix, $\mathbf{\Phi}$, are learned from the dataset (see next section for the general case of simultaneous learning). Both [37] and [38] provide a way of computing $\mathbf{\Phi}$ from data or a dictionary, although do not provide any hint on how $\mathbf{\Phi}$ might help the actual learning of the sparsifying basis $\mathbf{\Psi}$, and vice versa.

Our work starts with presenting an alternative to the framework in [37] for learning the projecting matrix from the previously learned dictionary. This alternative, as we show here, is computationally more efficient and produces significantly better results. We therefore start by briefly describing the algorithm in [37]. Motivated by the incoherence required by CS, as well as the fundamental conditions for optimal

[5] We use n for the signal dimension since we will work with patches.

performance of orthogonal matching pursuit and basis pursuit (see previous section), Elad proposed to reduce the *mutual coherence*, $\mu(\mathbf{D})$, of the equivalent dictionary $\mathbf{D} \coloneqq \mathbf{\Phi\Psi}$, $\mathbf{D} = [\mathbf{d}_1 \quad ... \quad \mathbf{d}_K]$, defined as [37]

$$\mu(\mathbf{D}) \coloneqq max_{i \neq j, 1 \leq i,j \leq K} \left\{ \frac{\mathbf{d}_i^{\mathrm{T}} \mathbf{d}_j}{\|\mathbf{d}_i\|_{\ell_2} \|\mathbf{d}_j\|_{\ell_2}} \right\}. \tag{7}$$

Hence, if the projection matrix $\mathbf{\Phi}$ is designed such that $\mu(\mathbf{\Phi\Psi})$ is minimal (recall that $\mathbf{\Psi}$ is fix for the moment), a larger number of signals would satisfy (5) and be successfully recovered from their linear projections.

Instead of minimizing (7), Elad proposes to work with

$$\mu_t(\mathbf{D}) = \frac{\sum_{i \neq j, 1 \leq i,j \leq K} (|g_{ij}| > t) \cdot |g_{ij}|}{\sum_{i \neq j, 1 \leq i,j \leq K} (|g_{ij}| > t)}, \tag{8}$$

where $g_{ij} = \tilde{\mathbf{d}}_i^{\mathrm{T}} \tilde{\mathbf{d}}_j$, $\tilde{\mathbf{d}}_i = \mathbf{d}_i / \|\mathbf{d}_i\|_{\ell_2}$, and t is a scalar that establishes the minimum value of $\mu_t(\mathbf{D})$. From Equation (8), then is obvious that $\mu_t \geq t$. Hence, t is the target value Elad proposes to minimize.

An alternative way of looking at the mutual coherence of the equivalent dictionary $\mathbf{D}$ is to consider the Gramm matrix, $\mathbf{G} \coloneqq \widetilde{\mathbf{D}}^{\mathrm{T}} \widetilde{\mathbf{D}}$, where $\widetilde{\mathbf{D}} = [\tilde{\mathbf{d}}_1 \quad ... \quad \tilde{\mathbf{d}}_K]$ is the equivalent dictionary $\mathbf{D}$ with all its columns normalized. Elad's idea is to minimize the largest absolute values of the off-diagonals in the corresponding Gramm matrix, while keeping the rank of the equivalent dictionary equal to $m << n$. This in turn minimizes μ_t. Instead of targeting $\mu_t(\mathbf{D})$, we address the problem of making any subset of columns in $\mathbf{D}$ as orthogonal as possible, or equivalently, $\mathbf{G}$ should be as close as possible to the identity matrix. We then directly target what we have learned from the RIP, which guarantees robustness of CS to noise and non-exact sparsity. As we show below, our proposed approach outperforms the one introduced in [37], especially for real not-exactly sparse signals (images). Computationally, the proposed algorithm is significantly more efficient than the one in [37]. Moreover, after introducing the proposed algorithm to achieve this close to the identity Gramm matrix, we also introduce the novel idea of simultaneously designing $\mathbf{\Phi}$ and $\mathbf{\Psi}$, which, to the best of our knowledge, has not been addressed before.

B. Learning the Sensing Projection Matrix

As mentioned above, and considering for the moment that the dictionary $\mathbf{\Psi}$ is known, we want to find the sensing matrix $\mathbf{\Phi}$ such that the corresponding Gramm matrix is as close to the identity as possible, i.e.,

$$\mathbf{\Psi}_{Kn}^{\mathrm{T}} \mathbf{\Phi}_{nm}^{\mathrm{T}} \mathbf{\Phi}_{mn} \mathbf{\Psi}_{nK} \approx \mathbf{I}_K.$$

Let us multiply both sides of the previous expression by $\boldsymbol{\Psi}$ on the left and $\boldsymbol{\Psi}^{\mathrm{T}}$ on the right, obtaining

$$\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} \approx \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}.$$

Let $\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ be the (known) eigen-decomposition of $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}$, then $\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \approx \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$, which is equivalent to $\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{V}\boldsymbol{\Lambda} \approx \boldsymbol{\Lambda}$. Let us denote $\boldsymbol{\Gamma}_{mn} \coloneqq \boldsymbol{\Phi}_{mn}\mathbf{V}_{nn}$, hence, $\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Lambda} \approx \boldsymbol{\Lambda}$. We want to compute $\boldsymbol{\Phi}(\boldsymbol{\Gamma})$ in order to minimize

$$\left\|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\right\|_{\mathrm{F}}^{2}. \tag{9}$$

Note that if $\boldsymbol{\Psi}$ is an orthonormal basis and $m = n$ (non standard in CS), then the previous equation would have an exact solution that produces zero error, i.e., $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}^{-\frac{\mathbf{1}}{\mathbf{2}}}$. However, since the dictionary is overcomplete, a critical aspect for achieving high sparsity and state-of-the-art image reconstruction [12], [18], [42], and in particular $m \ll n$ (and then $m \ll rank(\boldsymbol{\Lambda})$**),** we have to find an approximated solution for minimizing the error in Equation (9). We will achieve this starting from a random sensing matrix $\boldsymbol{\Phi}$ (and its corresponding $\boldsymbol{\Gamma}$), and progressively improving it in order to reduce this error.[6] This is detailed next.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the known diagonal matrix $\boldsymbol{\Lambda}$, ordered in decreasing order of magnitude, and $\boldsymbol{\Gamma}_{mn} = [\boldsymbol{\tau}_1 \quad \cdots \quad \boldsymbol{\tau}_m]^{\mathrm{T}}$. Then, Equation (9) becomes $\left\|\boldsymbol{\Lambda} - \sum_{i=0}^{m}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}$, where $\boldsymbol{\nu}_i = [\lambda_1\tau_{i,1} \quad \ldots \quad \lambda_n\tau_{i,n}]^{\mathrm{T}}$, or equivalently,

$$\left\|\boldsymbol{\Lambda} - \sum_{\substack{i=0,\\ i\neq j}}^{m}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^{\mathrm{T}} - \boldsymbol{\nu}_j\boldsymbol{\nu}_j^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}. \tag{10}$$

Let us define $\mathbf{E} \coloneqq \boldsymbol{\Lambda} - \sum_{i=0}^{m}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^{\mathrm{T}}$, $\mathbf{E}_j \coloneqq \boldsymbol{\Lambda} - \sum_{\substack{i=0,\\ i\neq j}}^{m}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^{\mathrm{T}}$, and let $\mathbf{E}_j = \mathbf{U}_j\boldsymbol{\Delta}_j\mathbf{U}_j^{\mathrm{T}}$ be the eigen-decomposition of $\mathbf{E}_j$. Then, Equation (10) becomes $\left\|\mathbf{E}_j - \boldsymbol{\nu}_j\boldsymbol{\nu}_j^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} = \left\|\sum_{k=1}^{n}\xi_{k,j}\mathbf{u}_{k,j}\mathbf{u}_{k,j}^{\mathrm{T}} - \boldsymbol{\nu}_j\boldsymbol{\nu}_j^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}$. If we

[6] Alternatively, we can obtain a closed solution to (9) of the form $\Gamma \coloneqq [\Gamma_1, \Gamma_2]$, where Γ_1 is a diagonal matrix obtained from the top *m* eigenvectors of Λ (elevated to the -1/2 power). While this provides a slightly faster algorithm than the *m*-steps here proposed, it produces virtually the same reconstruction results. Our proposed approach follows the algorithmic concepts of the KSVD and couple-KSVD described in the next section for dictionary learning, and shows how to progressively improve the classical random matrix of CS, also providing one possible solution to addressing the possible ambiguity provided by this closed form alternative Γ. (We thank Donald Goldfarb and Shiqian Ma for pointing out this closed solution and additional comments on the minimization of (9).)

set $\mathbf{v}_j = \sqrt{\xi_{1,j}}\mathbf{u}_{1,j}$, $\xi_{1,j}$ being the largest eigenvalue of $\mathbf{E}_j$ and $\mathbf{u}_{1,j}$ its corresponding eigenvector, then the largest error component in $\mathbf{E}$ is eliminated. Replacing $\mathbf{v}_j$ back in terms of $\boldsymbol{\tau}_j$ (the rows of the matrix we are optimizing for, $\boldsymbol{\Gamma} = \boldsymbol{\Phi}\mathbf{V}$),

$$[\lambda_1\tau_{j,1} \quad \dots \quad \lambda_n\tau_{j,n}]^{\mathrm{T}} = \sqrt{\xi_{1,j}}\mathbf{u}_{1,j}. \tag{11}$$

Since the matrix $\boldsymbol{\Lambda}$ is in general not full-rank, then for some $r \geq 0$, $\lambda_{n-r+1}, \dots, \lambda_n$ will be zero, and we can only update the $\tau_{j,1}, \dots, \tau_{j,n-r}$ components of $\boldsymbol{\tau}_j$. This derivation forms the basis of our algorithm for optimizing (9), see below for the exact steps.

Once we obtain $\hat{\boldsymbol{\Gamma}}$, then $\hat{\boldsymbol{\Phi}}$ can be easily computed following the relationship $\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{\Phi}}\mathbf{V}$, as $\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Gamma}}\mathbf{V}^{\mathrm{T}}$. Also, since we can only reduce $m \ll n$ components of the error matrix $\mathbf{E}$, and the error has a rank lower but close to n (recall that $\boldsymbol{\Psi}$ is almost an orthonormal basis for $\mathbb{R}^n$), then there is no actual hope to completely eliminate the error in (9). The proposed technique aims at reducing the largest m components of this error matrix $\mathbf{E}$.

In summary, the following are the steps of the proposed algorithm for optimizing the sensing matrix given the dictionary (using the notation defined above):

1. Initialize $\hat{\boldsymbol{\Phi}}$, for example, to a random matrix.
2. Find the eigen-decomposition $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathbf{T}}$ and r, the number of non-zero eigenvalues of $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}$.
3. Initialize $\hat{\boldsymbol{\Gamma}} \coloneqq \hat{\boldsymbol{\Phi}}\mathbf{V}$.
4. For j=1 to m
   Compute $\mathbf{E}_j$.
   Find the largest eigenvalue and corresponding eigenvector of $\mathbf{E}_j$, $\xi_{1,j}$ and $\mathbf{u}_{1,j}$.
   Use (11) to update the first r components of $\hat{\boldsymbol{\tau}}_j$ (thereby updating $\hat{\boldsymbol{\Gamma}}$).
5. Compute the optimal $\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Gamma}}\mathbf{V}^{\mathrm{T}}$.

C. Some Preliminary Experimental Results

As we will show in the experimental results, Section IV, this parameter-free algorithm is not only considerably faster than the one introduced in [37], but also significantly improves the reconstruction results and provides a fundamental building block for the simultaneous optimization of the dictionary and sensing matrix. Before these more detailed experimental results, let us now present some illustrative results showing the advantages of the proposed methodology over the algorithm proposed by [37]. For this, we will also consider below the (average) mutual coherence,

$$\mu_g(\mathbf{D}) := \frac{\sum_{i \neq j} g_{ij}^2}{K(K-1)}. \quad (12)$$

This is simply the mean square error that accounts for the off-diagonal elements in the Gramm matrix, while (8) only accounts for the maximum off-diagonal value.

Figure 1 compares, for three different sensing matrices $\mathbf{\Phi}$, the distribution of the absolute value of the off-diagonal elements of the Gramm matrix obtained using a dictionary $\mathbf{\Psi}$ learned from 440 natural images (see Section IV for more details on this dictionary). The three sensing matrices are a Gaussian random sampling matrix (as commonly used in CS), the sensing matrix obtained using Elad's algorithm (with parameters optimized to reduce $\mu_t(\mathbf{D})$; $\gamma = 0.6, t = 20\%$), and our new proposed algorithm.

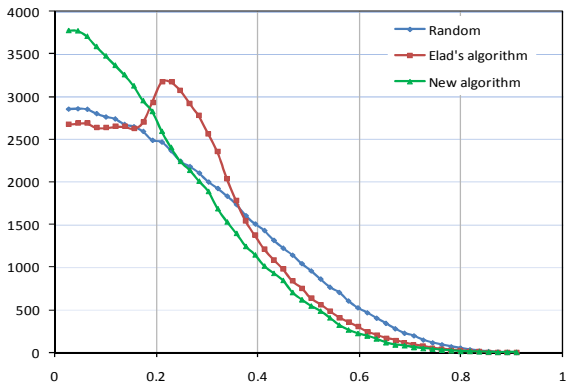


Figure 1: Distribution of the off-diagonal elements of the Gramm matrix obtained using a random sampling matrix, Elad's algorithm, and the new proposed algorithm.

Both our technique and the one proposed in [37] try to reduce the largest off-diagonal elements in the Gramm matrix, however, Elad's algorithm always presents a consistent artifact (see also [37]), where some off-diagonal elements in the Gramm matrix actually increase their value (notice the peak between 0.2 and 0.4 in Figure 1), which does not affect μ_t but it does affect negatively the RIP (see previous Section). Our proposed algorithm increases the frequency of the off-diagonal elements with low absolute value (between 0 and 0.2, Figure 1), and reduces the frequency of large absolute values in the Gramm matrix, better enforcing thus the RIP, since it tries to make the columns of $\mathbf{D}$ as close to orthogonal as possible.

This better behavior of the Gramm matrix (recall the RIP) is reflected in increased accuracy in the image reconstruction as well. Table 1 compares Elad's and our new proposed algorithm to design projection matrices, using patches/signals corresponding to synthetic data of fixed sparsity (as in [37]), patches

coming from real images pre-projected to have a fixed sparsity, and patches coming from real images without restricting their sparsity (compressible patches in contrast to sparse ones). The algorithm in [37] enforces a measure of incoherence, but not directly the RIP and its intuition, and does not actually outperform random sampling, in the mean square error (MSE) sense, when the patches do not have an exact fixed sparsity (see Table 1). The first row in Table 1 corresponds to 10000 synthetic signals obtained by combining at random S=4 columns of a randomly generated 64×256 dictionary. The second row corresponds to 6600 patches projected to have a fixed sparsity, S=6, obtained using OMP to reconstruct the real image patches; the dictionary $\mathbf{\Psi}$ of size 64×256 was trained using these patches and the KSVD algorithm (see next section). The last row corresponds to 6600 patches selected at random from 440 real images and a 64×256 dictionary $\mathbf{\Psi}$ trained with them using KSVD. The table reports average and variance results.

Table 1: Comparison of Elad's vs the new proposed algorithm.

S	m	Patches	Projection	time(s)	μ_t	σ_μ	μ_g	σ_μ	MSE	σ_{MSE}
4	16	Synthetic	Random	-	0.482	0.005	0.078	0.002	2.957	0.071
			Elad'algorithm	794.53	0.434	0.009	0.069	0.002	1.694	0.071
			New Algorithm	0.23	0.421	2.87E-04	0.059	1.37E-05	**1.549**	0.024
6	15	Image, fixed sparsity	Random	-	0.582	0.0139	0.115	0.0062	0.152	0.029
			Elad'algorithm	341.55	0.515	0.0146	0.095	0.0052	0.076	0.006
			New Algorithm	0.22	0.456	6.91E-04	0.063	5.09E-05	**0.022**	0.001
6	15	Real images	Random	-	0.535	0.010	0.097	0.004	0.331	0.028
			Elad'algorithm	426.85	0.487	0.013	0.085	0.004	0.357	0.022
			New Algorithm	0.80	0.442	4.22E-04	0.064	4.69E-05	**0.115**	0.002

From this table we note that $\mu_t(\mathbf{D})$**,** computed using (8), is similar in Elad's and our algorithm, but $\mu_g(\mathbf{D})$ is lower for the proposed algorithm. The reason is that (8) only takes into account the maximum off-diagonal value in the Gramm matrix, ignoring the value of the remaining terms, and hence the artifact introduced by Elad's algorithm (Figure 1). Finally, note the significant MSE improvements obtained with our proposed algorithm.

For illustration purposes, Table 1 also shows the running time of Elad's algorithm versus the proposed new algorithm, on a laptop with a single 1.6 Ghz processor and 1.5 Gb of RAM. Even though the time will change from one implementation and computer to another, this illustrates the significant computational advantage of our proposed technique, Elad's algorithm takes about 600 times longer.

Since Elad's algorithm is not robust in the presence of non-idealities such as non-exact sparsity, and its running time becomes prohibitive for practical applications, from now on when we refer to the non-random sampling matrix $\mathbf{\Phi}(\mathbf{\Psi})$ computed from a dictionary $\mathbf{\Psi}$, we refer to the new algorithm proposed here.

III. Simultaneously Learning the Dictionary and the Projection Matrix

It is now time to turn to the simultaneous learning of the sparsifying dictionary and the sensing matrix. This will be based on combining the just introduced approach for learning the sensing matrix with the KSVD algorithm for dictionary learning. We start then by briefly introducing this KSVD technique.

A. The KSVD algorithm

Recently, Aharon *et al.*, [18], [42] introduced a novel algorithm for learning overcomplete dictionaries to sparsely represent images. Let $\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_p]$ be an $n \times P$ matrix of P training square patches of length n pixels each, used to train an overcomplete dictionary $\mathbf{\Psi}$ of size $n \times K$, with $P >> K$ and $K > n$. The objective of the KSVD algorithm is to solve, for a given sparsity level S,

$$min_{\mathbf{\Psi},\mathbf{\Theta}} \|\mathbf{X} - \mathbf{\Psi\Theta}\|_F^2 \quad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S, \tag{13}$$

where $\mathbf{\Theta} = [\boldsymbol{\theta}_1 \quad \cdots \quad \boldsymbol{\theta}_p]$, and $\boldsymbol{\theta}_i$ is the sparse vector of coefficients representing the i^{th} patch in terms of the columns of the dictionary $\mathbf{\Psi} = [\boldsymbol{\psi}_1 \quad \cdots \quad \boldsymbol{\psi}_K]$. Starting from an arbitrary $\mathbf{\Psi}$, the KSVD algorithm progressively improves it in order to optimize the above expression, as described next.

Let $\mathbf{\Theta} = [\boldsymbol{\delta}_1 \quad \cdots \quad \boldsymbol{\delta}_K]^{\mathrm{T}}$, where $\boldsymbol{\delta}_i^{\mathrm{T}}$ are the rows of $\mathbf{\Theta}$**.** Then, as in the previous section, the error term in (13) can be decomposed as $\left\|\mathbf{X} - \sum_i \boldsymbol{\psi}_i \boldsymbol{\delta}_i^{\mathrm{T}}\right\|_F^2 = \left\|\mathbf{X} - \sum_{i \neq j} \boldsymbol{\psi}_i \boldsymbol{\delta}_i^{\mathrm{T}} - \boldsymbol{\psi}_j \boldsymbol{\delta}_j^{\mathrm{T}}\right\|_F^2$. Let us define $\mathbf{E} \coloneqq \mathbf{X} - \mathbf{\Psi\Theta}$ and $\mathbf{E}_j \coloneqq \mathbf{X} - \sum_{i \neq j} \boldsymbol{\psi}_i \boldsymbol{\delta}_i^{\mathrm{T}}$. Then, (13) can be rewritten as

$$min_{\mathbf{\Psi},\mathbf{\Theta}} \left\|\mathbf{E}_j - \boldsymbol{\psi}_j \boldsymbol{\delta}_j^{\mathrm{T}}\right\|_F^2 \quad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S. \tag{14}$$

At this point it is very tempting to obtain the SVD decomposition of $\mathbf{E}_j$ and eliminate the largest component of the error matrix $\mathbf{E}$ (see previous section). However, (14) requires also satisfying the sparsity constrain. Hence, let us define the set of all indices corresponding to the training patches that use the atom $\boldsymbol{\psi}_j$ for a given (temporary) dictionary $\boldsymbol{\psi}$ (this is determined simply using OMP or any other sparse representation technique, see below), i.e.,

$$\omega_j := \{p | 1 \le p \le P, \boldsymbol{\delta}_j^T(p) \neq 0\},\ 1 \le j \le K. \tag{15}$$

In matrix form, let now $\boldsymbol{\Omega}_j$ be a $P \times |\omega_j|$ matrix with ones on the $(\omega_j(i), i)$ entries and zero elsewhere. Then, (14) can be rewritten as,

$$min_{\boldsymbol{\Psi},\boldsymbol{\Theta}} \|\mathbf{E}_j\boldsymbol{\Omega}_j - \boldsymbol{\psi}_j\boldsymbol{\delta}_j^{\mathrm{T}}\boldsymbol{\Omega}_j\|_F^2, \tag{16}$$

where the sparsity set cannot be altered for now. Let $\mathbf{E}_j^{\mathrm{R}} := \mathbf{E}_j\boldsymbol{\Omega}_j$, $\boldsymbol{\delta}_{j,R}^{\mathrm{T}} := \boldsymbol{\delta}_j^{\mathrm{T}}\boldsymbol{\Omega}_j$, thereby, $\mathbf{E}_j^{\mathrm{R}}$ are just the columns of the error $\mathbf{E}_j$ corresponding to atom $\boldsymbol{\psi}_j$ and $\boldsymbol{\delta}_{j,R}^{\mathrm{T}}$ the rows of $\boldsymbol{\Theta}$, where the zeros have been removed. Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\mathbf{E}_j^{\mathrm{R}}$. Then, (16) becomes

$$min_{\boldsymbol{\Psi},\boldsymbol{\Theta}} \|\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} - \boldsymbol{\psi}_j\boldsymbol{\delta}_{j,R}^{\mathrm{T}}\|_F^2. \tag{17}$$

We now can eliminate the highest component of the error by defining

$$\boldsymbol{\psi}_j := \mathbf{u}_1,\ \ \boldsymbol{\delta}_{j,\mathrm{R}} := \sigma_1\mathbf{v}_1, \tag{18}$$

where σ_1 is the largest singular value of $\mathbf{E}_j^{\mathrm{R}}$ and $\mathbf{u}_1, \mathbf{v}_1$ are the corresponding left and right singular vectors. This then improves the dictionary atom $\boldsymbol{\psi}_j$ based on the patches that have used it when considering the temporary dictionary $\boldsymbol{\Psi}$**.** This continues in the same fashion for all the other columns.

In summary, the KSVD algorithm consists of the following key steps:

1. Initialize $\widehat{\boldsymbol{\Psi}}$
2. Repeat until convergence:
   - For $\widehat{\boldsymbol{\Psi}}$ fixed, solve (13) using OMP to obtain $\widehat{\boldsymbol{\Theta}}$,[7] i.e. $\widehat{\boldsymbol{\Theta}} = \mathrm{OMP}(\widehat{\boldsymbol{\Psi}}, \mathbf{X})$.
   - For j = 1 to K

     Compute ω_j and from there, $\mathbf{E}_j^{\mathrm{R}}, \boldsymbol{\delta}_{j,R}^{\mathrm{T}}$.

     Obtain the largest singular value of $\mathbf{E}_j^{\mathrm{R}}$ and the corresponding singular vectors.

     Update $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\Theta}}$ using (18).[8]

Experimentally, we found that initializing $\boldsymbol{\Psi}$ with an overcomplete dictionary using the Discrete Cosine Transform (DCT) [18] produces better results than with a zero mean, normalized random matrix. The results reported in Section IV use this initialization method.

[7] The original KSVD uses OMP for the sparse coding step, other sparsifying techniques could be used as well.

[8] As in a Gauss-Seidel type of approach, both the atom and the corresponding coefficients are updated at this step.

B. Coupled-KSVD

Let us consider now the problem of simultaneously training a dictionary $\mathbf{\Psi}$ and the projection matrix $\mathbf{\Phi}$, with the images available from a dataset. We define the following optimization problem:

$$min_{\mathbf{\Psi},\Phi,\mathbf{\Theta}} \{\alpha\|\mathbf{X}-\mathbf{\Psi\Theta}\|_F^2 + \|\mathbf{Y}-\mathbf{\Phi\Psi\Theta}\|_F^2\} \quad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S, \tag{19}$$

where $0 \leq \alpha \leq 1$ is a scalar that controls the weight of the error term $\|\mathbf{X}-\mathbf{\Psi\Theta}\|_F^2$ and $\mathbf{Y}$ are the linear samples given by

$$\mathbf{Y} = \mathbf{\Phi X} + \boldsymbol{\eta}, \tag{20}$$

considering $\boldsymbol{\eta}$ an additive noise added by the sensing system. Since $\mathbf{Y}=[\mathbf{y}_1 \quad \cdots \quad \mathbf{y}_P]$ is an $m\times P$ matrix with $m \ll n$, α is used in (19) to compensate for the larger value of the reconstruction error given by the term $\|\mathbf{X}-\mathbf{\Psi\Theta}\|_F^2$, and to give more importance to the projection error term, $\|\mathbf{Y}-\mathbf{\Phi\Psi\Theta}\|_F^2$, which is what is actually available at the reconstruction stage.[9] Notice that (19) can be rewritten as

$$min_{\mathbf{\Psi},\Phi,\mathbf{\Theta}} \left\|\begin{pmatrix}\alpha\mathbf{X}\\ \mathbf{Y}\end{pmatrix} - \begin{pmatrix}\alpha\mathbf{I}\\ \mathbf{\Phi}\end{pmatrix}\mathbf{\Psi\Theta}\right\|_F^2 \quad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S, \tag{21}$$

A possible way to solve (19),(21) consists in extending the KSVD algorithm to the coupled $\mathbf{X}$ and $\mathbf{Y}$ signals, together with the technique to adapt $\mathbf{\Phi}$ to $\mathbf{\Psi}$ introduced in the previous section. As in the KSVD algorithm, we start with an arbitrary dictionary, learn the sensing matrix most appropriate to it following the approach described in the previous section, and then simultaneously improve both of them. Let us define

$$\mathbf{Z} \coloneqq \begin{pmatrix}\alpha\mathbf{X}\\ \mathbf{Y}\end{pmatrix},\ \mathbf{W} \coloneqq \begin{pmatrix}\alpha\mathbf{I}\\ \mathbf{\Phi}\end{pmatrix}. \tag{22}$$

Then, (19),(21) can be rewritten as,

$$min_{\mathbf{\Psi},\Phi,\mathbf{\Theta}} \left\|\mathbf{Z}-\mathbf{D}_{\text{eq}}\mathbf{\Theta}\right\|_F^2 \quad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S, \tag{23}$$

where $\mathbf{D}_{\text{eq}} \coloneqq \mathbf{W\Psi} = \left[\mathbf{d}_1^{\text{eq}} \quad \ldots \quad \mathbf{d}_K^{\text{eq}}\right]$. As in KSVD, we can write

$$\left\|\mathbf{Z}-\sum_i \mathbf{d}_i^{\text{eq}}\boldsymbol{\delta}_i^{\text{T}}\right\|_F^2 = \left\|\mathbf{Z}-\sum_{i\neq j} \mathbf{d}_i^{\text{eq}}\boldsymbol{\delta}_i^{\text{T}} - \mathbf{d}_j^{\text{eq}}\boldsymbol{\delta}_j^{\text{T}}\right\|_F^2. \tag{24}$$

[9] While at the sparsifying dictionary and sensing matrix training step, we have available both the images, **X,** and their projections, **Y**; at the actual reconstruction step we have only the sensed data, **Y**, and the goal is to reconstruct from it the sparsest possible **X**, with the learned ($\mathbf{\Psi}, \mathbf{\Phi}$) (this is the standard CS/sparse-reconstruction scenario, but with our optimized pair ($\mathbf{\Psi}, \mathbf{\Phi}$))**.**

where $\boldsymbol{\delta}_i^{\mathrm{T}}$ are the rows of $\boldsymbol{\Theta}$ as defined previously for the standard KSVD algorithm. Let us now define $\mathbf{E} \coloneqq \mathbf{Z} - \mathbf{D}_{\mathrm{eq}}\boldsymbol{\Theta}$, $\mathbf{E}_j \coloneqq \mathbf{Z} - \sum_{i \neq j} \mathbf{d}_i^{\mathrm{eq}} \boldsymbol{\delta}_i^{\mathrm{T}}$. Considering also $\mathbf{E}_j^{\mathrm{R}}$, $\boldsymbol{\delta}_{j,R}^{\mathrm{T}}$, and $\boldsymbol{\Omega}_j$, as defined in the KSVD algorithm, then Equation (23) can be rewritten as

$$min_{\boldsymbol{\Psi},\Phi,\boldsymbol{\Theta}} \left\| \mathbf{E}_j^{\mathrm{R}} - \mathbf{d}_j^{\mathrm{eq}} \boldsymbol{\delta}_{j,R}^{\mathrm{T}} \right\|_F^2. \tag{25}$$

Similarly, let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\mathbf{E}_j^{\mathrm{R}}$, then (25) becomes

$$min_{\boldsymbol{\Psi},\Phi,\boldsymbol{\Theta}} \left\| \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} - \mathbf{d}_j^{\mathrm{eq}} \boldsymbol{\delta}_{j,R}^{\mathrm{T}} \right\|_F^2, \tag{26}$$

and the highest component of the (coupled) error can be eliminated defining

$$\mathbf{d}_j^{\mathrm{eq}} \boldsymbol{\delta}_{j,R}^{\mathrm{T}} \coloneqq \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathrm{T}}, \tag{27}$$

where σ_1 is the largest singular value of $\mathbf{E}_j^{\mathrm{R}}$ and $\mathbf{u}_1, \mathbf{v}_1$ are the corresponding left and right singular vectors. Now, since $\mathbf{d}_j^{\mathrm{eq}} = \begin{pmatrix} \alpha\mathbf{I} \\ \boldsymbol{\Phi} \end{pmatrix} \boldsymbol{\psi}_j$, Equation (27) is satisfied if

$$\begin{pmatrix} \alpha\mathbf{I} \\ \boldsymbol{\Phi} \end{pmatrix} \widehat{\boldsymbol{\psi}}_j = \mathbf{u}_1, \;\; \boldsymbol{\delta}_{j,R} = \sigma_1 \mathbf{v}_1, \tag{28}$$

where we have $m + n$ equations and n unknowns (the length of the $\boldsymbol{\psi}_j$ atom of dictionary $\boldsymbol{\Psi}$). Note here the importance of coupling the original images $\mathbf{X}$ and their corresponding sensed data $\mathbf{Y}$, if we would have not include $\mathbf{X}$, (28) would have m equations and n unknowns, hence, infinitely many solutions for $\boldsymbol{\psi}_j$ would satisfy (28). In other words, by introducing the regularizing condition that $\|\mathbf{X} - \boldsymbol{\Psi}\boldsymbol{\Theta}\|_F^2$ must also be minimized, we obtain a unique solution to (28) that best fits the training data $\mathbf{X}$ and its projection $\mathbf{Y}$. From (28), $\widehat{\boldsymbol{\psi}}_j$ can be computed using the pseudo-inverse as

$$\widehat{\boldsymbol{\psi}}_j = \left(\alpha^2\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} (\alpha\mathbf{I} \quad \boldsymbol{\Phi}^{\mathrm{T}})\mathbf{u}_1. \tag{29}$$

Since $\widehat{\boldsymbol{\psi}}_j$ computed using (29) does not necessarily have unit ℓ^2-norm, and the columns $\widehat{\boldsymbol{\psi}}_j$ of the learned dictionary $\widehat{\boldsymbol{\Psi}}$ should have unit ℓ^2-norm [42], we redefine $\widehat{\boldsymbol{\psi}}_j$ and $\boldsymbol{\delta}_{j,\mathrm{R}}$ as,

$$\widehat{\boldsymbol{\psi}}_j \leftarrow \widehat{\boldsymbol{\psi}}_j / \left\|\widehat{\boldsymbol{\psi}}_j\right\|_{\ell_2}, \; \boldsymbol{\delta}_{j,\mathrm{R}} \leftarrow \left\|\widehat{\boldsymbol{\psi}}_j\right\|_{\ell_2} \boldsymbol{\delta}_{j,\mathrm{R}}, \tag{30}$$

in order to keep the product $\mathbf{d}_j^{\mathrm{eq}} \boldsymbol{\delta}_{j,R}^{\mathrm{T}}$ on (27) unchanged.

We have now updated the dictionary $\boldsymbol{\Psi}$ and the corresponding sparse coefficients $\boldsymbol{\Theta}$ (repeating, as in the KSVD, the above procedure for all the atoms), considering $\boldsymbol{\Phi}$ fix. The feedback of the sensing matrix into

the update of the dictionary is provided using $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ as defined in Section II, i.e., learning the projection matrix from the just updated dictionary $\boldsymbol{\Psi}$. In turn, $\boldsymbol{\Psi}$ will be affected by $\boldsymbol{\Phi}$, as indicated on (28), in the next iteration. Additionally, OMP also uses the coupled signal $\mathbf{Z}$ to estimate $\boldsymbol{\Theta}$**.** Thereby, the whole learning procedure is simultaneous and exploits all the available data.

In summary, the proposed coupled-KSVD algorithm is,

1. Initialize $\widehat{\boldsymbol{\Psi}}$.
2. Repeat until convergence:
   - For $\widehat{\boldsymbol{\Psi}}$ fixed, compute $\widehat{\boldsymbol{\Phi}}\left(\widehat{\boldsymbol{\Psi}}\right)$ using the algorithm given on Section II.
   - For $\widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\Phi}}$ fixed, solve (21) using OMP to obtain $\widehat{\boldsymbol{\Theta}}$, i.e., $\widehat{\boldsymbol{\Theta}} = \mathrm{OMP}\left(\mathbf{D}_{\mathrm{eq}}, \mathbf{Z}\right)$.
   - For j = 1 to K

     Compute ω_j and from there $\mathbf{E}_j^{\mathrm{R}}, \boldsymbol{\delta}_{j,R}^{\mathrm{T}}$ using (22)-(24).

     Obtain the largest singular value of $\mathbf{E}_j^{\mathrm{R}}$ and the corresponding singular vectors.

     Update $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\Theta}}$ using (28), (29) and (30).

In the next section we evaluate the performance of the improved algorithm to compute the sensing projection matrix $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ from a given sparsifying dictionary $\boldsymbol{\Psi}$**,** and of the coupled-KSVD just introduced, and show their advantage over previously reported techniques.

IV. Experimental Results

In this section, we compare different methods to compute the dictionary, $\boldsymbol{\Psi}$, and sampling matrix, $\boldsymbol{\Phi}$, oriented to retrieve the original image patches from their linear measurements.[10] The methods considered are the classical training of a dictionary using KSVD, coupled with random sampling matrices as commonly used in CS; the proposed improved algorithm to learn $\boldsymbol{\Phi}$ from the already KSVD learned dictionary $\boldsymbol{\Psi}$ (Section II); and the new coupled-KSVD algorithm where we simultaneously learn both $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ from the data available (Section III). In particular, we compare the retrieval error of testing patches $\mathbf{X}$ extracted from real images, and reconstructed using OMP[11] with the equivalent dictionary $\mathbf{D} = \boldsymbol{\Phi}\boldsymbol{\Psi}$

[10] Recall that in the real CS-type scenario, once we have already learned the dictionary and sensing matrix, we have to recover the signal only from its linear projections.

[11] To be consistent with the KSVD-based dictionary/sensing training, we use OMP at the reconstruction step as well. As previously mentioned, we can replace this by other sparsifying techniques.

from their noisy linear samples, $\mathbf{Y} = \mathbf{\Phi X} + \boldsymbol{\eta}$, for the following $(\mathbf{\Phi}, \mathbf{\Psi})$ strategies:

- A dictionary $\mathbf{\Psi}$, learned from the real image data using the standard KSVD, and a Gaussian random sampling matrix $\mathbf{\Phi}$. This represents the standard CS scenario.
- A dictionary $\mathbf{\Psi}$, learned from the real image data using the standard KSVD, and then combined with the optimized sensing matrix $\mathbf{\Phi}(\mathbf{\Psi})$ computed from the dictionary, as indicated on Section II.
- A dictionary $\mathbf{\Psi}$, learned from the data using the coupled-KSVD with a fix Gaussian random sampling matrix $\mathbf{\Phi}$.[12]
- A dictionary $\mathbf{\Psi}$ and sampling matrix $\mathbf{\Phi}(\mathbf{\Psi})$, both learned from the data using the full coupled-KSVD.

The first two strategies are uncoupled, since the dictionary $\mathbf{\Psi}$ is learned using the classical KSVD, independently of $\mathbf{\Phi}$. The third strategy is semi-coupled, since the sampling projection matrix $\mathbf{\Phi}$ affects the learning process of $\mathbf{\Psi}$ through the samples $\mathbf{Y}$ (Section III), but not vice versa. The fourth strategy is completely coupled, since both $\mathbf{\Phi}$ and $\mathbf{\Psi}$ affect each other during the learning process: $\mathbf{\Phi}$ affects the learning of $\mathbf{\Psi}$ through the samples $\mathbf{Y}$**,** and in turn, $\mathbf{\Psi}$ affects the sampling matrix, since $\mathbf{\Phi}$ depends on $\mathbf{\Psi\Psi}^{\mathrm{T}}$ (Section II). In the following we refer to each strategy as uncoupled random (UR), uncoupled learning (UL), coupled random (CR), and coupled learning (CL), respectively.

The training data consists of 6600 8×8 patches obtained by extracting at random 15 patches from each one of the 440 images in the training set (250 images from the Berkeley segmentation data set [47] and 190 images from the *Labelme* data set [48]). The testing data consists of 120000 patches corresponding to all the non-overlapping patches of size 8×8 extracted from the remaining 50 images in the Berkeley dataset that are not in the training set.

The different strategies (UR, UL, CR, and CL), are evaluated in terms of the MSE of retrieval, defined as $\mathrm{MSE} := \|\mathbf{X} - \mathbf{\Psi\Theta}\|_F^2$, where $\mathbf{\Psi}$ is the dictionary learned from the training patches, $\mathbf{X}$ is the matrix of testing patches, and $\mathbf{\Theta}$ is obtained using OMP to solve

$$min_{\widehat{\mathbf{\Theta}}} \left\|\mathbf{Y} - \mathbf{\Phi\Psi}\widehat{\mathbf{\Theta}}\right\|_F^2 \qquad s.t.\ \forall i, \|\boldsymbol{\theta}_i\|_{\ell_0} \leq S, \tag{31}$$

[12] This means that we incorporate the sensing in the learning of the dictionary, but do not update the sensing matrix and keep it constant during the iterations of the coupled-KSVD algorithm.

i.e., $\widehat{\boldsymbol{\Theta}} = \mathrm{OMP}(\mathbf{D}, \mathbf{Y})$, where, $\boldsymbol{\Phi}$ is a random sampling matrix for the UR and CR strategies and a learned sampling matrix for the UL and CL strategies, and $\mathbf{Y}$ is a noisy version of the projected patches, as defined in (20). The noise added to the samples ranges from 0 to 25% in amplitude, for each patch, with increments of 5%. The parameter α in the coupled-KSVD was varied in the set $\left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}\right\}$. We use typical values for the other algorithm parameters: sparsity S = 4, 5, and 6; sampling dimension m = 2S, 2.5S, and 3S; patch dimension n =64 (8×8); and overcompleteness of K = 4n. These values are commonly used in learning overcomplete dictionaries and Compressive Sensing (see, for instance [1]-[8], [12], [18], [42]). We include here representative results from this large set of possible parameter combinations, see the supplementary material for numerous additional graphs and tables.

Figure 2 compares the average MSE of retrieval for the testing patches using S=6, m=12, n=64, and K=256; at different values of α and noise level, for the four training strategies. We clearly observe the significant advantage of learning the sensing matrix (coupled or uncoupled from the dictionary) over the more standard use of random projections.

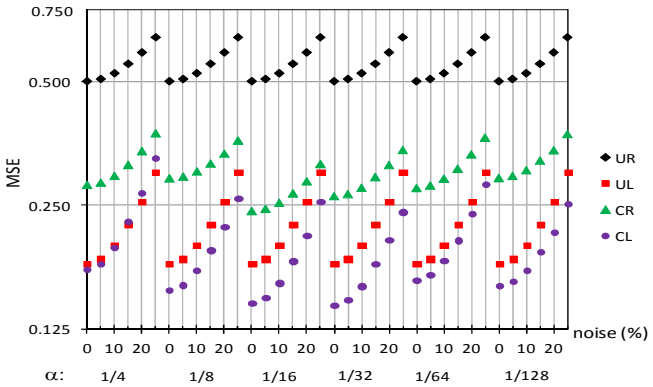


Figure 2: Retrieval MSE using the four training strategies at different noise levels and values of α, for S=6, m =12, n = 64, and K=256.

Figure 3a shows the retrieval MSE of CL relative to CR. Note that the MSE using coupled learning (CL) is almost 50% (a reduction of ~3 Db) of the MSE using semi-coupled learning, with a random projection matrix (as common in standard CS). The difference between CR and CL reduces as the noise level increases. Nevertheless, for noise levels below 20%, the MSE of CL is at least 30% lower than CR for α > ¼. Figure 3b shows the retrieval MSE of CL relative to UL. Here, the advantage of CL over UL is lower than in the previous case (Figure 3a), which indicates that a well-designed projection matrix as introduced in Section II, learned from the dictionary $\boldsymbol{\Psi}$, can do better than simply coupling the data using

a random sampling matrix (see also Figure 2). However, CL can still reduce the MSE with respect to UL as much as 20% (a reduction of ~1Db), which justifies its use.

From Figure 2 and Figure 3 it is also clear that the retrieval MSE and CL/CR and CL/UL ratios have a minimum at about $\alpha = \frac{1}{32}$. We comment more on this at the end of this section.

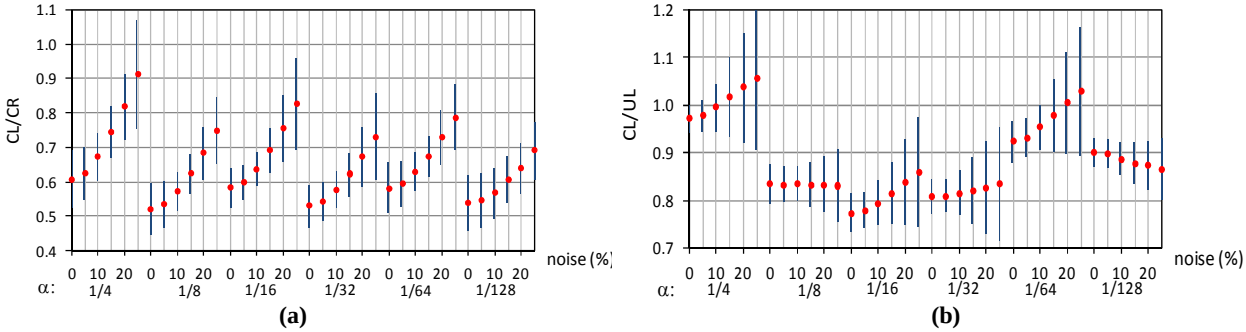


Figure 3: For S=6, m =12, n = 64, and K=256, different noise levels and values of α, a) Ratio between the retrieval MSE for CL and the retrieval MSE for CR, b) Ratio between the retrieval MSE for CL and the retrieval MSE for UL.

Figure 4 and Figure 5 show the retrieval MSE and the CL/CR, CL/UL ratios for *K*= 64, i.e., for a dictionary that is also a basis for the vector space of image patches. This is an interesting experiment, since the dictionary now is not overcomplete, being more in agreement with the majority of the theoretical results from the CS framework. These figures show that the proposed framework is also valid within this scenario, and as can be appreciated in these figures, the CL/CR and CL/UL ratios are even better than for the overcomplete case (Figure 2 and Figure 3), indicating that the proposed coupled learning of both **Ψ** and **Φ** can be even more influential with non-overcomplete dictionaries.

Due to space limitations, we cannot show here the results of all our experiments with all possible parameter variations (again, see supplementary material). However, a set of representative results is shown in Table 2, which indicates the best values of α that produced the minimum retrieval MSE and at the same time the best CL/CR and CL/UL ratios, for a representative noise level of 5%. In general, for overcomplete dictionaries with $K = 4n$, we found that the best values for α are $\frac{1}{32}, \frac{1}{16}$, and $\frac{1}{8}$ for $m = 2S$, $2.5S$, and $3S$, respectively, which indicates that as the number of samples is reduced, **Y** must have greater importance than **X** in the optimization. However, as detailed before, **X** has an important role in the coupling algorithm, to limit the number of possible solutions of the under-determined inversion problem with $m << n$, and given that in practice $m \geq 2S$, then we must have $\alpha > 0$.

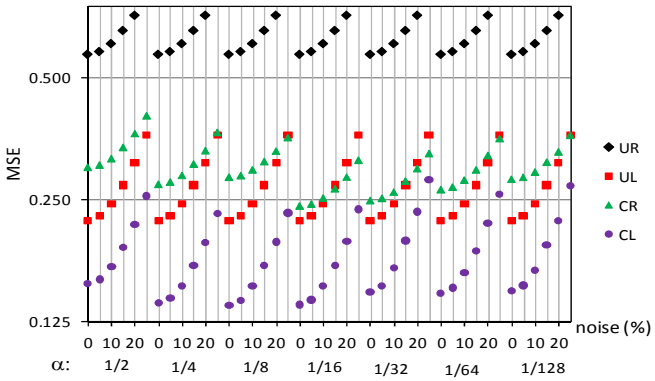
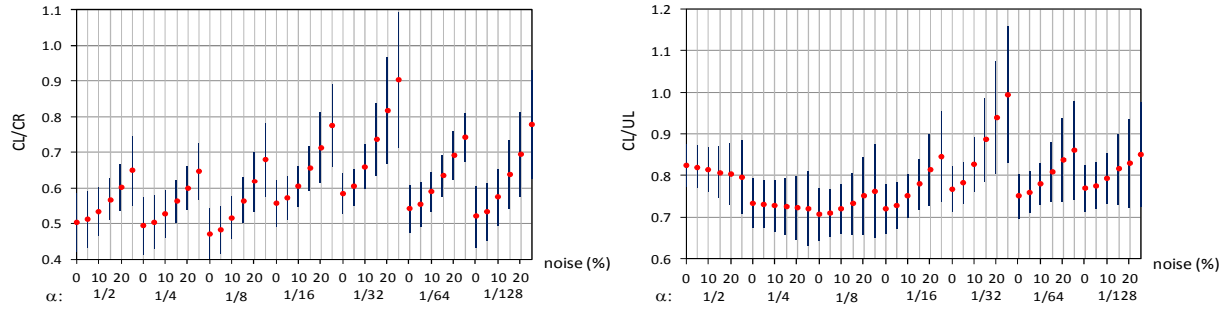


Figure 4: MSE of retrieval from projections using the four training strategies at different noise levels and values of α, for S=6, m =12, n = 64, and K=64.



(a) **(b)**

Figure 5: For S=6, m =12, n = 64, and K=64, different noise levels and values of α, a) Ratio between the MSE of retrieval from projections for CL and MSE of retrieval for CR, b) Ratio between the MSE of retrieval from projection for CL and the MSE of retrieval for UL.

Table 2: Sample results for the best values of α for a noise level of 5%.

K	S	m	α	MSE				CL/CR		CL/UL	
				UR	UL	CR	CL	mean	std	mean	std
256	4	8	1/32	0.5306	0.2154	0.3872	0.1890	0.4823	0.0880	0.8805	0.0317
		10	1/16	0.4980	0.1981	0.3113	0.1820	0.5746	0.0633	0.9215	0.0285
		12	1/8	0.4754	0.1838	0.2605	0.1731	0.6578	0.0451	0.9402	0.0266
	5	10	1/32	0.5029	0.1926	0.2827	0.1646	0.5720	0.0622	0.8497	0.0308
		12	1/16	0.4781	0.1782	0.2708	0.1614	0.5888	0.0626	0.9193	0.0235
		15	1/8	0.3951	0.1626	0.2303	0.1456	0.6267	0.0381	0.9092	0.0292
	6	12	**1/32**	0.5094	0.1848	0.2659	0.1465	**0.5436**	**0.0569**	**0.8090**	**0.0354**
		15	1/16	0.4203	0.1633	0.2254	0.1375	0.6016	0.0452	0.8399	0.0273
		18	1/8	0.3974	0.1477	0.2033	0.1287	0.6294	0.0261	0.8570	0.0266
64	6	12	**1/8**	0.5801	0.2275	0.2865	0.1409	**0.4814**	**0.0661**	**0.7107**	**0.0569**
		15	1/4	0.4601	0.1929	0.2241	0.1241	0.5483	0.0499	0.7357	0.0521
		18	1/2	0.3994	0.1838	0.2210	0.1243	0.5568	0.0400	0.8026	0.0314

For illustration purposes, Figure 6 shows one testing image consisting of non-overlapping 8×8 patches reconstructed from their noisy projections (5% level of noise) $\mathbf{Y}$ as $\hat{\mathbf{X}} = \boldsymbol{\Psi}\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = \mathrm{OMP}(\mathbf{D}, \mathbf{Y})$, and $\boldsymbol{\Psi}, \boldsymbol{\Phi}$ are obtained using either UR, UL, CR, or CL training strategies. The worst reconstruction case (Figure 6a) is obtained when $\boldsymbol{\Psi}$ is learned using classical KSVD and $\boldsymbol{\Phi}$ is simply a random sampling matrix (UR, standard CS scenario), followed by coupled-KSVD using a random sampling matrix (CR,

Figure 6c). Using a well-designed sampling matrix $\boldsymbol{\Phi}(\boldsymbol{\Psi})$ (UL, Section II) produces a good looking reconstruction of patches from their noisy projections (Figure 6b). Finally, using the coupled-KSVD (CL, Section III), an even better reconstruction from the noisy projections is obtained (Figure 6d). The better quality of Figure 6d over Figure 6b can be appreciated by the reduction of the artifacts, especially around the sharp edges on the top of the castle. Additional examples for the UR (standard CS framework) and the proposed CL technique are provided in Figure 7 (additional details on these figures, including the CR and UL cases, are included with the supplementary material).

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

Figure 6: One test image reconstructed from projected patches, without overlapping, with a noise level of 5%, using a) Uncoupled Random, b) Uncoupled Learning, c) Coupled Random, and d) Coupled Learning strategies to learn the dictionary and the projection matrix from the training patches. The retrieval MSE for these images is a) 1.1528, b) 0.4548, c) 0.6721, and d) 0.3769.

We have just presented the improved performance in terms of MSE and quality of the reconstructed patches, when using a well designed sampling matrix instead of random projection matrices, and also when exploiting coupled over uncoupled learning. Let us now compare the dictionaries learned from classic KSVD vs. coupled-KSVD with a random matrix and with a learned sampling matrix. We also want to compare the random projection matrices vs. the sampling matrices computed using the proposed uncoupled and coupled learning techniques. Let us start by comparing the equivalent dictionaries $\mathbf{D} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ for the four strategies considered here, in terms of the closeness of the Gramm matrix $\mathbf{G} = \widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}$ (see Section II) to the identity (as inspired by the RIP).

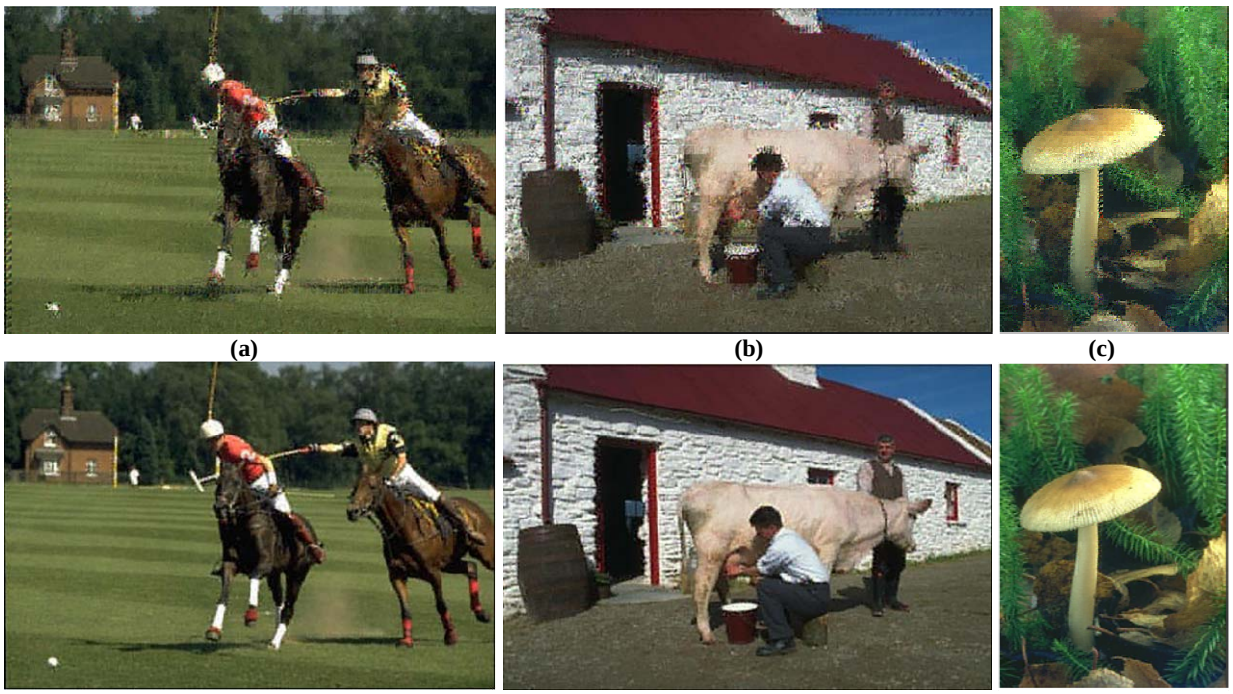


Figure 7: Additional examples of image recovery. Note the sharp improvement with our proposed CL framework (bottom row) when compared to the more classical UR scenario from CS (top row). The retrieval MSE for these images is a) 1.0539, b) 2.3455, c) 0.8207, d) 0.2389, e) 0.6707, and f) 0.2204. While here we sample at twice the sparsity rate, even sampling at four times the sparsity, the UR results are far from the CL ones at just twice the sparsity, both in visual quality and MSE (e.g., the MSE for the image in b) becomes 1.1707, while better than the 2.3455, still more than double the 0.6707 MSE obtained with the proposed approach at half the sampling rate).

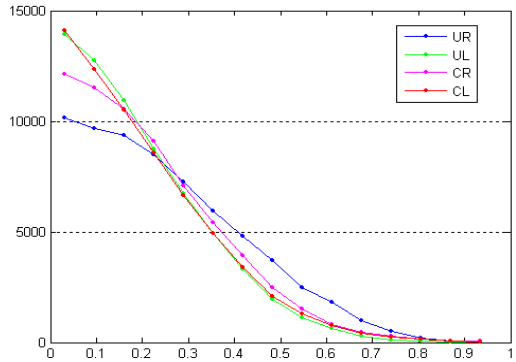


Figure 8: Distribution of the off-diagonal elements of the Gramm matrix obtained using UR, UL, CR and CL strategies.

Figure 8 shows the distribution of the off-diagonal elements of the Gramm matrix for each one of the strategies. The Gramm matrix of CR is closer to the identity than the Gramm matrix of UR, and the Gramm matrix of both UL and CL are closer to the identity than UR and UL respectively, but they are almost undistinguishable among themselves in terms of the distribution of the off-diagonal elements.

These results are in agreement with the improved performance observed for the UL and CL methods over the other two possible strategies that use random sampling matrices.

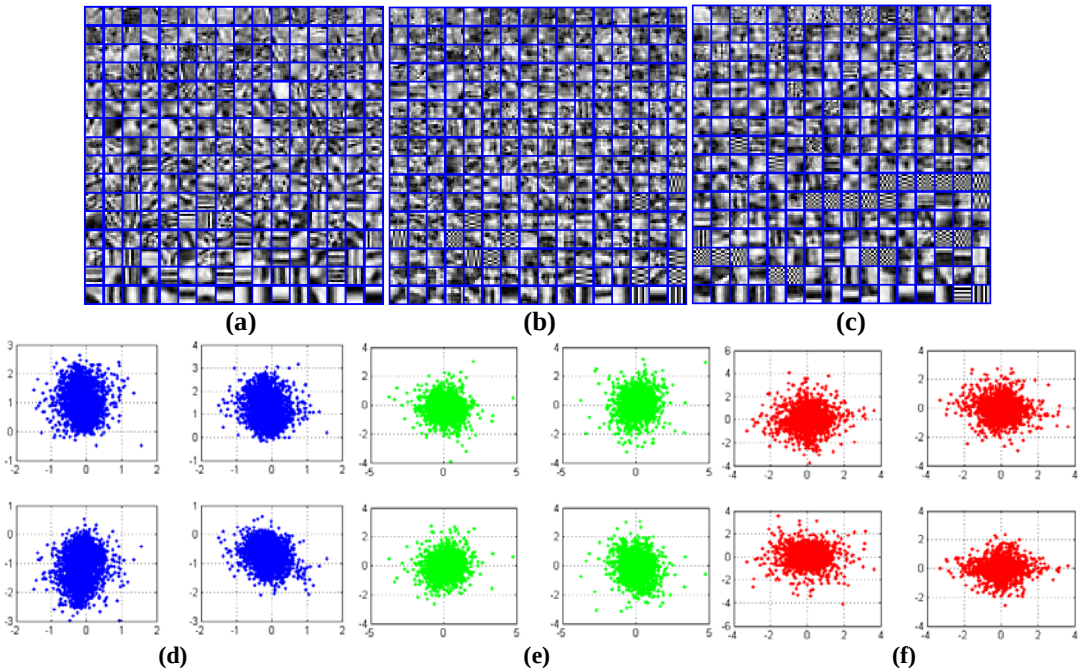


Figure 9: Visual representation of the learned dictionaries (above) and projection matrices (below) for the different learning strategies.

Figure 9 a), b) and c) show the K=256 atoms of the dictionaries learned with KSVD (used for both UR and UL), coupled-KSVD with a random sampling matrix (CR), and coupled KSVD with a learned sampling matrix (CL), respectively, represented conveniently here as $K = 16 \times 16$ images of size 8×8. We clearly see that the learned dictionaries are different. In order to compare the sampling matrices $\mathbf{\Phi}$ for the different learning strategies considered here, we use the same approach proposed by [38], i.e., good sampling matrices should produce signals that are as spread as possible in $\mathbb{R}^m$. Figure 9 d), e) and f) show scatter plots of the first row of $\mathbf{Y} = \mathbf{\Phi X}$ vs. rows two to five (for random, uncoupled following Section II, and coupled following Section III, respectively). It can be noticed here that UL produces samples $\mathbf{Y}$ with a larger spread in $\mathbb{R}^m$ than UR, and in turn CL produces samples that are more spread in $\mathbb{R}^m$ than UL (notice on Figure 9 d), e) and f) the change of scale on some scatter plots). This behavior also occurs for all the possible scatter plots among rows in $\mathbf{Y}$, but for limitations of space, we only present here the first four.

From Figure 9 one observes that besides making the Gramm matrix closer to the identity, the learned dictionaries should also be able of learning new patterns present only in the projected signals (and not in

the signal itself), and coupled-KSVD helps to introduce those new patterns. In addition, a well-designed sampling matrix can improve the spread of the projected signals, when compared to a random sampling matrix, which in turn improves the retrieval of the original signal from projections, thanks to the larger separability of those signals in $\mathbb{R}^m$. More formally, $\mathbf{\Phi}$ should maximize the mutual information $I(\mathbf{X},\mathbf{Y})$, between the signal $\mathbf{X}$ and its noisy projection $\mathbf{Y}$, which is equivalent to maximize the entropy of the output, $H(\mathbf{Y})$, in order to minimize the retrieval error of $\mathbf{X}$ from its noisy projections $\mathbf{Y}$ [38], [49].

V. Concluding Remarks

A computational framework for learning an optimal sensing matrix for a given sparsifying dictionary was introduced in this paper. This was complemented by a novel approach to simultaneously learn the sensing matrix and sparsifying dictionary from an image database. We showed that such learning leads to significantly improved reconstruction results when compared with more classical compressed sensing scenarios where random sensing matrices are used. The same framework can be used to learn the sparsifying dictionary while keeping the sensing matrix fix (see also [50]).

As mentioned in the introduction, the theoretical results for CS support the use of $\ell_1$optimization, while KSVD-type of algorithms have traditionally been based on OMP (for which the results are weaker). It is thereby important to further improve the results here presented using ℓ_1-based optimization approaches.

The framework here developed is based on image patches, as commonly exploited in image processing. While in principle we could work with entire images, this is computationally unfeasible. For tiny 32×32 images, following [51], we obtain results consistent with the work reported above for the patches, see Figure 10. Of course, images are much larger than this, and algorithms of the type of KSVD as here developed, or basically any dictionary learning approach, are virtually impossible. On the other hand, following once again the state of the art results for image enhancement via KSVD, we should work with overlapping patches (e.g., 8×8 or multiscale up to 20×20, see [52]). Ideally, we would like then to have the dictionary acting on all the overlapping $n \times n$ patches, with a unique sensing matrix globally acting on the $N \times N, n \ll N$, image. This will also permit to naturally include the multiscale framework developed in [52]. Results in this direction, as well as in the adaptation of the sensing to the task following [22], will be reported elsewhere.

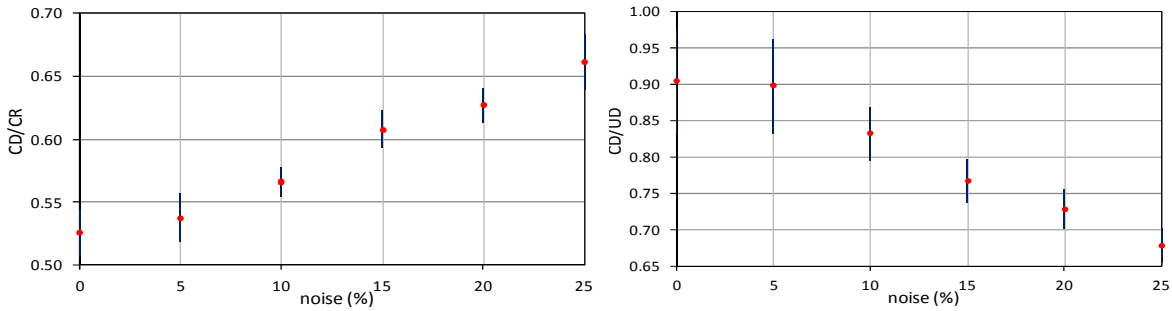


Figure 10: Sample MSE results for tiny images (left CL/CR, right CL/UL), which are consistent with those reported before for patches ($S = 125, m = 250, \alpha = 1/32$).

ACKNOWLEDGMENTS

The authors express their gratitude to Julien Mairal for providing his very efficient code implementing KSVD in C++, which served as our basis to develop the code used in our experiments. Also we wish to thank Prof. Michael Elad for making publicly available some useful code in Matlab [53]. The authors are partially supported by NSF, ONR, NGA, DARPA, ARO, and NIH.

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 5, pp. 489–509, Feb. 2006.

[2] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203- 4215, Dec. 2005.

[3] E. J. Candès, "Compressive sampling," in *Proc. Intl. Congress of Math.*, Madrid, Spain, 2006, pp. 1433-1452.

[4] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21-30, March 2008.

[5] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 14-20, March 2008.

[6] J. Haupt and R. Nowak, "Compressive sampling vs. conventional imaging," in *IEEE Intl. Conf. on Image Processing*, Atlanta, GA, 2006, pp. 1269-1272.

[7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289-1306, April 2006.

[8] Compressive Sensing resources. [Online]. Available: http://www.dsp.ece.rice.edu/cs/.

[9] Compressive Sensing group, "Compressive Imaging: A New Single Pixel Camera". [Online].

Available: http://www.dsp.ece.rice.edu/cs/cscamera/.

[10] M. F. Duarte, M.A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, March 2008.

[11] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," preprint, 2006.

[12] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53-69, Jan. 2008.

[13] S. Mallat and E. Le Pennec, "Sparse geometric image representation with bandelets," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 423–438, April 2005.

[14] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?" in *Proc. IEEE Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.

[15] K. Engan, S. O. Aase, and J. H. Husøy, "Frame based signal compression using method of optimal directions (MOD)," in *IEEE Intern. Symp. Circ. Syst.*, Orlando, FL, 1999, vol. 4, pp. 1-4.

[16] G. Peyre, "Sparse modeling of textures," Preprint Ceremade, 2007-15. [Online]. Available: http://www.ceremade.dauphine.fr/~peyre/publications/07-Preprint-Peyre-SparseModelingTextures.pdf

[17] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.

[18] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[19] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. New York, NY: Academic Press, 1999.

[20] M. Do and M. Vetterli, "Framing pyramids," *IEEE Trans. Image Process.*, vol. 51, no. 9, pp. 2329–2342, Sep. 2003.

[21] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proc. Data Compression Conference*, Snowbird, UT, 2000, pp. 523–544.

[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in Proc. *IEEE Computer Vision Pattern Recognition*, Anchorage, AK, June 2008.

[23] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Computer Vis. Pattern Recog.*, New York, NY, 2007, pp. 1-8.

[24] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," to appear, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.

[25] A. Battle, H. Lee, B. Packer and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in Proc. 24th *Intl. Conf. Machine Learning*, Corvalis, OR, 2007, pp. 759-766.

[26] D. L. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via l_1 minimization," in Proc. *Nat. Aca. Sci.*, vol. 100, no. 5, pp. 2197-2202, 2003.

[27] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320-3325, Dec. 2003.

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

[29] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annual Asilomar Conf. Signals, Systems, and Computers*, Los Alamitos, CA, 1993, vol. 1, pp. 40-44.

[30] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231-2242, Oct. 2004.

[31] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33-61, 1998.

[32] J. A. Tropp and A. C. Gilbert, "Signal recovery from partial information via Orthogonal Matching Pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655-4666, Dec. 2007.

[33] F. Murray and K. Kreutz-Delgado, "Sparse image coding using learned overcomplete dictionaries," in *IEEE Proc. Machine Learning for Signal Proc.*, Sao Luis, Brazil, 2004, pp. 579-588.

[34] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Information Theory*, to appear.

[35] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations Computational Mathematics*, to appear.

[36] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *J. Visual Communication and Image Representation*, vol. 19, no. 4, pp. 270-283, May 2008.

[37] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695-5702, Dec. 2007.

[38] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning compressed sensing," in *The Snowbird Learning Workshop*, Allerton, CA, 2007. [Online]. Available: http://www.cs.huji.ac.il/~yweiss/allerton-final.pdf

[39] L. Applebaum, S. Howard, S. Searle, and R. Calderbank, "Chirp sensing codes: deterministic compressed sensing measurements for fast recovery," preprint, 2008. [Online]. Available: http://www.dsp.ece.rice.edu/cs/.

[40] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. of Complexity*, vol. 23, no. 4-6, pp. 918-925, Aug. 2007.

[41] V. Chandar, "A negative result concerning explicit matrices with the restricted isometry property," preprint 2008. [Online]. Available: http://www.dsp.ece.rice.edu/cs/.

[42] M. Aharon, M. Elad, and A. Bruckstein, "The KSVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[43] E.P. Simoncelli, "Statistical models for images: compression restoration and synthesis," in *Proc. Conf. on Signals, Systems and Computers*, Asilomar, CA, 1997, pp. 673–678.

[44] D. Mumford and B. Gidas, "Stochastic models for generic images," *Quarterly of Applied Mathematics*, vol. 54, no. 1, pp. 85-111, March 2001.

[45] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17-33, Jan. 2003.

[46] J. Huang and D. Mumford, "Statistics of natural images and models," in *IEEE Conf. Computer Vision and Pattern Recognition*, Fort Collins, CO, 1999, vol. 1, pp. 1541-1547.

[47] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE 8th Int. Conf. Computer Vision*, Vancouver, Canada, 2001, vol. 2, pp. 416-423.
Data set Available: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/.

[48] B. Russell, A. Torralba, and W. T. Freeman, "Labelme, the open annotation tool," MIT, Computer Science and Artificial Intelligence Laboratory. [Online]. Available: http://labelme.csail.mit.edu/.

[49] J. F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Lett.* vol. 4, no. 4, pp. 112-114, Apr. 1997.

[50] G. Peyré, "Best basis compressed sensing," Preprint Ceremade 2007-20. [Online].
Available: http://www.ceremade.dauphine.fr/~peyre/publications/07-Preprint-Peyre-BestBasisCS.pdf.

[51] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large data set for non-parametric object and scene recognition," Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, Rep. MIT-CSAIL-TR-2007-024. [Online].
Available: http://people.csail.mit.edu/torralba/tmp/tiny.pdf.

[52] J. Mairal, G. Sapiro, and M. Elad "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Modeling and Simulation*, vol. 7, no. 1, pp. 214-241, April 2008.

[53] M. Elad, personal page. [Online]. http://www.cs.technion.ac.il/~elad/software/.