Approximation and Compression With Sparse Orthonormal Transforms

Osman Gokhan Sezer, *Member, IEEE*, Onur G. Guleryuz, *Member, IEEE*, and Yucel Altunbasak, *Fellow, IEEE*

***Abstract*— We propose a new transform design method that targets the generation of compression-optimized transforms for next-generation multimedia applications. The fundamental idea behind transform compression is to exploit regularity within signals such that redundancy is minimized subject to a fidelity cost. Multimedia signals, in particular images and video, are well known to contain a diverse set of localized structures, leading to many different types of regularity and to nonstationary signal statistics. The proposed method designs sparse orthonormal transforms (SOTs) that automatically exploit regularity over different signal structures and provides an adaptation method that determines the best representation over localized regions. Unlike earlier work that is motivated by linear approximation constructs and model-based designs that are limited to specific types of signal regularity, our work uses general nonlinear approximation ideas and a data-driven setup to significantly broaden its reach. We show that our SOT designs provide a safe and principled extension of the Karhunen–Loeve transform (KLT) by reducing to the KLT on Gaussian processes and by automatically exploiting non-Gaussian statistics to significantly improve over the KLT on more general processes. We provide an algebraic optimization framework that generates optimized designs for any desired transform structure (multiresolution, block, lapped, and so on) with significantly better n-term approximation performance. For each structure, we propose a new prototype codec and test over a database of images. Simulation results show consistent increase in compression and approximation performance compared with conventional methods.**

***Index Terms*— Sparse orthonormal transforms, sparse lapped transforms, sparse multi-resolution transforms, transform optimization, image compression, nonlinear approximation, machine learning, linear representation.**

I. Introduction

DATA representation with linear transforms is extensively used in many signal processing, reconstruction and compression applications [22], [46]. For a given class of signals, efficient transforms are those that enable close approximations of the signals using a small number of transform coefficients. Efficient transforms naturally lead to sparse representations of signals, allowing one to view, manipulate, and compress data using fewer degrees of freedom.

As the efficiency of a transform varies over classes of signals, methodologies that design good transforms for a given family of signals are highly sought after. Available transform design techniques can be viewed in terms of two categories. Model-based design techniques assume a specific type of regularity within the data samples and build analytical models of sample variations in localized neighborhoods. By using certain smoothness characteristics and approximation constructs, model-based methods try to condense signal variations into a few transform coefficients. Fourier transforms, wavelet transforms, the recent curvelet, bandelet, and contourlet transforms are well-known designs from this category [4], [5], [8], [14], [28], [33]. A nice property of model-based designs is that their optimality can be shown analytically. For example the Fourier transform can be shown to be optimal over stationary Gaussian signals, wavelet transforms over piecewise-smooth signals with point singularities, and the cited $*$-lets on piecewise smooth signals with discontinuities over curves [27].

The second category of design techniques is comparatively more generic and data-driven. Rather than making explicit demands on signal-domain regularity (piece-wise smoothness, etc.) algebraic restrictions are placed on transform coefficients. This type of design is directly in tune with the sparse representation observation as it algorithmically seeks transforms that grant the optimal n-term linear or nonlinear approximation of signals within a class. An important advantage of methods in this category is that they can exploit general forms of regularity beyond smoothness. Such regularity may be nonintuitive and difficult to capture with concise analytical models, making second category techniques the only alternative. Placing constraints in coefficient domain also allows the transform derivations to proceed over arbitrary collections of vectors with the aid of optimization algorithms. Well-known designs in this category include the Karhunen-Loeve transform (KLT), over-complete transforms obtained using the KSVD, as well as designs accomplished with various optimization algorithms [3], [6], [21], [53].

In this paper we provide a second category technique that enforces sparsity on transform coefficients by using the ℓ_0-norm. Influenced by recent designs [2], [3], [16] but focusing on orthonormal transforms and a novel clustering technique, we derive an algebraic method that results in efficient transforms for generic classes of signals with varying

Manuscript received April 1, 2013; revised March 7, 2014 and October 10, 2014; accepted March 3, 2015. Date of publication March 23, 2015; date of current version April 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles Boncelet.

O. G. Sezer is with the Mobile Processor Innovation Laboratory, Samsung Mobile, Richardson, TX 75082 USA (e-mail: osman@gatech.edu).

O. G. Guleryuz is with the Mobile Research Laboratory, LG Electronics, San Jose, CA 95119 USA (e-mail: guleryuz@ieee.org).

Y. Altunbasak is with the Scientific and Technological Research Council of Turkey, Ankara 06100, Turkey (e-mail: yucel.altunbasak@tubitak.gov.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2414879

statistical characteristics. Among other improvements, we show that transforms designed with the proposed method are nontrivial generalizations of the KLT as they substantially outperform it on general signals but reduce to it on Gaussian random processes. Hence applications motivated by KLT's well-known optimality over Gaussian processes can safely transition to the proposed method without incurring negative trade-offs. Beyond providing completely new designs, our method can also be used to significantly improve the efficiency of existing designs so that their applicability and structural properties can be extended to new types of signals. As we will see, one of the key properties of our work is its compression-friendly formulation which allows the resulting designs to serve effectively in compression codecs in addition to facilitating reconstruction-only applications.

Our work can be considered as part of the stream of research that started in relation to wavelet-based compression techniques [36], [41], [44], [51]. Early success of these techniques over images led researchers to analyze and quantify the classes of signals over which wavelet compression is optimal or near-optimal [12], [17]. Surprisingly, one of the ramifications of this analysis has been the clear suboptimality of the wavelet transform over realistic models of images, which are typically formulated as piecewise smooth processes with discontinuities along curves. Model-based work hence focused on the design of anisotropic transforms (multiresolution transforms such as [8], [33], [34], [45] as well as block-based designs [10], [18], [37], [52]) that were geared toward obtaining substantially better performance over directional singularities. On the data-driven side, noting the suboptimality of principal component analysis (PCA) methods and the KLT [13], researchers have concentrated on "independent" component analysis (ICA) and matching pursuit designs among others [20], [32]. More recently, with the help of [9] and [16] that provide algorithms for correctly finding signal expansions over overcomplete dictionaries, K-SVD and related designs have been proposed [3], [25]. All of these techniques significantly advance our understanding of signals and enable many interesting applications. Regardless, from a compression perspective it can be said that the following points remain unresolved.

- *Model Applicability:* For model-based work perhaps the most important issue is the real-world applicability of the assumed models. When one compares real-world performance improvements of anisotropic designs to that predicted by the underlying models, one encounters a significant gap. This is because real-world images are abundant with structures that are very different from directional singularities. Since regularity in such structures is not modeled nor exploited by model-based techniques, loss of compression performance ensues. To overcome this limitation the work in this paper is constructed from the ground up to be model-agnostic and broadly applicable to signals having varying structures and types of regularity.
- *Computational Complexity and Expansion Correctness:* Complexity and expansion correctness are issues that plague many otherwise excellent techniques in both categories. For example, pursuit and convex optimization techniques required in calculating signal expansions with over-complete dictionaries [3], [9], [25] place a very high computational burden on compression encoders. Moreover, unless one enforces dictionary coherency requirements in the design stage and unless target signals satisfy stringent statistical requirements [16], applicability of these techniques in finding the correct expansion coefficients over real-world signals becomes questionable. In comparison, our work uses orthonormal designs with which finding the correct signal expansion is computationally and algorithmically straightforward.
- *Quantization and R-D Optimization:* Orthonormality also has another, often overlooked, importance. The main thrust of recent work provides nonorthogonal and over-complete transform designs [3], [25]. These designs, while improving approximation performance, lose the conveniences of Parseval's theorem. Beyond making modern operational rate-distortion optimizations difficult, the nonorthogonality of these designs forces even simple scalar quantization to require elaborate search techniques for adequate operation (Figure 1). Due to its use of orthonormal transforms, our work is not encumbered by such issues and always leads to easy quantization and rate-distortion optimization problems. Furthermore, through a novel clustering technique, our algorithm enjoys these conveniences without sacrificing approximation performance.
- *Improvements and Compatibility:* Many existing techniques require a complete overhaul of the transform compression pipeline. This is done without taking advantage of existing designs on portions of the signal where such designs work well. One of the most important points of our methodology is that it is built to strictly improve performance over existing designs while preserving their desired structural properties (block-based, lapped, multi-resolution, separable, etc. [28], [36], [49]).

As compression and signal representation have shifted from PCA and linear approximation ($\mathcal{LA}$ - approximation using the first n transform coefficients) to nonlinear approximation ($\mathcal{NA}$ - approximation using the best n transform coefficients), many researchers have pointed to the optimality of nonlinear approximation using the *right basis* [12], [17], [27].

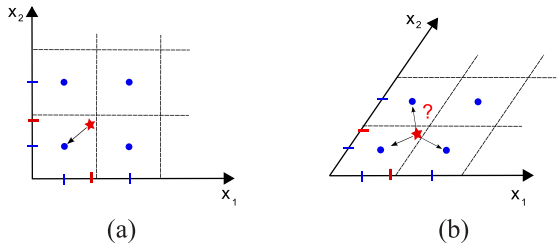


Fig. 1. Scalar quantization constellation for orthogonal and nonorthogonal systems in two dimensions. Stars represent real valued signal vectors and circles scalar quantization codewords. Two coordinate systems, a 2D orthogonal and a 2D nonorthogonal, are shown in (a) and (b) respectively. For the orthogonal setting shown in (a), scalar quantization results in the observed signal (star) to be quantized to the indicated codeword (circle), which guarantees minimum distortion. In (b) however, straightforward scalar quantization will not necessarily result in the codeword obtaining the minimum distortion, requiring a search algorithm. For further details and an effective search algorithm the reader is referred to [31].

Finding the right basis however has proven to be a very difficult task in itself as model-based work is susceptible to model failures and generally-formulated work to important nuances of compression. Our work tries to overcome these problems by providing a compression-friendly formulation while allowing for a robust and general representation. The remaining, all-important factor that can profoundly influence real-world success is the optimization of the formulation over real-world signal classes. We address this factor using a carefully-defined optimization process with annealing-like steps. Hence, in many ways our work can be considered as an important attempt at the validation of general nonlinear approximation for high performance compression.[1]

The remainder of the paper is organized as follows: In Section II we provide the motivation and definition of the sparse orthonormal transform. After considering basic properties, we compare the SOT to the KLT over Gaussian processes in Section II-C and show that they are equivalent. In Section II-D we derive the basic SOT design algorithm and proceed to comparing the SOT to the KLT over non-Gaussian processes where we observe that the SOT can discern structure even in cases where the KLT is degenerate. Section III introduces our main transform design algorithm which uses classification and annealing-like steps. The details of block, lapped, and multiresolution specializations of this algorithm are provided in Sections III-C.1 through III-C.3. We provide experimental validation of the proposed algorithm in Section IV. Final conclusions and discussions are given in Section V.

II. Sparse Orthonormal Transforms

A. Preliminaries

Suppose signals are lexicographically ordered into N-dimensional vectors. Let x $(N \times 1)$ denote a random signal coming from a zero-mean random process and let $E[.]$ denote expectation. Consider an arbitrary reconstruction basis $\mathbf{G}$ $(N \times L)$, $L \geq N$, having the i^{th} column g_i $(N \times 1)$, with $g_i^T g_i = 1$. Throughout assume that $rank(\mathbf{G}) = N$. Since we will be doing exemplary comparisons between the SOT and the KLT it is fruitful to define basic approximation constructs.

Definition 1 ($\mathcal{LA}$): The linear approximation of x with $\mathbf{G}$ is defined as

$$\hat{x}_{\mathcal{L}}(\mathbf{G}, n) = \sum_{i=1}^{n} c_i g_i, \tag{1}$$

where $1 \leq n \leq N$, and

$$(c_1, \ldots, c_n) = \underset{(\alpha_1,\ldots,\alpha_n)}{\arg\min} ||x - \sum_{i=1}^{n} \alpha_i g_i||^2, \tag{2}$$

are the expansion coefficients associated with the basis vectors g_i, $i = 1, \ldots, n$.

Remark: Note that the *index set* of the approximating coefficients is fixed to $\{1, \ldots, n\}$ independent of x so that the approximation with n coefficients is obtained using the first n columns of $\mathbf{G}$. For $\mathcal{LA}$ one typically sets $L = N$ so that with the help of the rank condition one has $\hat{x}_{\mathcal{L}}(\mathbf{G}, N) = x$. The well-known orthonormal $\mathcal{LA}$-induced design is the KLT. For example nonorthogonal designs the reader is referred to [24].

Definition 2 ($\mathcal{NA}$): The nonlinear approximation of x with $\mathbf{G}$ is defined as

$$\hat{x}_{\mathcal{N}}(\mathbf{G}, n) = \mathbf{G}c, \ \ni ||c||_0 = n \tag{3}$$

where $||.||_0$ denotes the ℓ_0-norm, i.e., the number of nonzero components of a vector, and

$$c = \underset{\alpha}{\arg\min} ||x - \mathbf{G}\alpha||^2, \ \ni ||\alpha||_0 = n \tag{4}$$

are the expansion coefficients.

Remark: Note that unlike the $\mathcal{LA}$ case, the index set of the approximating coefficients is dependent on x and approximation with n coefficients can involve any n columns of $\mathbf{G}$. Observe also that $||x - \hat{x}_{\mathcal{N}}(\mathbf{G}, n)|| \leq ||x - \hat{x}_{\mathcal{L}}(\mathbf{G}, n)||, \ \forall n$. Designs such as wavelets, curvelets, contourlets, and the KSVD are examples [3], [8], [11], [14].

Whether one is motivated by LA or NA, transition from approximation to compression can be done by invoking the familiar transform coding recipe. The $\mathcal{LA}$-motivated encoder sends n, followed by the first n coefficients, which are scalar quantized and entropy coded. The $\mathcal{NA}$-motivated encoder must send the index set of non-zero coefficients followed by the coefficient value data. Overall distortion for both cases is due to the errors introduced by the approximation and quantization processes. Under mild assumptions $\mathcal{LA}$ bit-rate can be approximated to be linear with n. This observation is not true for $\mathcal{NA}$ in general but, interestingly, for nontrivial models of multimedia signals and by using associated optimal or near-optimal basis, one can show that $\mathcal{NA}$ bit-rate can also be approximated to be linear with n [12], [17], [27], [48]. With these results and under associated conditions, compression distortion-rate performance can be approximated to asymptotically track n-term approximation performance for both cases.[2] Of course, establishing a strong duality between approximation and compression leads one to the more difficult problem of finding optimally approximating transforms.

Definition 3 (KLT): Let $\mathbf{K} = E[xx^T]$ denote the covariance matrix of the stochastic process. Assume $\mathbf{K}$ has the eigendecomposition $\mathbf{K} = \mathbf{PDP}^T$, where $\mathbf{P}$, $\mathbf{P}^T\mathbf{P} = \mathbf{1}$, is the matrix of orthonormal eigenvectors, i.e., the Karhunen-Loeve transform, and $\mathbf{D}$ is the diagonal matrix of eigenvalues. Assume that eigenvalues have been ordered in decreasing order so that the first column of the KLT corresponds to the largest eigenvalue.

It is well-known that the KLT optimizes n-term linear approximation performance, i.e.,

$$E[||x - \hat{x}_{\mathcal{L}}(\mathbf{P}, n)||^2] \leq E[||x - \hat{x}_{\mathcal{L}}(\mathbf{G}, n)||^2], \ \forall \mathbf{G}. \tag{5}$$

For $\mathcal{NA}$ on the other hand we not only have,

$$E[||x - \hat{x}_{\mathcal{N}}(\mathbf{P}, n)||^2] \leq E[||x - \hat{x}_{\mathcal{L}}(\mathbf{P}, n)||^2], \tag{6}$$

but as the above alluded to work advertises, on multimedia-like signals with $\mathcal{NA}$-optimal basis, $\mathbf{G}^*$, one can obtain

$$E[||x - \hat{x}_{\mathcal{N}}(\mathbf{G}^*, n)||^2] << E[||x - \hat{x}_{\mathcal{N}}(\mathbf{P}, n)||^2], \tag{7}$$

[1] Our early work appeared in [39] and [41].

[2] The reader should consult [12], [17] for related caveats.

while maintaining comparable rate and quantization distortion to the KLT. In essence significantly better performance than that of $\mathcal{LA}$ with the KLT can be had by switching to $\mathcal{NA}$ and using the $\mathcal{NA}$-optimal basis [8], [12], [33], [33], [47], [48]. The substantial recent interest in $\mathcal{NA}$-optimal designs can be tied to these results.

In Sections II-C and II-E we will see that results like (7) are expected to hold *only* when the underlying stochastic process deviates strongly from a Gaussian. This is because for Gaussian processes the KLT is $\mathcal{NA}$-optimal as well as being $\mathcal{LA}$-optimal (Proposition 1). Given the abundance of signals that can be closely approximated with Gaussian processes and given the prevalence of Gaussian-like signal segments even on multimedia data, it is hence not completely unexpected that the KLT remains popular in its various incarnations, especially via the DCT [4]. In comparison, the number and success of $\mathcal{NA}$ optimizing designs for general processes have so far been limited. Our aim in this article will be to provide designs that allow one to tap into the performance advantages of $\mathcal{NA}$-optimal designs *where they exist* without losing the convenience and performance of the KLT. In particular, over Gaussian processes, where KLT is optimal, we will see that our SOT designs analytically reduce to the KLT. Over more general processes our designs, through $\mathcal{NA}$, will utilize non-Gaussian statistics to significantly improve over the KLT. As will become clear, the SOT and the SOT design algorithm presented in this paper are intended to be compatible and safe upgrades to the KLT and PCA for data with unknown distribution.

B. Motivation and Definition of the SOT

Definition 4 ($\mathcal{NA}$ Cost): For a given class of signals and a scalar parameter, $\lambda \geq 0$, we define the λ-level nonlinear approximation cost of a reconstruction basis $\mathbf{G}$ $(N \times L)$ as

$$\mathcal{C}_{\mathcal{N}}(\mathbf{G}, \lambda) = E[\min_{\alpha}\{||x - \mathbf{G}\alpha||^2 + \lambda||\alpha||_0\}], \tag{8}$$

where α $(L \times 1)$ are the expansion coefficients of x and $||.||_0$ denotes the ℓ_0 norm. We will refer to the first term on the right side of (8) as the *distortion* cost and the second term as the *complexity* cost.

Remark: Note that as λ increases, the impact of the complexity cost increases and approximations to x become coarse. As λ decreases, the impact of the complexity term decreases, and approximations to x become more faithful.

The SOT definition we provide below will determine orthonormal basis that minimize (8). Observe that without any structural constraints on $\mathbf{G}$ one would prefer a reconstruction basis with the number of columns, L, satisfying $L \gg N$ so that the complexity term is reduced as much as possible at a given level of distortion. For example, for interesting values of λ, one can imagine gain-shape vector quantization codebooks [21] that have codewords on the sphere to be good candidates for columns of $\mathbf{G}$. This trivialization of the reconstruction basis happens because the cost in (8) accounts for the complexity of the expansion in an analytically compelling but artificial way. In order to utilize the mathematical conveniences of the zero-norm while retaining useful correspondences between complexity and bitrate one must impose structural constraints on the basis.[3]

One way to impose structure is to constrain L and design basis that minimize (8). The work in [3] proceeds along this line albeit using the ℓ_1 norm. As mentioned in the introduction, the resulting designs are very valuable from an approximation perspective but bring about further constraints that are not always conducive for compression applications. In particular: (i) Utilizing the ℓ_1 norm can only guarantee sparse coefficient vectors under constraints on the reconstruction basis and over restricted class of signals (see coherency and sparsity conditions in [16]). One must incorporate these constraints into the minimization and verify that the class of signals one is interested in satisfy the conditions at meaningful levels of distortion; (ii) Being nonorthogonal, the resulting designs lose the advantages of Parseval's theorem. Expansion coefficients α of a signal x have to be found via solving, $\min_{\alpha}\{||x - \mathbf{G}\alpha||^2 + \lambda||\alpha||_1\}$, which can only be done using computationally complex optimization algorithms; (iii) Adequate quantization of the expansion coefficients requires the solution of an integer program [24], [30].

In this paper we bypass these issues by using the ℓ_0 norm, setting $L = N$, and fixing $\mathbf{G}$ to be orthonormal so that $\mathbf{G}^T\mathbf{G} = \mathbf{1}$. This relieves the above points since (i) no norm change is required, (ii) expansion coefficients can be obtained by applying $\mathbf{G}^T$ followed by thresholding, and (iii) quantization is reduced to straightforward scalar quantization.

Discussion: It is important to note that these conveniences are *not* coming at a significant loss in representation accuracy in contexts that target parsimonious representation of multimedia-like signals. In compression, theoretical results on multimedia-like signals point to optimality or near optimality of well-chosen orthonormal designs [12], [15], [17], i.e., the right orthonormal transform is expected to lose little over such signals due to the orthonormality constraint. In compressed sensing, the fundamental representation assumption being made is that the signals being compressively sampled are sparse with respect to an orthonormal transform (typically taken as the wavelet transform [9]). The optimal nonlinear approximating orthonormal transform, i.e., the SOT as defined below, is in fact the best transform for the compressive sensing framework.

We thus arrive at the following definition which introduces the sparse orthonormal transform.

Definition 5 (SOT): For a given parameter, $\lambda \geq 0$, the sparse orthonormal transform, $\mathbf{G}_\lambda$, is the orthonormal transform that minimizes the λ-level nonlinear approximation cost among all orthonormal transforms, i.e.,

$$\mathcal{C}_{\mathcal{N}}(\mathbf{G}_\lambda, \lambda) \leq \mathcal{C}_{\mathcal{N}}(\mathbf{H}, \lambda), \quad \forall \mathbf{H} \ni \mathbf{H}^T\mathbf{H} = \mathbf{1}. \tag{9}$$

Remark: Note that the SOT is λ-dependent and that we immediately have $\mathcal{C}_{\mathcal{N}}(\mathbf{G}_\lambda, \lambda) \leq \mathcal{C}_{\mathcal{N}}(\mathbf{P}, \lambda)$.

[3]Note that typical analysis of the duality of approximation and compression on continuous-time signals in Sobolev and Besov spaces also uses orthonormal basis functions [12].

Given the above definition, in order to better understand the SOT let us establish its relationship to the KLT by briefly considering Gaussian processes.

C. SOT Over Gaussian Processes: SOT vs. KLT

Proposition 1 (SOT vs. KLT on Gaussian Processes): Suppose the signals of interest are obtained as realizations of a zero-mean Gaussian process with covariance matrix $\mathbf{K}$. As before, let $\mathbf{K} = \mathbf{PDP}^T$ denote the eigenvalue problem of $\mathbf{K}$ with $\mathbf{D}$ as the diagonal matrix of eigenvalues and $\mathbf{P}$ the orthonormal matrix of eigenvectors, i.e., the Karhunen-Loeve transform. Then, for any $\lambda \geq 0$,

$$\mathcal{C}_{\mathcal{N}}(\mathbf{P}, \lambda) = \mathcal{C}_{\mathcal{N}}(\mathbf{G}_\lambda, \lambda), \tag{10}$$

i.e., SOT and KLT accomplish the same optimal cost. Suppose further that the eigenvalues of $\mathbf{K}$ are distinct and the KLT is unique. Then, for $0 < \lambda < E[x^T x]$, i.e., for meaningful levels of λ, and up to column re-orderings,

$$\mathbf{P} = \mathbf{G}_\lambda, \tag{11}$$

i.e., the sparse orthonormal transform is also unique, independent of λ, and equal to the Karhunen-Loeve transform.

Proof: See Appendix I. ∎

Remark: With Proposition 1 it is clear that over Gaussian processes the SOT reduces to the KLT. Thus if one traded off the KLT for the SOT, one would lose nothing over Gaussian signals which is the domain of optimality of the KLT. In Section II-E we will see that over non-Gaussian processes the SOT provides benefits far beyond the KLT. One can hence argue that switching to the SOT and the SOT design algorithm provides a no-compromise alternative to the KLT and PCA.

The next section introduces the basic algorithm for SOT derivation which will allow more in-depth analysis and comparisons.

D. Basic Algorithm for SOT Derivation

The basic algorithm for SOT derivation is composed of two steps of alternating minimizations that optimize (8). The first step (Proposition 2) accomplishes the inner minimization for a fixed transform, i.e., $\min_c\{||x - \mathbf{G}c||^2 + \lambda||c||_0\}$, whereas the second step (Proposition 3) optimizes the transform for a given set of coefficients, i.e., $\min_{\mathbf{G}} E[||x - \mathbf{G}c||^2]$.

Proposition 2 (Optimal Expansion Coefficients): Suppose $\mathbf{G}$ $(N \times N)$ is a given orthonormal transform having the i^{th} column g_i. Consider the minimization $\min_\alpha\{||x - \mathbf{G}\alpha||^2 + \lambda||\alpha||_0\}$. Let $c = \mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})$, where for $i = 1, \ldots, N$,

$$c_i = \mathcal{T}(g_i^T x, \lambda^{1/2}) = \begin{cases} g_i^T x, & |g_i^T x| \geq \lambda^{1/2} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Then

$$||x - \mathbf{G}c||^2 + \lambda||c||_0 \leq ||x - \mathbf{G}\alpha||^2 + \lambda||\alpha||_0, \ \forall\alpha. \tag{13}$$

Proof: Since $\mathbf{G}$ is orthonormal, the minimization can be accomplished independently for each vector component, i.e., $\min_\alpha\{||x-\mathbf{G}\alpha||^2+\lambda||\alpha||_0\} = \sum_{i=1}^N \min_{\alpha_i}\{||g_i^T x-\alpha_i||^2+\lambda||\alpha_i||_0\}$. Considering the inner cost function with α_i set to zero and non-zero values establishes the result. ∎

Remark: The reader interested in solutions for other ℓ_p-based complexity cases is referred to [26].

Using Proposition 2 for the inner minimization, the $\mathcal{NA}$ cost of Equation 8 becomes,

$$\begin{aligned}
&\mathcal{C}_{\mathcal{N}}(\mathbf{G}, \lambda) \\
&= E[||x - \mathbf{G}\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||^2] + \lambda E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||_0], \\
&= E[||\mathbf{G}^T x - \mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||^2] + \lambda E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||_0], \\
&= E[||\mathbf{G}^T x||^2] - E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||^2] \\
&\quad + \lambda E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||_0], \\
&= E[||x||^2] - \Big\{E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||^2] \\
&\qquad - \lambda E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||_0]\Big\},
\end{aligned} \tag{14}$$

where we have used that $\mathbf{G}$ is orthonormal and that the product of a coefficient with its thresholded version results in the square of the latter. The SOT must hence maximize,

$$\begin{aligned}
&E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||^2] - \lambda E[||\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})||_0] \\
&\quad = \sum_{i=1}^N E[(||g_i^T x||^2 - \lambda)\big||g_i^T x| \geq \lambda^{1/2}]Pr(|g_i^T x| \geq \lambda^{1/2}),
\end{aligned} \tag{15}$$

where $Pr(|g_i^T x| \geq \lambda^{1/2})$ is the probability that the i^{th} coefficient magnitude will exceed the threshold. Rather than trying to solve the highly nonlinear (15) in a single step we will solve for the optimal transform for a given set of coefficients, redetermine the optimal coefficients for the solved transform, and so on.

Proposition 3 (Optimal Transform): Suppose $\mathbf{H}$ $(N \times N)$ is an orthonormal transform. For random vectors x and α, consider the cost $E[||x - \mathbf{H}\alpha||^2]$. Set $\mathbf{Y} = E[\alpha x^T]$ and let $\mathbf{Y} = \mathbf{USV}^T$ denote its singular value decomposition, where $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal matrices of singular vectors and $\mathbf{S}$ is the diagonal matrix of singular values. Let

$$\mathbf{G} = \mathbf{VU}^T. \tag{16}$$

Then $\mathbf{G}$ is orthonormal and

$$E[||x - \mathbf{G}\alpha||^2] \leq E[||x - \mathbf{H}\alpha||^2], \ \forall \mathbf{H} \ni \mathbf{H}^T\mathbf{H} = \mathbf{1}. \tag{17}$$

Proof: See Appendix II. ∎

Algorithm 1, as illustrated in the panel, combines Propositions 2 and 3.

Remark: Note that, since the steps applying Propositions 2 and 3 both minimize $\hat{\mathcal{C}}_{\mathcal{N}}(\mathbf{G}, \lambda)$, the ensemble nonlinear approximation cost is non-increasing and convergence is guaranteed.

Proposition 4 (Stable Points): Assume a large enough ensemble so that ensemble averages reflect statistical averages. Let $\mathbf{G}$ $(N \times N)$ denote an orthonormal transform. If $\mathbf{G}$ is a stable point of Algorithm 1 then $E[\mathbf{G}\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})x^T]$ is a symmetric positive semi-definite matrix. Conversely, suppose $E[\mathbf{G}\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})x^T]$ is a symmetric positive semi-definite matrix with distinct eigenvalues. Then, $\mathbf{G}$ is a stable point of Algorithm 1.

Algorithm 1 Basic SOT

Given $\lambda > 0$ and an initial orthonormal transform $\mathbf{H}_0$, let x^j $(N \times 1)$, $j = 1, \ldots, J$, denote a given training ensemble of vectors drawn from a zero-mean random process. Let $\hat{E}[.]$ denote expectation over the ensemble and let $\hat{\mathcal{C}}_{\mathcal{N}}(.,.)$ denote the ensemble nonlinear approximation cost. Let $\Delta(.,.) \geq 0$ denote a scalar difference function and fix $\epsilon > 0$.

1) *Initialization:* Set $\mathbf{H} = \mathbf{H}_0$.
2) *Optimal Coefficients (Proposition 2)*: For $j = 1, \ldots, J$, $c^j = \mathcal{T}(\mathbf{H}^T x^j, \lambda^{1/2})$.
3) *Optimal Transform (Proposition 3)*: Set $\hat{\mathbf{Y}} = \hat{E}[cx^T]$, obtain the svd $\hat{\mathbf{Y}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^T$, set $\mathbf{G} = \hat{\mathbf{V}}\hat{\mathbf{U}}^T$
4) *Repeat until convergence criterion is met:* If $\Delta(\hat{\mathcal{C}}_{\mathcal{N}}(\mathbf{G}, \lambda), \hat{\mathcal{C}}_{\mathcal{N}}(\mathbf{H}, \lambda)) > \epsilon$, set $\mathbf{H} = \mathbf{G}$ and go to step 2.
5) Output $\mathbf{G}$.

Proof: Suppose $\mathbf{G}$ is a convergence point of the algorithm. Using Proposition 2 we know that optimal coefficients for $\mathbf{G}$ over x are $\mathcal{T}(\mathbf{G}^T x)$. Affecting Proposition 3, we obtain $\mathbf{Y} = E[\mathcal{T}(\mathbf{G}^T x)x^T]$ having the singular value decomposition $\mathbf{Y} = \mathbf{USV}^T$ and with $\mathbf{G}$ satisfying $\mathbf{G} = \mathbf{VU}^T$. This directly leads to $\mathbf{G}E[\mathcal{T}(\mathbf{G}^T x)x^T] = \mathbf{VU}^T\mathbf{USV}^T = \mathbf{VSV}^T$ and establishes the result. For the converse, since $\mathbf{G}E[\mathcal{T}(\mathbf{G}^T x)x^T]$ is symmetric positive semi-definite with distinct eigenvalues its eigen decomposition is unique. We hence have $\mathbf{G}E[\mathcal{T}(\mathbf{G}^T x)x^T] = \mathbf{GY} = \mathbf{GUSV}^T$ which establishes $\mathbf{GU} = \mathbf{V}$ and $\mathbf{G} = \mathbf{VU}^T$, i.e., Proposition 3. ■

Remark: Note that in the case of a $\mathbf{G}$ that results in a symmetric and positive semi-definite matrix, $\mathbf{G}E[\mathcal{T}(\mathbf{G}^T x)x^T]$, with *repeated* eigenvalues the optimal transform is not unique even though $\mathbf{G}$ satisfies Proposition 3. For convenience in discussion we will expand the notion of algorithm stable points to cover such cases.

Corollary 5: An orthonormal transform $\mathbf{G}$ that results in statistically independent transform coefficients, if it exists, is a stable point of Algorithm 1.

Proof: Since $\mathbf{G}$ results in statistically independent coefficients one immediately has that the cross correlation of thresholded and unthresholded transform coefficients, $E[\mathcal{T}(\mathbf{G}^T x, \lambda^{1/2})x^T]\mathbf{G}$, is diagonal and positive semi-definite. Using $\mathbf{Y} = E[\mathcal{T}(\mathbf{G}^T x)x^T]$ one can easily see that $\mathbf{G}$ satisfies Proposition 3. ■

Remark: As we will see below, despite being a stable point, an orthonormal transform that results in independent coefficients can have a nonlinear approximation cost that is worse than one with dependent coefficients. Surprisingly, transforms resulting in independent coefficients (and ICA) do not necessarily provide optimally sparse decompositions.

E. SOT Over Non-Gaussian Processes: SOT vs. KLT

From Definition 3 we know that the KLT results in a diagonal coefficient covariance matrix, i.e., $\mathbf{P}^T E[xx^t]\mathbf{P}$ is diagonal. Of course, over Gaussian processes it is well known that KLT has the much stronger property of resulting in statistically independent coefficients [43]. Generally speaking, transforms that result in decorrelated coefficients are not necessarily stable points of Algorithm 1. On the other hand, transforms that result in independent coefficients are stable points but not necessarily points of global optimality (Corollary 5). For the specific case of a Gaussian process we know through Proposition 1 that the KLT is in effect equivalent to the SOT. In this section we consider examples that highlight that the confluence of desirable KLT properties, namely decorrelation, independence, and SOT equivalence, quickly disappears when one moves to non-Gaussian processes. While the KLT's desirable properties vanish, the SOT keeps recovering the structure in the data.

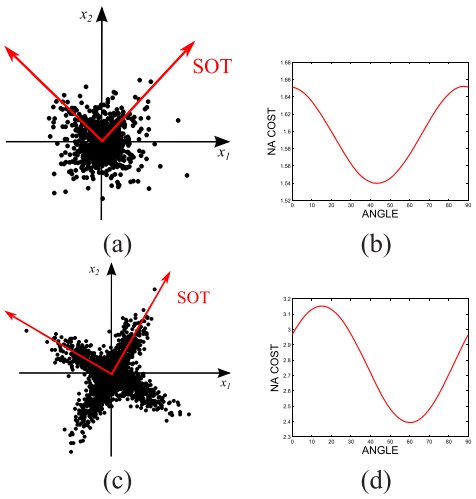


Fig. 2. Laplacian and mixture Gaussian probability distributions with $\mathcal{NA}$ cost of transforms over the associated random processes (Definition 4, $\lambda = 4$). (a) Point cloud from $f_L(R_{\pi/4}\, x)$ (coordinate system of the SOT is superimposed) and (b) $\mathcal{NA}$ cost of transforms over the Laplacian process as a function of the direction angle (each angle, counter-clockwise with respect to x_1 axis, represents an orthonormal transform). (c) Point cloud from $f_M(R_{\pi/3}\, x)$ and (d) $\mathcal{NA}$ cost of transforms over the mixture Gaussian process as a function of the direction angle.

Consider the Laplacian and mixture Gaussian probability densities given by

$$f_L(x) = \frac{1}{2} e^{-\sqrt{2}(|x_1|+|x_2|)} \tag{18}$$

$$f_M(x) = \frac{1}{4\pi}(e^{-(x_1^2/4+x_2^2/.25)/2} + e^{-(x_1^2/.25+x_2^2/4)/2}) \tag{19}$$

and let $R_\theta = \begin{bmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{bmatrix}$ denote the $2D$ rotation matrix. Note that both densities lead to $2D$ processes with covariance matrices that are scalar multiples of the identity matrix. In the following examples consider the densities $f_L(R_{\pi/4}\, x)$ and $f_M(R_{\pi/3}\, x)$. The examples illustrate the weakness of decorrelation (and hence the weakness of the KLT) as a means of uncovering the structure in the data for general cases. In all examples decorrelation is trivial to obtain as any orthonormal transform decorrelates. While any transform qualifies as the KLT, the SOT is well defined and points to the underlying structure in data.

Figure 2 shows scatter plots from two 2D distributions and associated nonlinear approximation costs of

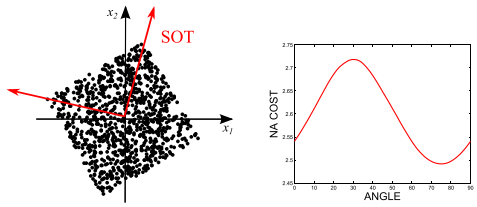


Fig. 3. Uniform distribution with $\mathcal{NA}$ cost of transforms over the associated random process (Definition 4, $\lambda = 4$). (a) Point cloud from the 2D Uniform distribution rotated by 30-degrees (coordinate system of the SOT is superimposed) and (b) $\mathcal{NA}$ cost of transforms over the uniform process as a function of the direction angle (each angle, counter-clockwise with respect to x_1 axis, represents an orthonormal transform).

2D orthonormal transforms. In the top row of the figure the distribution is $f_L(R_{\pi/4}\ x)$, i.e., a 2D separable Laplacian oriented to 45-degrees. The bottom row corresponds to the distribution $f_M(R_{\pi/3}\ x)$, i.e., an equal mixture of two Gaussians, one aligned to 60-degrees and the other to 150-degrees. For both cases the covariance matrix is a scaler multiple of the identity matrix and the KLT is hence not unique (any 2D orthonormal transform qualifies to be the KLT). Parameterizing 2D orthonormal transforms by the angle their axis makes with respect to x_1 direction it is clear that, due to symmetry, one only needs to consider $\mathcal{NA}$ costs for angles from 0 to 90-degrees. The $\mathcal{NA}$ costs shown on the right side illustrate that the minimal costs are accomplished at 45 and 60-degrees respectively. Note that unlike the covariance matrices and the KLT, which are degenerate, the $\mathcal{NA}$ costs have well-defined minimums and hence the SOT is well-defined for both cases. It is clear that the SOT is able to find the underlying structure within these distributions. Note also that for the Laplacian case an orthonormal transform that provides independent coefficients exists and the SOT coincides with it. These observations are in line with Corollary 5.

The next example in Figure 3 shows a uniform distribution where the SOT and the transform that provides independent coefficients differ. The KLT is again degenerate. At 30-degrees one obtains the transform that generates independent coefficients. Interestingly, as can be seen in Figure 3 (b), this transform obtains *the worst* $\mathcal{NA}$ cost! The SOT on the other hand points to diagonal direction at 75 degrees. Further insight into this example can be gained by considering Figure 4, where the $2D$ uniform distribution is decomposed into two distributions. These distributions are so that the uniform random vectors can be considered to be coming from a mixture of the implied stochastic processes. The top row of Figure 4 illustrates a circularly symmetric region for which that there is no preference for an optimal direction. All orthonormal transforms accomplish the same $\mathcal{NA}$ cost. In the bottom row of Figure 4 however, due to the accumulation of the distribution on the diagonals, the optimal transform can be seen to be aligned with the diagonal direction.Using Proposition 2 for a given λ, consider the point cloud depicted at the bottom. One can see two types of points: (i) Points for which the transform will retain two coefficients and accomplish zero distortion; (ii) Points for

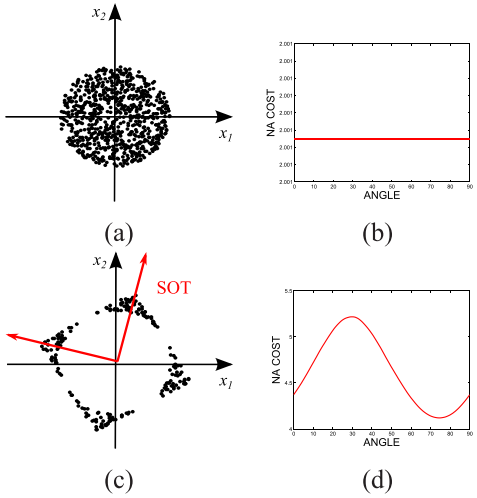


Fig. 4. The uniform distribution in Figure 3 split into two parts. (a) A distribution that is uniform within a circle resulting in a circular point cloud around the center and (b) the flat $\mathcal{NA}$ cost of transforms indicating that all transforms have equivalent cost. (c) The remaining portion resulting in a distribution concentrated in the corners and (d) the $\mathcal{NA}$ cost minimized by the SOT at 75-degrees.

which the transform will retain fewer coefficients resulting in distortion. For the same number of retained coefficients, the SOT aligns itself with the corners and ensures smaller distortion for points of type ii.

This difference between the SOT and the transform with independent components essentially means that for the same level of distortion, SOT furnishes a sparser decomposition of the data compared to a transform with independent components. One can hence conclude that transforms with independent components do not always provide optimally sparse decompositions.

III. Main Algorithm

The main algorithm for sparse orthonormal transforms augments Algorithm 1 with classification and an annealing-like step to increase the designs' robustness and to address optimization related concerns. We discuss these in turn before deriving the algorithm in Section III-C.

A. Classification

One of the key attractions of nonlinear approximation is that it allows efficient representation of a broader set of processes compared to linear approximation [7]. For example if one considers high-fidelity approximation with $n \ll N$ coefficients, $\mathcal{LA}$ with a transform would tackle random processes that result in random vectors with average energy concentrated on the *first* n basis vectors of the transform (a linear subspace). $\mathcal{NA}$ on the other hand can tackle processes with energy concentrated on *any* n basis vectors of the transform (a combinatorial $\binom{N}{n}$ union of linear subspaces, resulting in non-convex, star-shaped sets [23]). Given this representation superiority for the $\mathcal{NA}$ optimizing SOT, the important question is whether a single family of SOTs (tessellated by λ) is adequate for representing real-world multimedia-like signals.

Orthonormal designs over continuous time, piecewise smooth process models hint that a single orthonormal transform may be capable of obtaining theoretically optimal performance (see [15], [19]). In practice however one has to contend with discrete-time images and statistical learning related issues. The richness of structure in discrete-time images has so far precluded discrete-time analogs of continuous time designs from obtaining performance at the desired level. Since SOTs have to be learned from data one also has to incorporate statistical concerns about transform size. In Sections III-C.1 through III-C.3 we provide SOT designs having block, lapped, and multi-resolution structure using training sets obtained from images. It is clear that small-sized transforms lead to easier learning problems but may not sufficiently capture the local regularity. Large sizes on the other hand lead to harder learning problems and may also expose the transform to too diverse a structure. Memory and computational issues also constrain the choice of transform locality. In this paper we reduce the impact of transform size and increase the robustness to structure diversity by classifying the target random process into K classes and using a separate SOT for each. The following definition illustrates the classification of a vector for a given set of transforms.

Definition 6 (Best $\mathcal{NA}$ Transform): Let $\mathbf{G}^k$, $k = 1, \ldots, K$ denote K orthonormal transforms. Given a random vector x define the label, $\mathcal{I}(x)$, $\mathcal{I}(x) \in \{1, \ldots, K\}$, such that for $k = 1, \ldots, K$,

$$\min_{\alpha} \|x - \mathbf{G}^{\mathcal{I}(x)}\alpha\|_2^2 + \lambda\|\alpha\|_0 \leq \min_{\alpha} \|x - \mathbf{G}^{k}\alpha\|_2^2 + \lambda\|\alpha\|_0. \tag{20}$$

The best λ-level nonlinear approximating transform for x is then $\mathbf{G}^{\mathcal{I}(x)}$.

Remark: Observe that except for cases involving ties, which can be resolved randomly, we have

$$\mathcal{I}(x) = \arg\min_{k} (\min_{\alpha} \|x - \mathbf{G}^{k}\alpha\|_2^2 + \lambda\|\alpha\|_0). \tag{21}$$

Discussion: It is important to note that Definition 6 classifies each signal to use one of the K orthonormal transforms. Through (21), $\mathcal{I}(x)$ partitions the vector space into K disjoint regions. Unlike work that builds a dictionary as a union of orthonormal basis [25] and work that uses group sparsity metrics [42], [50], observe that each x is still represented by an orthonormal transform after classification with (21). With the referenced work, one has to ensure that the composite dictionary satisfies coherency constraints and the data is sufficiently sparse just to be able to solve for the sparse decomposition coefficients. One also has to further compromise due to the many disadvantages of nonorthogonal transforms and due to computational issues. Our classification on the other hand optimally uses one of the K orthonormal transforms without needing further assumptions on the dictionary and sparsity level of the data. This is consistent with the discussion in Sections I and II-B and allows our work to enjoy the convenience of using orthonormal transforms without sacrificing approximation performance. Compression and approximation related implications are discussed in Sections IV-D through IV-G.

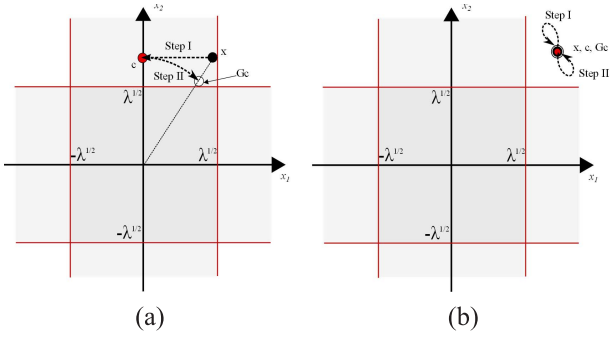
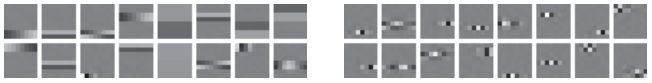


Fig. 5. Influence of a point, x, during SOT optimization (Proposition 2, step I, followed by Proposition 3, step II) assuming an initial transform of identity. (a) The point x gets thresholded to c in step I. The incentive x provides for transform optimization via $||x - \mathbf{G}c||^2$ is a rotation from c toward x as shown in step II. (b) When the point does not get thresholded it provides an incentive for the status quo.

B. Annealing

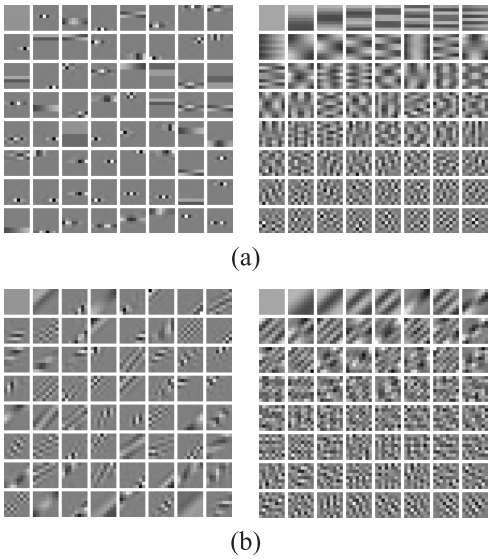
The scalar λ value in the SOT formulation (Definition 5) establishes the design of transforms for a particular complexity level. Since transform coefficients with absolute values smaller than $\sqrt{\lambda}$ are thresholded to zero, a natural way for selecting λ during the design is to relate it to the compression level via the scalar quantization step size. One can likewise select λ in a denoising context by relating it to the noise standard deviation. Given this target value one can proceed with the SOT design using Algorithm 1. As illustrated in Figure 5 however, from an optimization perspective, there is a difference between small and large values of λ and it is in fact prudent to accomplish the SOT optimization at a sequence of λ values.

Figure 5 illustrates two cases where a data sample provides an incentive to change the transform (Figure 5 (a)) and provides an incentive to keep the transform unchanged (Figure 5 (b)). Using the intuition provided by the figure one can see that applying Algorithm 1 with a very small λ will introduce marginal changes since most data samples will provide the incentive of Figure 5 (b). This of course brings about the adverse impact of generating locally optimal transforms. Our main algorithm, hence starts from large λ values which reduces the resistance of the system to changes in the transform by pushing most of the data into the thresholding zone. The derived transform is then used as the initial transform for the next step which solves the optimization again with a reduced λ. As λ decreases, large-energy samples move out of the thresholding zone, basis vectors mostly responsible for representing such large-energy clusters stabilize and change little for the rest of the iterations. In this fashion λ serves as temperature or energy state in a way similar to various annealing procedures carried out in optimization. To visualize the effect of annealing on 8×8 image blocks with horizontal structures, we have applied Algorithm 1 starting from a large λ with subsequent reductions. Figure 6 shows basis vectors that stabilized in higher λ values (Figure 6 (a)) and lower λ values (Figure 6 (b)) during the annealing process. Observe that the basis vectors in Figure 6 (a) corresponding



(a) (b)

Fig. 6. Sparse orthonormal transform designed for horizontal direction. (a) Basis vectors that stabilized at larger λ values. (b) Basis vectors that stabilized at smaller λ values.



(a) (b)

Fig. 7. Sparse orthonormal transforms (left column) and Karhunen Loeve transforms (right column). The transforms were designed over training blocks initially classified with gradient angles at (a) 0-degrees, (b) 45-degrees.

to high energy levels tend to have low-frequency in horizontal direction. (The full complement of the SOT basis vectors are shown in Section IV, Figure 7.)

Algorithm 2 Main Algorithm for Sparse Orthonormal Transforms

Given $\lambda > 0$, the number of classes, $K \geq 1$, and a set of initial transforms, $\mathbf{H}_0^k,\ k = 1, \ldots, K$, let $\mathcal{S}$ be a training ensemble of vectors drawn from a zero-mean random process. Let $\Delta(.,.) \geq 0$ denote a scalar difference function, fix $\epsilon > 0$, $\delta\lambda > 0$, and $\lambda_{max} >> \lambda$.

1) *Initialization:*
   a) Partition training set $\mathcal{S}$ into K sub-classes, $\mathcal{S}_k, k = 1, \ldots, K$.
   b) Set $\mathbf{H}^k = \mathbf{H}_0^k, k = 1, \ldots, K$.
2) *Transform Update:* Set $\lambda_t = \lambda_{max}$, $\forall k \in \{1, ..., K\}$,
   a) $\mathbf{G}^k = \mathbf{H}^k$.
   b) *Optimal transform using $\mathcal{S}_k$, $\mathbf{H}_0 = \mathbf{G}^k$, and λ_t*: Algorithm 1, steps 1-5. Output $\rightarrow \mathbf{G}^k$.
   c) *Annealing step*: $\lambda_t = \lambda_t - \delta\lambda$.
   d) *Repeat until cooled down*: If $\lambda_t > \lambda$ go to step 2b.
3) *Reclassification:*
   a) *Relabel data:* $\forall x \in \mathcal{S}$ obtain $\mathcal{I}(x) = \text{argmin}_k(\min_\alpha \|x - \mathbf{G}^k\alpha\|_2^2 + \lambda\|\alpha\|_0)$.
   b) *Update sub-classes:* $\forall k \in \{1, ..., K\}$, $\mathcal{S}_k = \{x | \mathcal{I}(x) = k\}$.
4) Overall Convergence Check:
   a) $\hat{\mathcal{C}}_\mathcal{N}(\mathbf{G}^1, \ldots \mathbf{G}^K, \lambda) = \sum_{k=1}^K \hat{E}[\min_\alpha\{||x - \mathbf{G}^k\alpha||^2 + \lambda||\alpha||_0\} | x \in \mathcal{S}_k]$.
   b) *Repeat until convergence criterion is met:* If $\Delta(\hat{\mathcal{C}}_\mathcal{N}(\mathbf{G}^1, \ldots \mathbf{G}^K, \lambda), \hat{\mathcal{C}}_\mathcal{N}(\mathbf{H}^1, \ldots \mathbf{H}^K, \lambda)) > \epsilon$, set $\mathbf{H}^k = \mathbf{G}^k, k = 1, \ldots, K$, and go to step 2a.
   c) Output $\mathbf{G}^k, k = 1, \ldots, K$.

C. Algorithm

Algorithm 2, as illustrated in the panel, combines Propositions 2 and 3 with clustering (Section III-A) and annealing (Section III-B) to design Sparse Orthonormal Transforms.

Remark: Note that, similar to Algorithm 1, the classified ensemble nonlinear approximation cost, $\hat{\mathcal{C}}_\mathcal{N}(\mathbf{G}^1, \ldots \mathbf{G}^K, \lambda)$, is non-increasing and convergence is guaranteed.

In this paper we used image gradients to accomplish the initial training set partitioning and used the identity matrix as the initial transform, i.e., $\mathbf{H}_0^k = \mathbf{1}$. While we have not done so here, a good alternative is to use the KLT of each class as the initializing transform in Step 2b (Section IV contains further simulation related details). With our annealing procedure we didn't observe any differences between the two initializers on our data set. For applications that require few annealing steps the KLT initializer may be preferred. Our software that implements the algorithm can be found in [1]. The following three subsections consider the structural specializations of the algorithm to block, lapped, and multiresolutional.

1) Block Transform Specialization (SOT): In order to generate transforms defined over image blocks Algorithm 2 is applied with a training set consisting of blocks extracted from images. The resulting SOT has block structure (Figure 7) and applications proceed by applying it over image blocks.

2) Sparse Lapped Transform Specialization (SLT): Lapped transforms have been proposed as an efficient way for reducing the compression artifacts of block-based transforms [28], [29]. While lapped transforms have spatially overlapping basis functions the structure of generated transform coefficients is very similar to block transforms. The extension of the described formulation to lapped transforms is hence straightforward. We start with an initial transform such as a lapped orthogonal transform (LOT) or a lapped bi-orthogonal transform (LBT). In the lapped formulation, the same iterative optimization process is performed as discussed. However, the training data is different. For the lapped case, the training blocks x^j $(N \times 1)$ correspond to the lapped transform coefficients, i.e., $x^j = \mathbf{A}_F y^j$, where $\mathbf{A}_F$ $(N \times N_L)$ is the forward lap transform and y^j $(N_L \times 1$ is the extended j^{th} block. (Observe that one typically has $N_L = 4N$ indicating the doubled horizontal and vertical extent of lapped transform basis functions on the image plane.) In this fashion, the SOTs computed by Algorithm 2, $\mathbf{G}^k,\ k = 1, \ldots, K$, become

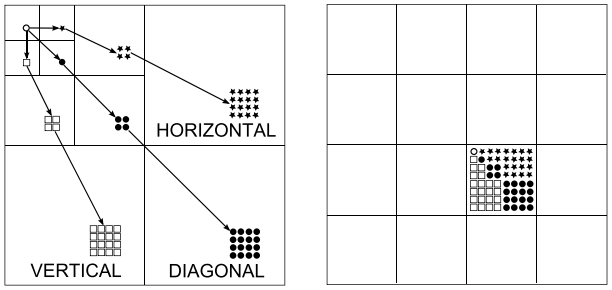


Fig. 8. Wavelet coefficient grouping for SMT. (a) Sub-bands of a three-level discrete wavelet transform. (b) The sub-bands mapped to blocks of wavelet coefficients.

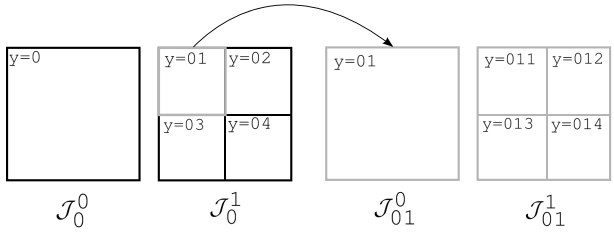


Fig. 9. Quad-tree segmentation (labels of segments are in the top-left corners). The abbreviations for sparsity-distortion cost of an encoding unit are given at the bottom.

transforms defined in the lapped transform domain. In order to obtain the overall transform in pixel domain one computes, $\mathbf{A}_I\mathbf{G}^k$, where $\mathbf{A}_I$ $(N_L \times N)$ is the inverse lap transform. For this specialization initial classification is done using the y^j.

3) Sparse Multiresolutional Transform Specialization (SMT): The extension of the proposed optimization method to wavelet transform is also formulated in the transform domain. The major difference between SLT and SMT is the way the coefficients are grouped. In SLT, the coefficients of the original lapped transform are directly mapped into a new transform space. Wavelet transforms can be handled in the same manner but we have found that better performance is observed when the coherence of the wavelet coefficients at different subbands is taken into account as follows. Consider a multiscale decomposition of an image with a 2D discrete wavelet transform. Beyond the lowest frequency subband, such a decomposition represents an image in terms of subbands having three different orientations: vertical, diagonal, and horizontal (Figure 8-(a)). Imagine the "tree" of wavelet coefficients as depicted in Figure 8 (a) (for details on tree structured coefficient groupings refer to [49]). Observe that using localized wavelet basis the tree shown in Figure 8-(a) can be considered to correspond to the spatial region/block of the image shown in Figure 8-(b). Since the wavelet coefficients of a region have strong coherence among the subbands with same orientation, we have defined vectors of wavelet coefficients for each subband orientation. For example the tree illustrated in Figure 8-(a) is decomposed into three vectors for horizontal, vertical, and diagonal orientations. Each of these vectors have 21 components with a 21×21 SOT derived for each vector (the low frequency sub-band is left as is). While these three vectors have different SOTs associated with them their classification label is joint and initially based on the gradient of the spatial location that the tree corresponds to as illustrated in Figure 8 (b).

IV. Simulations

In this paper three prototype image codecs corresponding to block, lapped, and multi-resolution transform structures are designed in order to evaluate the compression performance of the proposed approach. As an extension, to validate the representation efficiency of the new proposed method over the KLT, a KLT-based codec is also implemented.

A. Dictionary Design

The experiments start with a dictionary learning step which is performed off-line using Algorithm 2. For this purpose, we used $\sim 50,000$ blocks extracted from a training set of natural images. Block size is set to 8×8. In order to enable the SLT specialization of Section III-C.2, we extracted extended blocks of size 16×16 centered around the aforementioned blocks and applied the LBT as $\mathbf{A}_F$ [29]. For the SMT specialization of Section III-C.3 we used a 3-level CDF-9/7 wavelet transform. The training images were not used in the simulation results we report.

As the initial classification heuristic, blocks were classified based on the eight image gradient directions, varying from 0 to 157.5-degrees, with $180/8 = 22.5$-degree intervals. This process results in eight classes. The SOTs designed using Algorithm 2 that correspond to four of these are shown in Figure 7. Our optimization is not designed to preserve the initial gradient based classification, however, the final optimized basis functions do have directional structure as illustrated. In addition to these eight transforms we include the DCT as a ninth transform in the rate-distortion optimization stage for the SOT-based image codec. For SLT and SMT codecs an identity transform is included resulting in nine compression classes and transforms for each codec.

B. Dictionary Adaptation

Having a library of transforms requires appropriate adaptation to the structure of data. The method proposed here adapts transforms in a sparsity-distortion optimal fashion using quad-trees and a CART-like algorithm [33]. For ease of discussion let us consider the SOTs designed for 8×8 blocks and a given number of classes. SLT and SMT specializations follow in a straightforward way. Figure 9 shows the segmentation of an encoding unit with labels and costs for the shown segments. For a non-partitioned segment (or a leaf node, denoted by the superscript "0") the encoding cost is determined via,

$$\mathcal{J}_{\mathrm{Y}}^0 = \min_k \left(\sum_{\forall x \in \mathcal{Q}_{\mathrm{Y}}} \min_{\alpha} \|x - \mathbf{G}^k\alpha\|_2^2 + \lambda\|\alpha\|_0 \right) + E^0 \quad (22)$$

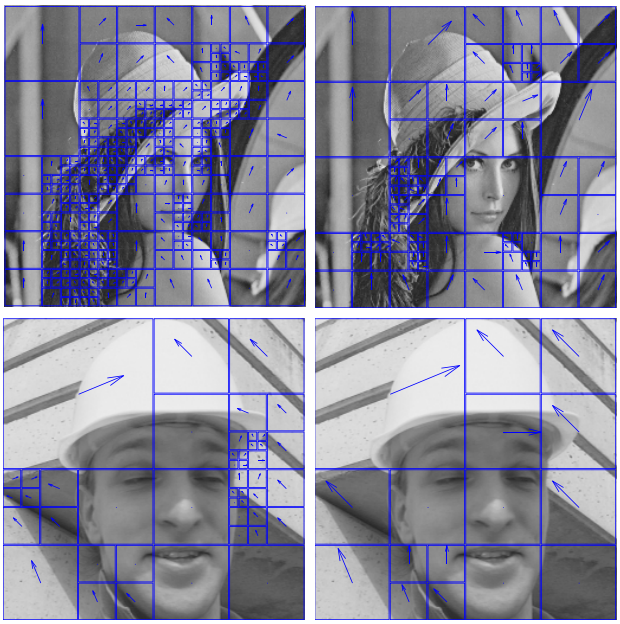


Fig. 10. Quad-tree-based classification results for $\lambda = 625$ (left column) and $\lambda = 2500$ (right column) for the images *lena* (top row), and *foreman* (bottom row). Each segment is such that all 8×8 blocks within utilize the same transform. The eight arrow directions correspond to the eight different SOTs and the dot symbol corresponds to the DCT.

where $\mathcal{Q}_{\mathtt{Y}}$ determines all 8×8 blocks in the segment $\mathtt{Y}$ and E^0 is the cost of signaling the class information of the segment, i.e., the cost of encoding the class of the transform used in the segment. In order to determine if the segment $\mathtt{Y}$ should be a leaf node or should be further divided (denoted by the superscript "1"), $\mathtt{Y}$ is segmented into four parts and the associated cost is calculated via,

$$\mathcal{J}^1_{\mathtt{Y}} = \sum_{j=1}^{4} \min(\mathcal{J}^1_{\mathtt{Y}j}, \mathcal{J}^0_{\mathtt{Y}j}) + E^1, \tag{23}$$

where E^1 is the cost of signaling the subdivision. For $\mathtt{Y}$ to be a leaf node, the above two costs should satisfy $\mathcal{J}^1_{\mathtt{Y}} > \mathcal{J}^0_{\mathtt{Y}}$. Using this condition the algorithm generates a quad-tree segmentation of the image such that all blocks within each segment are of the same class and thus utilize the same transform.

Figure 10 shows the quad-tree partitioning obtained by the described transform adaptation method using the SOT and two λ values. The arrows represent the initial gradient direction for the class of the transform that is used to encode all 8×8 blocks within that particular segment. Observe that as the λ value gets larger the adaptation of the transforms gets coarser, i.e., larger segments are generated. It is important to note that this structure and the fact that the structure remains to be geometric is discovered by the classification resulting from an algebraic optimization process.

C. Image Codecs

The proposed block-transform codec divides image into 8×8 non-overlapping blocks and finds coefficients of each block with the transform provided by the quad-tree segmentation algorithm. For lapped- and wavelet-transform codecs, transformation happens in lapped/wavelet domain where the original coefficients are replaced with the coefficients of new transforms. We utilize scalar quantization and an entropy coder that encodes coefficients based on significance similar to SPIHT. An important issue after the SOT, SLT, or SMT is evaluated is the ordering of the transform coefficients. If the coefficients are not placed in a suitable ordering before entropy coding with a SPIHT-like coder there may be a significant rate penalty. Since the significance of a coefficient is related to its energy, we incorporated an expected-energy-based ordering of coefficients as follows. First the basis vectors of the SOT (and SLT) of each class are ordered with respect to their energy level by calculating the variances of associated coefficients within the respective class. The coefficients are then organized into 64-subbands based on their energy levels similar to [40], [49]. A different approach is employed for the SMT-based codec in which the coefficients with the same sub-band orientation are ordered in decreasing energies from coarse to fine scales in the first 3-levels [38]. Since it is common to have 5-level wavelet decomposition, an additional two-level CDF 9/7 decomposition is applied to the low frequency components of the three-level representation. Finally, the coefficients of the SOT, SLT and SMT are quantized with a uniform dead-zone quantizer which is followed by entropy coding with a SPIHT-like encoder. The quad-tree segmentation and associated class information is also included in the entropy-coded bit-stream.

TABLE I

COMPRESSION PERFORMANCE IN TERMS OF BJONTEGAARD DISTORTION (BD) RATE AND PSNR VALUES FOR BLOCK-, LAPPED- AND MULTIRESOLUTION-STRUCTURE CODECS

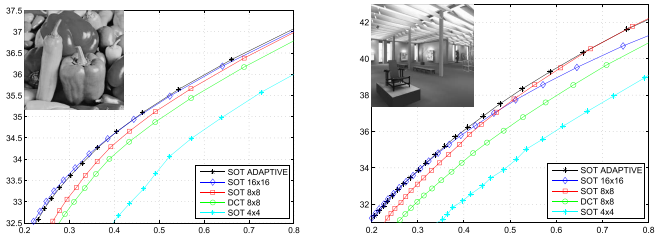
	BLOCK SOT vs. DCT		LAPPED SLT vs. LBT		WAVELET SMT vs. D97	
	BD-rate	BD-psnr	BD-rate	BD-psnr	BD-rate	BD-psnr
lena	-6.41	0.34	-1.66	0.08	-3.17	0.14
barbara	-2.76	0.20	-0.03	0.002	-7.20	0.48
museum	-12.74	1.17	-7.01	0.58	-8.18	0.66
boat	-3.51	0.18	-1.10	0.05	-3.47	0.16
cameraman	-11.22	0.86	-6.08	0.46	-1.84	0.13
foreman	-15.25	0.97	-5.78	0.31	-3.76	0.18
chair	-11.11	1.06	-6.20	0.55	-8.91	0.72
peppers	-5.77	0.25	-5.10	0.18	-3.55	0.13
goldhill	-0.55	0.02	-0.20	0.01	-4.83	0.19

D. Compression Results

The experiments were conducted with a standard set of test images augmented with a few computer-generated images. All images are 512×512 except for *foreman* and *cameraman* images, which are 256×256. Test images *museum* and *chair* are computer generated, the rest are natural images. Structure-wise, *peppers, foreman, cameraman* and the computer-generated images have strong directional structure, *barbara* has distinctly anisotropic textures. Compression results are given in Table I, where we

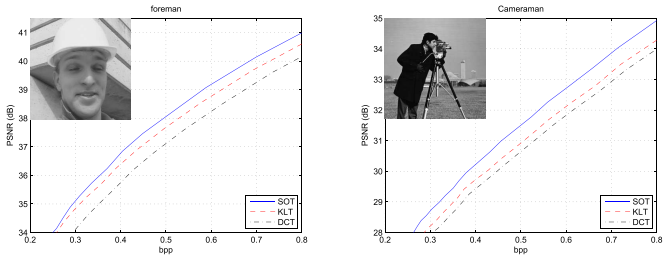
TABLE II
DENOISING PERFORMANCE (IN *dBs* PSNR) OF GLOBALLY TRAINED KSVD AND SOT DICTIONARIES AGAINST ZERO MEAN GAUSSIAN NOISE

	Lena		Barbara		Peppers256		Boat		House		Cameraman	
σ	KSVD	SOT	KSVD	SOT	KSVD	SOT	KSVD	SOT	KSVD	SOT	KSVD	SOT
2	43.21	**43.44**	42.42	**43.49**	42.63	**43.31**	41.74	**43.03**	44.27	**44.31**	42.43	**43.84**
5	38.47	**38.66**	37.20	**38.10**	37.62	**38.10**	36.64	**37.22**	38.83	**39.28**	37.46	**38.22**
10	35.41	**35.71**	33.06	**34.33**	34.28	**34.73**	33.53	**33.84**	35.65	**35.95**	33.49	**34.06**
15	33.60	**33.89**	30.60	**32.04**	32.34	**32.72**	31.62	**31.96**	34.03	**34.19**	31.32	**31.74**
20	32.25	**32.50**	28.86	**30.37**	30.92	**31.25**	30.22	**30.60**	32.81	**32.83**	29.84	**30.20**
25	31.19	**31.35**	27.57	**29.07**	29.78	**30.07**	29.15	**29.52**	**31.78**	31.67	28.73	**29.06**
30	30.31	**30.37**	26.56	**28.02**	28.86	**29.07**	28.29	**28.62**	**30.89**	30.64	27.87	**28.15**



(a) (b)

Fig. 11. PSNR (dB) versus rate (bpp) curves for *peppers 512 × 512* (a) and *museum* (b) images.



(a) (b)

Fig. 12. PSNR (dB) versus rate (bpp) curves for KLT versus SOT comparisons. Curves are for *foreman* (a) and *cameraman* (b) images.

compare the compression performance of baseline codecs obtained using the DCT, LBT, and the wavelet transform with respect to corresponding proposed method with same structure. Note that the proposed methods (SOT, SLT, and SMT) consistently outperform the conventional approaches (DCT, LBT, and CDF 9/7). Interested readers can find similar observations for video compression with SOT in [39].

E. Compression Results With Adaptive Block Sizes

Another important aspect of compression is adapting to the scale of the local structure. Wavelet transforms are quite successful in capturing large scale dependencies using multi-scale representations. To get close to the compression performance of multi-scale representations with a block transform one can change the support size of the transform adaptively depending on the local structure. For example, around fine image details the reduction of block size may help capture local variations better. Around larger regions of uniform statistics larger block sizes may improve performance. To implement this adaptivity, SOTs with three different block sizes were trained (4×4, 8×8, and 16×16.) Next, these transforms were incorporated into a block-based codec in which the quad-tree segmentation is altered to accommodate block size adaptation. Basically, for each segment the block size of the transform that gives best sparsity-distortion cost is selected. Quantization and entropy coding is kept the same as the fixed-block-size transform coder. (In order to utilize the same entropy coder the coefficients are ordered into formations of two, three, and four-deep trees for 4×4, 8×8 and 16×16 block sizes, respectively.) The rate-distortion performance of 8×8 DCT, SOT with 4×4, 8×8, and 16×16 block sizes, and SOT with block size adaptation is provided in Figure 11. Observe that the block size adaptation follows best representation available for that bitrate.

F. KLT vs. SOT

Our last compression experiment compares the coding efficiency of the KLT with that of the SOT (see Sections II-C and II-E). For this purpose we have designed an image codec that uses the KLTs of each class (see Figure 7) in a fashion similar to the above described codecs. We have used the same sparsity-distortion based adaptation method as previously described. Figure 12 shows comparisons of KLT, DCT, and SOT-based codecs. Observe that the SOT uniformly outperforms the KLT on natural images. Note also that the performance gap becomes larger on synthetic images with sharp edges.

G. Denoising

Finally we consider denoising results that compare the KSVD with the SOT. Denoising results are intended as a means for gauging approximation performance of similarly motivated techniques that differ from structural and optimization viewpoints (see Sections I and II-B). On a given image region the KSVD-based denoiser chooses a representation from a large number of (overcomplete) basis vectors by solving a convex problem that, under certain conditions, can discover the underlying sparsity correctly. Over the same image region the SOT denoiser picks one orthonormal transform out of several (using (21)) and proceeds to approximate the region using basis vectors from that transform only (in effect, using Proposition 2). Given the orthonormal transform the SOT-denoiser correctly solves the non-convex sparse approximation problem. Hence it is important to

see that by committing to orthonormal transforms the SOT denoiser, armed with the proposed classification method, does not lose approximation performance. Table II illustrates denoising results using globally trained (i.e., not per-image adaptive) KSVD and SOT transforms, where zero mean Gaussian noise is added to corresponding image with given σ values. As illustrated the SOT together with the proposed classification provides KSVD-caliber approximation performance.

V. Conclusion

We have presented a transform design method that produces sparse orthonormal transforms by jointly optimizing classification and l_0-norm induced sparsity of transform coefficients. Our method results in transforms that can have block, lapped, or multiresolution structure and can be deployed in ways that are conformant with existing designs, in fashions that strictly improve existing algorithms. The provided compression and denoising results clearly demonstrate the usefulness of the proposed work over a diverse range of image features, including anisotropic and geometrical structures. These results can be further improved by including the designed transforms within operational rate-distortion algorithms and the applicability range can be expanded to include other reconstruction applications. In particular, techniques like compressed sensing can use our orthonormal designs (for example, the SMT of Section III-C.3 instead of the commonly used wavelet transform [9]) as the model orthonormal transform.

Given the established equivalence of the SOT with the KLT over Gaussian processes, our method can be seen as a natural extension of the KLT to non-Gaussian data. Hence applications motivated by KLT's well-known optimality over Gaussian processes can safely be altered to use the proposed method without incurring negative trade-offs and without worrying about the Gaussianity of the underlying data. With the proposed work KLT's benefits over Gaussian data are seamlessly blended with the power of a nonlinear approximation optimizing transform over more general data sets. Our formulation ensures performance improvements and compatibility over existing designs allowing for a robust transition to the proposed method.

Appendix I
Proof of Proposition 1

Assume $\mathbf{G}$ $(N \times N)$ is orthonormal and g_i denotes its i^{th} column. We prove Proposition 1 in two steps. In Proposition 6 below we show that a metric of the form $\mathcal{W}(\mathbf{G}) = \sum_{i=1}^{N} f(g_i^T \mathbf{K} g_i)$, where $f(.)$ is a concave function and $\mathbf{K}$ is a covariance matrix, i.e., a real-symmetric positive semidefinite matrix, is minimized by the KLT, i.e., with $\mathbf{G} = \mathbf{P}$. In the second step we show that $\mathcal{C}_{\mathcal{N}}(\mathbf{G}, \lambda)$ is such a metric.

Proposition 6: Let x $(N \times 1)$ denote a zero-mean random vector with covariance matrix $E[xx^T] = \mathbf{K}$. Let $\mathbf{K} = \mathbf{PDP}^T$ denote the eigen-decomposition of $\mathbf{K}$, with $\mathbf{P}$ the orthonormal matrix of eigenvectors, i.e., the KLT, and $\mathbf{D}$ the diagonal matrix of eigenvalues, $d_1, \ldots, d_N$. Suppose $f(.)$ is a given concave function. For an orthonormal transform $\mathbf{G}$ $(N \times N)$ define

$$\mathcal{W}(\mathbf{G}) = \sum_{i=1}^{N} f(g_i^T \mathbf{K} g_i), \tag{24}$$

where g_i denotes the i^{th} column of $\mathbf{G}$. Then,

$$\begin{aligned}\mathcal{W}(\mathbf{G}) &= \sum_{i=1}^{N} f(g_i^T \mathbf{K} g_i) \geq \sum_{i=1}^{N} f(d_i) \\ &= \sum_{i=1}^{N} f(p_i^T \mathbf{K} p_i) = \mathcal{W}(\mathbf{P}),\end{aligned} \tag{25}$$

i.e., equation (24) is minimized by the Karhunen-Loeve transform. Furthermore, if f is strictly concave and the eigenvalues of $\mathbf{K}$ are distinct, then the KLT is the unique minimizer of (24).

Proof: Using $\mathbf{K} = \mathbf{PDP}^T$, let $r_i = \mathbf{P}^T g_i$. Observe that $r_i^T r_i = \sum_{j=1}^{N} r_{i,j}^2 = 1$. We have

$$f(g_i^T \mathbf{K} g_i) = f(r_i^T \mathbf{D} r_i) \tag{26}$$

$$= f(\sum_{j=1}^{N} r_{i,j}^2 d_j). \tag{27}$$

Since f is concave, using Jensen's inequality [43], we obtain

$$f(g_i^T \mathbf{K} g_i) \geq \sum_{j=1}^{N} r_{i,j}^2 f(d_j). \tag{28}$$

Let $\mathbf{Q}$ be the diagonal matrix with the diagonal entries $\mathbf{Q}_{i,i} = f(d_i)$. Note that the right side of (28) becomes

$$\sum_{j=1}^{N} r_{i,j}^2 f(d_j) = r_i^T \mathbf{Q} r_i \tag{29}$$

$$= g_i^T \mathbf{PQP}^T g_i. \tag{30}$$

Let $Tr[.]$ denote the trace of a matrix. The optimization function in (24) can now be written as

$$\mathcal{W}(\mathbf{G}) = \sum_{i=1}^{N} f(g_i^T \mathbf{K} g_i) \geq \sum_{i=1}^{N} g_i^T \mathbf{PQP}^T g_i \tag{31}$$

$$= Tr[\mathbf{G}^T \mathbf{PQP}^T \mathbf{G}] \tag{32}$$

$$= Tr[\mathbf{Q}] \tag{33}$$

$$= \sum_{i=1}^{N} f(d_i), \tag{34}$$

which establishes (25). Observe that with strictly concave f and distinct d_i, equality in (28) is achieved if and only if r_i is a vector with a single non-zero component, or up to column re-orderings, $r_{i,j} = \delta_{i,j}$, and $g_i = p_i$. ■

Using Definition 4 we have

$$\begin{aligned}\mathcal{C}_{\mathcal{N}}(\mathbf{G}, \lambda) &= E[\min_{\alpha}\{||x - \mathbf{G}\alpha||^2 + \lambda||\alpha||_0\}] \\ &= \sum_{i=1}^{N} E[\min_{\alpha_i}\{||g_i^T x - \alpha_i||^2 + \lambda||\alpha_i||_0\}].\end{aligned} \tag{35}$$

The inner minimization can be straightforwardly solved (see also Proposition 2) to yield

$$\alpha_i = \mathcal{T}(g_i^T x, \lambda^{1/2}) = \begin{cases} g_i^T x, & |g_i^T x| \geq \lambda^{1/2} \\ 0, & \text{otherwise.} \end{cases} \tag{36}$$

As Proposition 1 is stated in terms of a zero-mean Gaussian random process with covariance matrix $\mathbf{K}$, $g_i^T x$ becomes a zero-mean Gaussian random variable with variance $\sigma_i^2 = g_i^T \mathbf{K} g_i$. Without loss of generality assume $\lambda > 0$ (for $\lambda = 0$, $\mathcal{C}_{\mathcal{N}}(\mathbf{G}, 0) = 0$ for any $\mathbf{G}$) and $\mathbf{K}$ is positive-definite (otherwise, simply remove the empty subspaces and consider the reduced dimensional process with positive-definite covariance). We thus have,

$$\begin{aligned} &E[\min_{\alpha_i}\{||g_i^T x - \alpha_i||^2 + \lambda||\alpha_i||_0\}] \\ &\quad = \frac{2}{(2\pi\sigma_i^2)^{1/2}}\left[\int_0^{\lambda^{1/2}} z^2 e^{-\frac{z^2}{2\sigma_i^2}} dz + \lambda \int_{-\infty}^{-\lambda^{1/2}} e^{-\frac{z^2}{2\sigma_i^2}} dz\right] \\ &\quad = f(\sigma_i^2) = f(g_i^T \mathbf{K} g_i), \end{aligned} \tag{37}$$

where λ-dependency in the last line is suppressed for notational convenience. Given Proposition 6 above, in order to prove Proposition 1 it suffices to show that $f(.)$ defined in (37) is concave. Using (37) we have,

$$\begin{aligned} f(t) &= \frac{2}{(2\pi t)^{1/2}}\left[\int_0^{\lambda^{1/2}} z^2 e^{-\frac{z^2}{2t}} dz + \lambda \int_{-\infty}^{-\lambda^{1/2}} e^{-\frac{z^2}{2t}} dz\right] \\ &= \frac{2}{(2\pi)^{1/2}}\left[t\int_0^{(\frac{\lambda}{t})^{1/2}} z^2 e^{-\frac{z^2}{2}} dz + \lambda \int_{-\infty}^{-(\frac{\lambda}{t})^{1/2}} e^{-\frac{z^2}{2}} dz\right]. \end{aligned} \tag{38}$$

A function having second order derivatives is concave (strictly concave) if and only if its second derivative is non-positive (negative) [35]. Note that $f(t)$ is differentiable and

$$\begin{aligned} &f'(t) \\ &= \frac{2}{(2\pi)^{1/2}}\left[\int_0^{(\frac{\lambda}{t})^{1/2}} z^2 e^{-\frac{z^2}{2}} dz + \cancel{(\lambda e^{-\frac{\lambda}{2t}} - \lambda e^{-\frac{\lambda}{2t}})}\left(\left(\frac{\lambda}{t}\right)^{1/2}\right)'\right] \end{aligned} \tag{39}$$

$$\begin{aligned} f''(t) &= \frac{2}{(2\pi)^{1/2}}\left(\frac{\lambda}{t} e^{-\frac{\lambda}{2t}}\right)\left(\left(\frac{\lambda}{t}\right)^{1/2}\right)' \\ &= \frac{1}{(2\pi)^{1/2}}\left(\frac{\lambda}{t}\right)^{5/2} e^{-\frac{\lambda}{2t}}\left(-\frac{\lambda}{t^2}\right) \qquad (40) \\ &< 0, \qquad (41) \end{aligned}$$

where the last line follows since $\lambda > 0$ and $t > 0$. We hence have that

$$\mathcal{C}_{\mathcal{N}}(\mathbf{G}, \lambda) = \sum_{i=1}^{N} f(g_i^T \mathbf{K} g_i), \tag{42}$$

with f strictly concave. Proposition 1 now follows with the aid of Proposition 6.

APPENDIX II
PROOF OF PROPOSITION 3

The proposition calls for the minimization of $E[||x - \mathbf{H}\alpha||^2$ in terms of orthonormal $\mathbf{H}$. Let $Tr[.]$ denote the trace of a matrix. We have,

$$\begin{aligned} E[||x - \mathbf{H}\alpha||^2 &= E[x^T x] - 2E[x^T \mathbf{H}\alpha] + E[\alpha^T \mathbf{H}^T \mathbf{H}\alpha] \\ &= E[x^T x] - 2Tr[E[\alpha x^T]\mathbf{H}] + E[\alpha^T \alpha]. \end{aligned} \tag{43}$$

The minimization $\min_{\mathbf{H}} E[||x - \mathbf{H}\alpha||^2$ thus becomes equivalent to the maximization $\max_{\mathbf{H}} Tr[E[\alpha x^T]\mathbf{H}]$. Let $\mathbf{Y} = E[\alpha x^T]$ and let $\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ denote its singular value decomposition, where $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal matrices of singular vectors and $\mathbf{S}$ is the diagonal matrix of singular values.

$$\begin{aligned} Tr[E[\alpha x^T]\mathbf{H}] &= Tr[\mathbf{Y}\mathbf{H}] \\ &= Tr[\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{H}] \\ &= Tr[\mathbf{S}\mathbf{V}^T\mathbf{H}\mathbf{U}] \\ &= Tr[\mathbf{S}\mathbf{R}], \end{aligned} \tag{44}$$

where $\mathbf{R} = \mathbf{V}^T\mathbf{H}\mathbf{U}$ is orthonormal. The maximization $\max_{\mathbf{H}} Tr[E[\alpha x^T]\mathbf{H}]$ now becomes $\max_{\mathbf{R}} Tr[\mathbf{S}\mathbf{R}]$. Since $\mathbf{S}$ is diagonal we have

$$\begin{aligned} Tr[\mathbf{S}\mathbf{R}] &= \sum_{i=1}^{N} \mathbf{S}_{i,i}\mathbf{R}_{i,i} \\ &\leq \sum_{i=1}^{N} \mathbf{S}_{i,i}, \end{aligned} \tag{45}$$

where the inequality follows since $\mathbf{R}$ is orthonormal and $|\mathbf{R}_{i,j}| \leq 1$, $\forall i, j \in \{1, \ldots, N\}$. Assuming all singular values are greater than zero, the The inequality becomes an equality if the optimal transform $\mathbf{R}^* = \mathbf{1}$. (Observe that when all singular values are greater than zero this condition becomes an if and only if condition.) We hence obtain,

$$\mathbf{H}^* = \mathbf{V}\mathbf{U}^T, \tag{46}$$

which establishes the proposition.

REFERENCES

[1] *Proof-of-Concept Software*. [Online]. Available: http://eeweb.poly.edu/~onur/source.html

[2] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," *Proc. SPIE*, vol. 5914, pp. 327–339, Aug. 2005.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[4] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[5] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Process.*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[6] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vis. Res.*, vol. 37, no. 23, pp. 3327–3338, 1997.

[7] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numer.*, vol. 15, pp. 257–325, May 2006.

[8] E. Candes and D. Donoho, *Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects With Edges*. Nashville, TN, USA: Vanderbilt Univ. Press, 1999.

[9] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[10] C.-L. Chang and B. Girod, "Direction-adaptive partitioned block transform for image coding," in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 145–148.

[11] A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 45, no. 5, pp. 485–560, 1992.

[12] A. Cohen, I. Daubechies, O. G. Guleryuz, and M. T. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1895–1921, Jul. 2002.

[13] A. Cohen and J.-P. D'Ales, "Nonlinear approximation of random functions," *SIAM J. Appl. Math.*, vol. 57, no. 2, pp. 518–540, Apr. 1997.

[14] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.

[15] D. L. Donoho, "Ridge functions and orthonormal ridgelets," *J. Approx. Theory*, vol. 111, no. 2, pp. 143–179, 2001.

[16] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[17] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2435–2476, Oct. 1998.

[18] A. Drémeau, C. Herzet, C. Guillemot, and J. J. Fuchs, "Sparse optimization with directional DCT bases for image compression," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 1290–1293.

[19] M. J. Fadili and J.-L. Starck, "Curvelets and ridgelets," in *Encyclopedia of Complexity and Systems Science*, vol. 3, R. Meyers, Ed. New York, NY, USA: Springer-Verlag, 2009.

[20] A. J. Ferreira and M. A. T. Figueiredo, "Class-adapted image compression using independent component analysis," in *Proc. Int. Conf. Image Process. (ICIP)*, vol. 1. Sep. 2003, pp. I-625–I-628.

[21] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. New York, NY, USA: Springer-Verlag, 1992.

[22] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 9–21, Sep. 2001.

[23] O. G. Guleryuz, E. Lutwak, D. Yang, and G. Zhang, "Information-theoretic inequalities for contoured probability distributions," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2377–2383, Aug. 2002.

[24] O. G. Guleryuz and M. T. Orchard, "Optimized nonorthogonal transforms for image compression," *IEEE Trans. Image Process.*, vol. 6, no. 4, pp. 507–522, Apr. 1997.

[25] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decomposition," in *Proc. 15th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, pp. 293–296.

[26] D. A. Lorenz, "Convergence rates and source conditions for Tikhonov regularization with sparsity constraints," *J. Inverse Ill-Posed Problems*, vol. 16, no. 5, pp. 463–478, 2008.

[27] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. New York, NY, USA: Academic, 2009.

[28] H. S. Malvar, *Signal Processing With Lapped Transforms*. Norwood, MA, USA: Artech House, 1992.

[29] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 1043–1053, Apr. 1998.

[30] P. Moulin, "A multiscale relaxation algorithm for SNR maximization in nonorthogonal subband coding," *IEEE Trans. Image Process.*, vol. 4, no. 9, pp. 1269–1281, Sep. 1995.

[31] P. Moulin, "A multiscale relaxation algorithm for SNR maximization in nonorthogonal subband coding," *IEEE Trans. Image Process.*, vol. 4, no. 9, pp. 1269–1281, Sep. 1995.

[32] R. Neff and A. Zakhor, "Matching pursuit video coding. I. Dictionary approximation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 13–26, Jan. 2002.

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

[34] G. Peyré and S. Mallat, "Discrete bandelets with geometric orthogonal filters," in *Proc. 12th IEEE Int. Conf. Image Process.*, Sep. 2005, pp. 65–68.

[35] R. T. Rockafellar, *Convex Analysis* (Princeton Mathematical). Princeton, NJ, USA: Princeton Univ. Press, 1970.

[36] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.

[37] I. W. Selesnick and O. G. Guleryuz, "A diagonally-oriented DCT-like 2D block transform," *Proc. SPIE, Wavelets Sparsity XIV*, vol. 8138, p. 81381R, Aug. 2011.

[38] O. G. Sezer, Y. Altunbasak, and O. G. Guleryuz, "A sparsity-distortion-optimized multiscale representation of geometry," in *Proc. 17th IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 2717–2720.

[39] O. G. Sezer, R. Cohen, and A. Vetro, "Robust learning of 2D separable transforms for next-generation video coding," in *Proc. IEEE Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2011, pp. 63–72.

[40] O. G. Sezer, O. Harmanci, and O. G. Guleryuz, "Sparse orthonormal transforms for image compression," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 149–152.

[41] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.

[42] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "Collaborative hierarchical sparse modeling," in *Proc. 44th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2010, pp. 1–6.

[43] H. Stark and J W. Woods, *Probability and Random Processes With Applications to Signal Processing*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[44] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, 2nd ed. Boston, MA, USA: Kluwer, 2001.

[45] V. Velisavljevic, B. Beferull-Lozano, M. Vetterli, and P. L. Dragotti, "Directionlets: Anisotropic multidirectional representation with separable filtering," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1916–1933, Jul. 2006.

[46] M. Vetterli, "Wavelets, approximation, and compression," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 59–73, Sep. 2001.

[47] M. B. Wakin, J. K. Romberg, H. Choi, and R. G. Baraniuk, "Wavelet-domain approximation and compression of piecewise smooth images," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1071–1087, May 2006.

[48] C. Weidmann and M. Vetterli, "Rate distortion behavior of sparse sources," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 4969–4992, Aug. 2012.

[49] Z. Xiong, O. Guleryuz, and M. T. Orchard, "A DCT-based embedded image coder," *IEEE Signal Process. Lett.*, vol. 3, no. 11, pp. 289–290, Nov. 1996.

[50] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. B, Statist. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[51] M. T. Orchard, Z. Xiong, and K. Ramchandran, "Space-frequency quantization for wavelet image coding," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 677–693, May 1997.

[52] B. Zeng and J. Fu, "Directional discrete cosine transforms—A new framework for image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 3, pp. 305–313, Mar. 2008.

[53] J. Zepeda, C. Guillemot, and E. Kijak, "Image compression using the iteration-tuned and aligned dictionary," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 793–796.

Osman Gokhan Sezer received the B.S. degree in electrical engineering from Bogazici University, Istanbul, Turkey, in 2003, the M.S. degrees in electrical engineering from Sabanci University, Istanbul, in 2005, and the Georgia Institute of Technology, Atlanta, in 2008, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology.

He joined the Texas Instruments Research and Development Center as a member of the Technical Staff in 2011, where he conducted research on imaging, video, and computer vision problems. Later, he became a system lead on lidar technologies for advanced driver assistance systems. Since 2014, he has been with the Mobile Processor Innovation Laboratory, Samsung Mobile, where he is involved in research on intersection of computer vision and image processing. His inventions with the MPI Laboratory enable first wide spread adoption of superresolution techniques in consumer cameras. His research interests include computer vision, computational imaging, and machine learning for sparse signal and image processing. He was a recipient of the Best Student Paper Award in the SPIE Visual Communication and Image Processing Conference, San Jose, CA, in 2006, and the Texas Instruments Graduate Fellowship from 2007 to 2011.

Onur G. Guleryuz received the B.S. degrees in electrical engineering and physics from Bogazici University, Istanbul, Turkey, in 1991, the M.S. degree in engineering and applied science from Yale University, New Haven, CT, in 1992, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana—Champaign (UIUC), Urbana, in 1997.

He was a Research Assistant with the Image Formation and Processing Group, Department of Electrical and Computer Engineering, UIUC, from 1992 to 1995. From 1995 to 1997, he was a Research Assistant with the Department of Electrical Engineering, Princeton University, Princeton, NJ. From 1997 to 2001, he was an Assistant Professor with the Department of Electrical Engineering, NYU Polytechnic School of Engineering, Brooklyn, NY, where he continued to serve as a Research Assistant Professor until 2012. He was with the Epson Palo Alto Laboratory, Epson Research and Development, Palo Alto, CA, from 2000 to 2004, DoCoMo Communications Laboratories USA, Inc., Palo Alto, from 2004 to 2011, and Futurewei Technologies, Inc., Santa Clara, CA, from 2011 to 2013. He is currently with LG Electronics USA, Inc., San Jose, CA.

Dr. Guleryuz's research interests include statistical signal processing, computer vision, and information theory. He received the National Science Foundation Career Award, the Seiko-Epson Corporation President's Award for Research and Development, the DoCoMo Communications Laboratories Research of the Year Award and the President's Award, and the IEEE Signal Processing Society Best Paper Award. He served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was a member of the IEEE Signal Processing Society Multimedia Signal Processing Technical Committee, where he chaired the Awards Subcommittee. He serves on the IEEE Signal Processing Society Image, Video, and Multidimensional Signal Processing Technical Committee.

Yucel Altunbasak was born in Kayseri in 1971. He received the degree from the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, in 1992, and the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Rochester, New York.

He was appointed as a Consultant Assistant Professor with Stanford University, while employed with Hewlett-Packard's Palo Alto Laboratories, Silicon Valley, CA, as a Research Engineer, in 1996. After three years in Silicon Valley, he returned to academic life as an Assistant Professor with the Department of Electrical and Computer Engineering, Georgia Institute of Technology. From 2009 and to 2011, he served as a Rector of the TOBB University of Economics and Technology, Turkey. In addition to his academic work, he has continuously worked on collaboration with the industry. He licensed and successfully prototyped a MPEG and processing device for a satellite and cable TV company. While working as a Senior Advisor to the company Vestel, he initiated and was the driving force behind an image processing technology called Pixellence, which received the Special Jury Award of the Turkish Industry and Business Association. He has supervised 19 Ph.D. students, and authored over 190 papers and 50 patents/patent applications. Since 2011, he has served as the 9th and current President of the Scientific and Technological Council of Turkey. He received the Full Professorship from the Georgia Institute of Technology in 2009.

Prof. Altunbasak has received numerous awards and memberships, and has served as an Editor of several leading research journals and chaired many industrial associations.