CROSS-SCALE PREDICTIVE DICTIONARIES FOR IMAGE AND VIDEO RESTORATION

Vishwanath Saragadam, Aswin C. Sankaranarayanan, Xin Li

Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT

We propose a novel signal model, based on sparse representations, that captures cross-scale features for visual signals. We show that cross-scale predictive model enables faster solutions to sparse approximation problems. This is achieved by first solving the sparse approximation problem for the downsampled signal and using the support of the solution to constrain the support at the original resolution. The speedups obtained are especially compelling for high-dimensional signals that require large dictionaries to provide precise sparse approximations. We demonstrate speedups in the order of $10 - 20\times$ for denoising and up to $9\times$ speed-ups for compressive sensing of images and videos.

***Index Terms*—** sparse approximation, orthogonal matching pursuit, K-SVD, dictionary, multi-scale

1. INTRODUCTION

Visual signals exhibit strong correlation across scales that can often be modeled and exploited to enhance image processing algorithms. An important example of this idea is the multi-scale coding of images using the wavelet-tree model which provides both a sparse as well as a predictive model for the occurrence of non-zero wavelet coefficients across scales [19]. Specifically, the wavelet tree model arranges the wavelet coefficients of an image onto a rooted subtree. Under such an organization, the dominant non-zero coefficients form a connected rooted ub-tree[4], i.e., children of a node with small wavelet coefficients are expected to take small values as well. The wavelet tree model is central to many compression [16], sensing [6, 7], and processing algorithms [4]. In spite of the elegant results for images, there are no known predictive as well as sparsifying transforms for visual signals like videos.

Dictionary learning provides an alternate approach to wavelets in terms of enabling sparse representations [15]. The goal here is to *learn* an overcomplete dictionary such that the training dataset can be expressed as a sparse linear combination of the elements/atoms of the dictionary. An example of this approach is the K-SVD algorithm [2]. The reliance on machine learning, as opposed to analytic constructions as in the case of wavelets, provides immense flexibility towards obtaining a dictionary that is tuned to the specifics of a particular signal class or application.

Fig. 1: Left to right: Bayer image, image reconstructed using OMP, and image reconstructed using our proposed method. While OMP takes 16 minutes, our proposed method takes only 1.5 minutes, a speed-up of $10\times$.

In spite of a large body of work devoted to learning sparse representations, there is little work devoted to learning predictive models — similar to the wavelet tree model — that exploit correlations across spatial and temporal scales. We address this gap by proposing a novel multi-scale dictionary model for videos that naturally enables cross-scale prediction. Given the set of sparsifying dictionaries – one for each scale – the non-zero support patterns of a signal and its downsampled counterparts are constrained to only exhibit specific predetermined patterns. Hence, we show that this naturally enables cross-scale prediction that can be used to speed-up algorithms like OMP. We term our algorithm *zero tree OMP*. Further, we propose a simple training method, which is a modified form of K-SVD training method, to obtain dictionaries that are consistent with our model. Finally, we empirically verify that the model works through simulations on images and videos.

Prior work. Sparse representation of visual signals is widely used in compressive sensing (CS), where a signals is sensed from far-fewer measurements than its dimensionality [5]. Most relevant to our paper is the work of Hitomi et al. [11] where a sparsifying dictionary is used on video patches to recover high-speed videos from low-frame rate sensors. Hitomi et al. also demonstrated the accuracy enabled by large dictionaries; specifically, they obtain remarkable results with a dictionary with $100,000$ atoms for video patches of dimension $N = 7 \times 7 \times 36 = 1764$. As expected, sparse approximation using algorithms like orthogonal matching pursuit (OMP) with such large dictionaries is slow.

A number of techniques have been devoted to speeding up different aspects of the problem. For problems in high-

dimensionality, i.e. large N, one approach is to embed to work on random projections of the dictionary [18]. In the context of high-dimensional data, it is typical to have dictionaries with a very large number of atoms [11], i.e., $T \gg N$. Here, the search for the atom closest to the residue becomes the most time-consuming step. One approach to speeding up OMP is by using approximate nearest neighbors and shallow-tree based matching [3, 10]. Another approach is to restrict the search space by imposing a tree structure on sparse coefficients [13]. Speed up in OMP has also been obtained through parallel implementation of the search for atoms [8], and through tweaking the least squares step [9]. However, such methods only improve the constants in complexity, thus providing lesser improvements for larger dictionaries.

There also exist a few multi-scale dictionary models in literature. Jayaraman et al. [17] provide a multi-level representation of images patches which provides speed-ups by picking frequently used dictionary atoms first. Jenatton et al. [12] present a hierarchical dictionary learning mechanism, where they impose a tree structure on the sparsity, but not much has been said about speed ups obtained.

2. PROPOSED SIGNAL MODEL

Notation. We denote vectors in bold font and scalars/matrices in capital letters. A vector is said to be K-sparse if it has at most K non-zero entires. The list of indices of non-zero entries of a sparse vector is termed its support; the support of a vector $\mathbf{s}$ is denoted as $\Omega_{\mathbf{s}}$. The ℓ_0-norm of a vector is the number of non-zero entries. Finally, given a dictionary $D \in \mathbb{R}^{N \times T}$ and a support set Ω, $D_{|\Omega}$ refers to the matrix of size $N \times |\Omega|$ formed by selecting columns of D corresponding to the elements of Ω; similarly, given a vector $\mathbf{s}$, $\mathbf{s}_{|\Omega}$ refers to an $|\Omega|$-dimensional vector formed by selecting entries in $\mathbf{s}$ corresponding to Ω.

Proposed cross-scale predictive sparse model. We propose a signal model that predicts the support of a signal across scales (see Figure 2). For simplicity, we first present the model for a two-scale scenario.

Given a collection of signals, $\mathcal{X} \subset \mathbb{R}^N$, our proposed signal model consists of two sparsifying dictionaries $D_{\text{high}} \in \mathbb{R}^{N \times T_{\text{high}}}$ and $D_{\text{low}} \in \mathbb{R}^{N_{\text{low}} \times T_{\text{low}}}$ that satisfy the following three properties.

- *Sparse approximation at the finer scale.* A signal $\mathbf{x} \in \mathcal{X}$ enjoys a K_{high}-sparse representation in D_{high}, i.e, $\mathbf{x} \approx D_{\text{high}}\mathbf{s}_{\text{high}}$ with $\|\mathbf{s}_{\text{high}}\|_0 \leq K_{\text{high}}$.
- *Sparse approximation at the coarser scale.* Given $\mathbf{x} \in \mathcal{X}$ and a known downsampling operator $W : \mathbb{R}^N \mapsto \mathbb{R}^{N_{\text{low}}}$, the downsampled signal $\mathbf{x}_{\text{low}} = W\mathbf{x}$ enjoys a sparse representation in D_{low}, i.e., $\mathbf{x}_{\text{low}} \approx D_{\text{low}}\mathbf{s}_{\text{low}}$ with $\|\mathbf{s}_{\text{low}}\|_0 \leq K_{\text{low}}$. The downsampling operator W is domain specific.
- *Cross-scale prediction.* The support of $\mathbf{s}_{\text{high}}$ is constrained by the support of $\mathbf{s}_{\text{low}}$; specifically, $\Omega_{\mathbf{s}_{\text{high}}} \subset f(\Omega_{\mathbf{s}_{\text{low}}})$, where the mapping $f(\cdot)$ is known a priori.

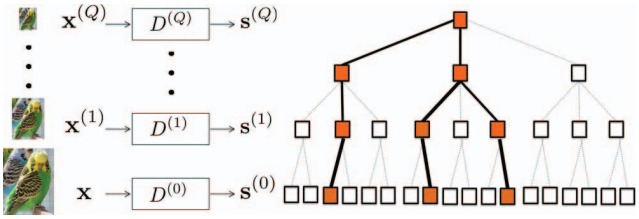

Fig. 2: Proposed cross-scale signal model with sparse coefficients across scales forming a rooted subtree. A child can be nonzero only if the parent is non-zero.

We make a few observations. First, $T_{\text{high}} \gg T_{\text{low}}$ since $N \gg N_{\text{low}}$. With the increase of dimension of the signal, more complex patterns emerge which require larger number of redundant elements. Second, since the computational time for OMP is proportional to the number of atoms in the dictionary, constraining the search space can help speed-up the algorithm. Armed with this insight, the proposed model obtains speed-ups by first solving a sparse approximation problem at the coarser scale and subsequently exploiting the cross-scale prediction property to constrain the support at finer scales.

Cross-scale mapping. We use a simple strategy for the cross-scale mapping f. Let $Q = T_{\text{high}}/T_{\text{low}}$ (assuming T_{high} and T_{low} are chosen to ensure Q is an integer). The cross-scale prediction map is defined using this simple rule

$$i \in \Omega_{\mathbf{s}_{\text{low}}} \implies (i-1)Q + \{1, 2, \ldots, Q\} \subset f(\Omega_{\mathbf{s}_{\text{low}}})$$

Each element of the support $\Omega_{\mathbf{s}_{\text{low}}}$ in the coarser scale controls the inclusion/exclusion of a *non-overlapping block* of locations for the sparse vector in the finer scale. As a consequence, the cardinality of $f(\Omega_{\mathbf{s}_{\text{low}}})$ is simply QK_{low}.

Solving inverse problems under the proposed signal model. We now detail the procedure for solving a sparse approximation problem using the proposed signal model (see Figure 2). Specifically, we seek to recover $\mathbf{x} \in \mathcal{X}$ from a set of linear measurements $\mathbf{y} \in \mathbb{R}^M$ of the form

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{e} = \Phi D_{\text{high}}\mathbf{s}_{\text{high}} + \mathbf{e},$$

where $\Phi \in \mathbb{R}^{M \times N}$ is the measurement matrix and $\mathbf{e}$ is the measurement noise. As indicated earlier, we obtain $\mathbf{s}_{\text{high}}$ using a two-step procedure.

Step1 — Sparse approximation at the coarser scale. We first solve the following sparse approximation problem:

$$(P_{\text{low}}) \qquad \widehat{\mathbf{s}}_{\text{low}} = \arg\min_{\mathbf{s}_{\text{low}}} \|\mathbf{y} - \Phi U D_{\text{low}}\mathbf{s}_{\text{low}}\|_2 \quad \text{s.t.} \quad \|\mathbf{s}_{\text{low}}\|_0 \leq K_{\text{low}}.$$

Here, $U : \mathbb{R}^N_{\text{low}} \mapsto \mathbb{R}^N$ is an upsampling operator such that WU is an identity map on $\mathbb{R}^{N_{\text{low}}}$. In all our experiments, we used a uniform down sampler and a nearest neighbour up sampler specific to the domain of the signal.

This first step recovers a low-resolution approximation to the signal, $\mathbf{x}_{\text{low}} = D_{\text{low}}\widehat{\mathbf{s}}_{\text{low}}$.

Step 2 — Sparse approximation at the finer scale. Armed with the support $\widehat{\Omega} = \Omega_{\widehat{\mathbf{s}}_{\text{low}}}$, we can solve for $\mathbf{s}_{\text{high}}$ by solving:

$$(P_{\text{high}}) \qquad (\widehat{\mathbf{s}}_{\text{high}})_{|f(\Omega)} = \arg\min_{\alpha} \|\mathbf{y} - \Phi(D_{\text{high}})_{|f(\Omega)}\alpha\|_2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq K_{\text{high}}.$$

The sparse approximation problems in both steps are solved using OMP. The proposed mapping across scales for the sparse support forms a zero tree, where a coefficient is zero if the corresponding coefficient at coarser scale is zero. Hence we refer to our algorithm as zero tree OMP.

Theoretical speed-up. Let $C(N, T, K)$ be the amount of time required to solve a sparse-approximation problem using OMP for a dictionary of size $N \times T$ and sparsity level K, given by [14]

$$C(N, T, K) = O(NTK + TK + K^4 + K^3N).$$

Hence, obtaining $\mathbf{s}_{\text{high}}$ directly from $\mathbf{x}$ would require $C(N, T_{\text{high}}, K_{\text{high}})$ computations. In contrast, our proposed two-step solution using cross-scale prediction has a computational cost of $C(N, T_{\text{low}}, K_{\text{low}}) + C(N, QK_{\text{low}}, K_{\text{high}})$.

For dictionaries with a large number of atoms, i.e., large T, and small values for sparsity level K, the linear dependence on N dominates the total computation time. Here, the speed-up provided by our algorithm is approximately $T_{\text{high}}/(T_{\text{low}} + K_{\text{low}}Q)$.

Learning cross-scale sparse models. We can learn the dictionaries $(D_{\text{high}}, D_{\text{low}})$ with a simple modification to the K-SVD algorithm. We first learn the coarse-scale dictionary D_{low} by applying K-SVD to downsampled training data $X_{\text{low}} = [W\mathbf{x}_1, \ldots W\mathbf{x}_n]$, by-product of which are $\{\Omega_{s_{\text{low}}, K_{\text{low}}}\}$. We then learn the fine-scale dictionary D_{high} by replacing OMP with zero tree OMP in the K-SVD algorithm.

Figure 3 shows an example of the learned low resolution atoms and the corresponding high resolution atoms. Observe that constraining the sparse support of the high resolution approximation alone learns patches which are very similar in appearance to the low resolution patches, which is in strong favor of our signal model.

3. EXPERIMENTAL RESULTS

We compare zero tree OMP using our proposed two-scale dictionaries against traditional OMP on dictionaries learnt using K-SVD. We used uniform downsampling operator specific to images and videos, which avoided any aliasing artifacts. We compare both the run time and approximation accuracy for images and videos. We quantify approximation accuracy using recovered SNR that is defined as follows: given a signal $\mathbf{x}$ and its estimate $\widehat{\mathbf{x}}$, SNR $= 20\log_{10}(\|\mathbf{x}\|/\|\mathbf{x} - \widehat{\mathbf{x}}\|)$.

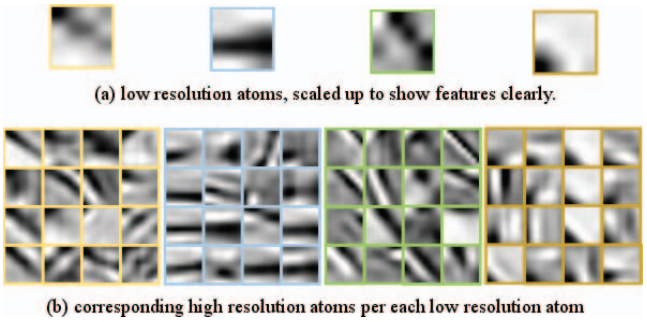

Fig. 3: Visualization of select low resolution atoms and their corresponding atoms in the high resolution dictionary.

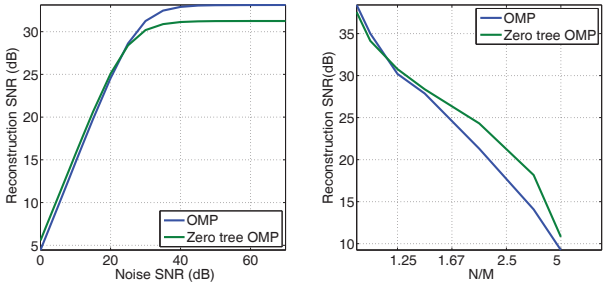

Fig. 4: Approximation accuracy for image applications. The left plot is for denoising and the right plot is for image inpainting(N/M is the number of unknown pixels per each known).

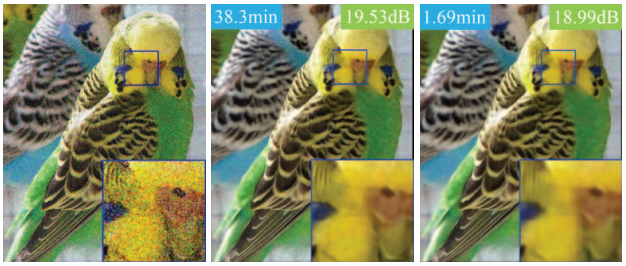

Fig. 5: Visualization of results for image denoising. From left to right: noisy image with SNR of 10dB, recovered image using proposed method, and recovered image using K-SVD learned dictionary. We obtain a speed-up of $22\times$ with hardly any reduction in accuracy.

Images. Figure 1 shows demosaicing of the Bayer pattern using both, OMP and zero tree OMP. We trained an 8192 atom high resolution dictionary on 24×24 Kodak True color RGB images [1] and 512 atom low resolution dictionary on the patches downscaled to 12×12. We compare this against 8192 atom single scale dictionary. It took 16 minutes for the single scale, whereas only 1.5 minutes for the two scale dictionary. Figure 4 shows performance metrics in terms of recovered SNR for denoising and inpainting, as compared against traditional OMP. Figure 5 shows image denoising at an SNR of 10dB. We perform denoising with the trained RGB dictionaries of 24×24 patch and with a patch overlap of 18 pixels.

Signal Class	N	N_{low}	T_{low}	T_{high}	K_{low}	K_{high}	Speedup	Model Accuracy (dB)	K-SVD Accuracy (dB)
Images	8x8	4x4	64	1024	8	8	4.10	20.67	21.98
	24x24x3	12x12x3	512	8192	8	8	22.6	19.64	20.57
Videos	8x8x16	4x4x8	512	8192	16	16	15.87	22.62	24.09
	8x8x16	4x4x8	512	8192	14	16	15.80	22.75	24.09
	8x8x32	4x4x16	512	8192	16	16	23.81	20.72	21.36
	8x8x16	4x4x8	512	16384	16	16	16.89	21.84	23.27

Legend	
N	Size of the high resolution signal
N_{low}	Size of the low resolution signal
T_{low}	Number of atoms in low resolution dictionary
T_{high}	Number of atoms in high resolution dictionary
K_{low}	Sparsity used in low resolution dictionary
K_{high}	Sparsity used in high resolution dictionary
Speed up	Ratio of time taken for single scale approximation by time taken for two scale approximation

Table 1: Table with speed-up for various dictionary sizes, patch sizes and sparsity. The speed-up shown are for solving sparse approximation problems and quantify the ratio of time taken by OMP using a K-SVD learnt dictionary to zero tree OMP on the proposed model. Also shown are approximation errors on training dataset for both K-SVD and the proposed algorithm.

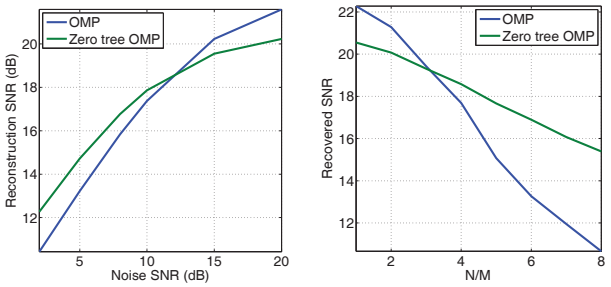

Fig. 6: Approximation accuracy for video applications. The left plot is for denoising and the right plot is for video compressive sensing(N/M is the number of frames recovered from each coded image).

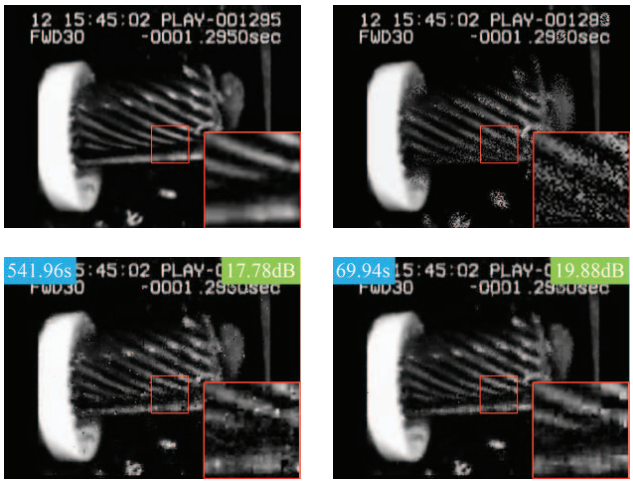

Fig. 7: Visualization of video compressive sensing. We simulated the architecture proposed in Hitomi et al. [11] where a single coded image is obtained by temporal sampling of 8 frames. Given this single coded image, we recover 8 frames (at $8\times$ frame-rate) by solving an inverse problem. Clockwise from top left: ground truth video frame; coded image from 8 frames; recovered video frame using proposed method; recovered frame using K-SVD learned dictionary. We obtain a speed-up of $9\times$ with a small increase in accuracy.

Our method is $22\times$ faster, with little difference in accuracy.

Videos. We trained an 8192 atom high resolution dictionary for $8 \times 8 \times 16$ video patches and 512 atom low resolution dictionary for the patches downscaled to $4 \times 4 \times 8$. We compared the trained dictionaries against an 8192 atom single scale dictionary obtained using K-SVD. We maintained the same sparsity across all the dictionaries. Figure 6 shows the performance of our proposed method and conventional K-SVD+OMP for denoising and video compressive sensing where we implemented the temporal sampling method proposed in Hitomi et al. [11]. Visualization of the recovered frames is shown in Figure 7. The increase in accuracy can be attributed to the video having low frequency components which is better captured by the low resolution dictionary.

Summary. Table 1 and Figures 4 and 6 quantify the performance of the proposed signal model and those obtained using K-SVD for a wide range of parameters as well as signals. Across the board, we observe that the proposed framework provides approximations that are as good as those obtained with K-SVD, but with speed-ups of $4 - 20\times$. The speed-ups obtained are comparable to results in [3] with higher approximation accuracies for our proposed method.

As a result of speed-up of the sparse coding step, we get significant speed-ups during the training phase $(2 - 10\times)$ using modified K-SVD, which makes it feasible to deal with very large scale problems.

4. CONCLUSION AND DISCUSSIONS

We presented a signal model that enables the cross-scale predictability for visual signals. Our method is appealing because of the simple extension to OMP and K-SVD algorithms while providing significant speed-ups at little or no loss in accuracy. The computational gains provided by our algorithm are especially significant for problems involving high-dimensional dictionaries with a large number of atoms.

5. ACKNOWLEDGEMENT

This work has been supported in part by Intel Corporation.

6. REFERENCES

[1] Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`. Accessed: 2015-10-05.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, 2006.

[3] A. Ayremlou, T. Goldstein, A. Veeraraghavan, and R. G. Baraniuk. Fast sublinear sparse representation using shallow tree matching pursuit. *arXiv preprint arXiv:1412.0680*, 2014.

[4] R. G. Baraniuk. Optimal tree approximation with wavelets. In *SPIE Intl. Symp. Optical Science, Engineering, and Instrumentation*, 1999.

[5] R. G. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4), 2007.

[6] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Information Theory*, 56(4):1982–2001, 2010.

[7] S. Deutsch, A. Averbush, and S. Dekel. Adaptive compressed image sensing based on wavelet modeling and direct sampling. In *SAMPTA*, 2009.

[8] Y. Fang, L. Chen, J. Wu, and B. Huang. GPU implementation of orthogonal matching pursuit for compressive sensing. In *IEEE Intl. Conf on Parallel and Distributed Systems (ICPADS)*, 2011.

[9] M. Gharavi-Alkhansari and T. S. Huang. A fast orthogonal matching pursuit algorithm. In *IEEE Intl. Conf. Acoustics, Speech, Signal Processing*, 1998.

[10] R. Gribonval. Fast matching pursuit with a multiscale dictionary of gaussian chirps. *IEEE Trans. Signal Processing*, 49(5), 2001.

[11] Y. Hitomi, J. Gu, M. Gupta, T. Mitsunaga, and S. K. Nayar. Video from a single coded exposure photograph using a learned over-complete dictionary. In *IEEE Intl. Conf. Computer Vision*, 2011.

[12] R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Intl. Conf., Machine Learning*, 2010.

[13] C. La and M. N. Do. Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In *IEEE Intl. Conf. Image Processing*, 2006.

[14] B. Mailhé, R. Gribonval, F. Bimbot, and P. Vandergheynst. A low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries. In *IEEE Intl. Conf. Acoustics, Speech, Signal Processing*, 2009.

[15] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

[16] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing*, 41(12):3445–3462, 1993.

[17] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias. Learning stable multilevel dictionaries for sparse representations. *IEEE Trans. Neural Networks and Learning Systems*, PP(99), 2014.

[18] S. N. Vitaladevuni, P. Natarajan, and R. Prasad. Efficient orthogonal matching pursuit using sparse random projections for scene and video classification. In *IEEE Intl. Conf. Computer Vision*, 2011.

[19] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Appl. Comp. Harmonic Analysis*, 11(1):89–123, 2001.