Point Coding: Sparse Image Representation with Adaptive Shiftable-Kernel Dictionaries

Doru C. Balcan
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
Email: dbalcan@cs.cmu.edu

Michael S. Lewicki
Electrical Engineering and Computer Science Dept.
Case Western Reserve University
Cleveland, OH
Email: michael.lewicki@case.edu

***Abstract*—This paper addresses the problem of adaptively deriving optimally sparse image representations, using an dictionary composed of shiftable kernels. Algorithmic advantages of our solution make possible the computation of an approximately shift-invariant adaptive image representation. Learned kernels can have different sizes and adapt to different scales. Coefficient extraction uses a fast implementation of Matching Pursuit with essentially logarithmic cost per iteration. Dictionary update is performed by solving a structured least-squares problem either by algebraic characterization of pseudoinverses of structured matrices, or by superfast interpolation methods. Kernels learned from natural images display expected 2D Gabor aspect (localization in orientation and frequency), as well as other structures commonly occurring in images (*e.g.*, curved edges, or cross patterns), while when applied to newspaper text images, kernels tend to reproduce printed symbols or groups thereof.**

I. Introduction

Efficient image representation is an important problem, due to its practical applications [9] and to its potential as a principled approach to modeling natural vision [23]. Capturing the high structure of the visual signals and encoding it compactly is a challenging task. For example, relevant visual content can appear at any spatial position and scale, which explains the success of image coders based on multiscale representations such as wavelets [28], [2]. However, even wavelets prove suboptimal in modeling certain structure frequently occurring in images (among other things, sharp edges at arbitrary orientations) and several new representations have recently been designed to fill this gap (see for instance [3], [4], [7]).

In spite of such progress, it is still not clear what makes an optimally efficient code for images in general: smooth surfaces and short straight edges may be the optimal features for some image classes (*e.g.*, natural scenes), but not for others (faces, cartoons, fingerprints, various textures, or medical images). Furthermore, according to Shannon's source coding theorem, a representation is inherently suboptimal for a given class of signals of interest, unless it captures the probability density of the data. This suggests that better representations can be obtained by learning more general and flexible dictionaries, which reflect the statistical structure of special image classes. In this paper, we focus on adaptively deriving dictionaries generated by a set of relatively small image patterns (hereby called "kernels"), shifted at arbitrary positions. The goal is to find such a set of kernels for which any signal in the target class has a sparse linear representation. Each coefficient in the sparse set represents a triple: one component is its value, the second is the index k of a kernel, while the remaining one is a point p in signal space - the location of the shifted kernel k. Therefore, from now on we choose to refer to the problem above as *Point Coding*. The usual approach to computing a solution is by minimizing a two-term cost function: the first term measures the fidelity of reconstruction (usually, an *error* term), while the second term stands for some form of regularization (in our case, *sparseness*). By employing different choices for these two terms, various algorithms can emerge, each with its own technical challenges and advantages.

This area of research has been particularly active in recent years, and several interesting directions have been explored. For example, we mention here the significant work in [5], [22], [13], [1]. Let us mention that the Point Coding problem is by no means particular to images; in fact approaches to sparse adaptive representation of general types of signals exist which focus on sounds [27], [17], [10], and even to combined audio-visual signals [19]. The goal of the present paper is to bring forward computationally efficient methods for designing very general, (approximately) shift-invariant adaptive representations of images. The generality refers to the fact that the kernel sizes can be *truly arbitrary*, the user giving up the control over this issue to the kernel-learning procedure and this process is even able to produce a multiscale representation if optimality demands it (in a similar fashion to Spike Coding [26]). In such conditions, computational efficiency is afforded by exploiting the (still high) structure of the optimization problem, using tools imported from structured matrix algebra [12], [15]. Namely, the coefficient extraction step uses a fast implementation of Matching Pursuit with essentially logarithmic cost per iteration, while dictionary update is performed by solving a highly structured least-squares problem, either by algebraic characterization of pseudoinverses of certain structured matrices [11], or by fast interpolation methods [30].

The paper is organized as follows. Section II contains the mathematical formulation of the problem. Next we describe the stages of the solution: the sparse coefficient extraction and the dictionary update. We include our experimental results in

section V and present our conclusions in the final section.

II. The Point Coding Problem

In this section we shall formulate the Point Coding problem mathematically. Let us start by introducing the appropriate notation.

Let $\Phi = \{\phi_1, ..., \phi_K\}$ be a set of 2D (rectangular) kernels, of possibly different sizes $m_k \times n_k$, normalized to unit Frobenius norm, and let $\mathbf{f} = \left(\mathbf{vec}(\phi_1)^T, \ldots, \mathbf{vec}(\phi_K)^T\right)^T$ the ensemble obtained by concatenating their vectorized versions[1]. For each kernel $\phi_k \in \Phi$ and for $p \in I\!N^2$, we denote $\phi_{k,p}$ the translated version of the kernel such that its upper-left corner lies at position p. (In this paper, we shall work exclusively with finite-size images, which means that if ϕ_k is entirely contained within an $M \times N$ image, it can only be shifted into $(M - m_k + 1) \times (N - n_k + 1)$ positions.)

For all k and p, the coefficient of the shifted kernel $\phi_{k,p}$ will be denoted $s_{k,p}$. Under the linear additive noisy model assumption, for any image x we can write:

$$x = \sum_{k=1}^{K} \sum_{p \in P_k} s_{k,p} \cdot \phi_{k,p} + \epsilon = \widehat{x}(s, \Phi) + \epsilon \qquad (1)$$

where P_k is the set of all occurrences of kernel ϕ_k in the representation[2]. As a measure of representation accuracy, we hereby consider the (squared) reconstruction error:

$$F_x(\Phi, s) = ||x - \widehat{x}(s, \Phi)||_2^2 = ||x - \sum_{k=1}^{K} \sum_{p=1}^{d_k} s_{k,p} \cdot \phi_{k,p}||_2^2 \quad (2)$$

where for all k, $d_k = |P_k|$ is the number of shifted versions of kernel ϕ_k, while imposing the sparsity restriction translates in minimizing the number of non-zeros in s. Therefore, for a fixed signal x, we should solve the following optimization problem

$$\begin{aligned} \min_{\Phi, s} \quad & ||s||_0 \\ \text{s.t.} \quad & F_x(\Phi, s) < \epsilon \end{aligned}$$

for $\epsilon \geq 0$ or equivalently:

$$\min_{\Phi, s} \quad F_x(\Phi, s) + \lambda ||s||_0 \qquad (3)$$

for some $\lambda > 0$.

The optimization problems above are NP-hard (see for instance [6]). Therefore, we attempt to approximately find a solution via an iterative, alternating procedure. First, we find a sparse set of points corresponding to a fixed dictionary and a preset level of precision; then, for a fixed set of coefficients, update the dictionary to better fit the data. Finding the sparsest linear approximation in a general dictionary is also NP-hard [20]; however, suboptimal approaches (like greedy) proved quite satisfactory in practice. Therefore, for the first step we choose to employ Matching Pursuit [18]. The second step, adapting the dictionary to the signal structure seems simpler, since it only requires solving a quadratic (*i.e.*, convex) optimization problem in Φ (or $\mathbf{f}$). In the following, we shall separately describe each of the two steps.

[1] For a matrix M, $\mathbf{vec}(M)$ is the set of all the entries in the matrix, stored column-wise.

[2] For brevity, we will further refer both to Φ and to $\mathbf{f}$ as the *encoding dictionary*. Also, we will refer to the coefficients $s_{k,p}$ as *points*. Note again that one point is determined not only by the value of the coefficient, but also by its corresponding kernel and by the position where it occurs.

III. Matching Pursuit

The Matching Pursuit (MP) algorithm [18] is a greedy iterative procedure whose goal is to identify a decomposition of a given vector as a linear combination of elements of a dictionary. If the dictionary is an orthogonal vector set and the signal is indeed a sparse combination of atoms, MP is guaranteed to find this sparse set. In general, this method only serves as an approximation to the sparsest set problem (see for instance [20]).

The main practical challenge in using Matching Pursuit with a high-dimensional, highly overcomplete dictionary is the large cost of the update and of identifying the next atom, maximally correlated with the residual. In the case of a shiftable-kernel dictionary with small kernels, approaches presented in the Matching Pursuit Toolkit (MPTK) [16] and in Sallee [25] shrink this cost to essentially logarithmic in the size of the signal (we assume that the number of kernels and their sizes are constant). The difference between the two approaches is that in MPTK this logarithmic cost reflects searching through a binary tree to find for the next maximum, while in Sallee's work this is due to maintaining a heap which holds the maximum correlation coefficients of the kernels with the signal, grouped into equal-length, adjacent blocks. Thus, every pair of *delete-max* and *insert* operations, corresponding to into the heap maintain the desired cost.

We employed this second approach, slightly adapting it to the 2D case. Namely, we compute the correlation coefficients of the image with all the kernels in our dictionary and we divide this correlation map into (roughly square) blocks of equal size. At each step, we only update a small number of blocks (namely, 4) and therefore only need to search for a small number of *new maxima*. The heap structure admittedly helps avoid most of the work; careful storage of the correlation matrix can further help by enhancing data locality and thus avoid costly memory operations. As a typical result, decomposing a 256×256 image to 30dB using a dictionary of 25 Gabor-looking 8×8 kernels can be executed in as little as 6.4 seconds on a G5 Mac computer (with a MEX C implementation of Matching Pursuit). We intend to provide a more thorough description and analysis of the above procedure in a future paper.

Once MP computes a sparse set of coefficients (now considered fixed), we proceed to optimizing the kernels to better fit the signals.

IV. Dictionary Update

In the following we describe the optimal dictionary for a given set of points as the mode of the posterior distribution, in a similar fashion to [26], [24]:

$$p(x|\Phi) = \int p(x|\Phi, s)p(s)ds \tag{4}$$

where for integration we marginalize over all possible point sets. We can approximate the integral above with $p(x|\Phi, s')p(s')$, where s' is the set of coefficients produced by Matching Pursuit. Then, assuming an additive Gaussian noise model $\epsilon \sim N(0, \sigma_\epsilon I)$, for every kernel ϕ_k:

$$\frac{\partial}{\partial \phi_k} \log(p(x|\Phi)) = \frac{\partial}{\partial \phi_k}\{\log(p(x|\Phi, s')) + \log(p(s'))\} \tag{5}$$

$$= \frac{-1}{2\sigma_\epsilon}\frac{\partial}{\partial \phi_k}||x - \sum_{k'=1}^{K}\sum_{p=1}^{n_{k'}} s_{k',p} \cdot \phi_{k',p}||_2^2 \tag{6}$$

$$= \frac{1}{\sigma_\epsilon}\sum_{p=1}^{n_k} s_{k,p} \cdot [x - \widehat{x}(s, \Phi)]_{k,p} \tag{7}$$

where $[x - \widehat{x}(s, \Phi)]_{k,p}$ denotes the restriction of the error image $x - \widehat{x}(s, \Phi)$ on the support of $\phi_{k,p}$. This immediately gives us a learning rule for the MAP dictionary and we could employ any (stochastic) gradient based method to perform the optimization.

Let us observe that maximizing the posterior with respect to Φ means minimizing the (squared) reconstruction error, which is simply a quadratic form of the ensemble $\mathbf{f}$. Indeed, if we denote by y_i the vectorized version of training image x_i, let us denote by $S^{(i)} = [S^{(i,1)}, \ldots, S^{(i,K)}]$ the matrix corresponding to the linear mapping

$$\widehat{x_i}(s^{(i)}, \Phi) = S^{(i)} \cdot \mathbf{f} \tag{8}$$

and so, the optimization problem we need to solve reduces to minimizing the following cost function:

$$Q(\mathbf{f}) = \sum_{i=1}^{I} ||x_i - S^{(i)}\mathbf{f}||_2^2 \tag{9}$$

$$= \sum_{i=1}^{I}\left(||x_i||_2^2 - 2x_i^T S^{(i)}\mathbf{f} + \mathbf{f}^T S^{(i)T} S^{(i)}\mathbf{f}\right) \tag{10}$$

$$= ct. + \sum_{i=1}^{I}\left(-2x_i^T S^{(i)}\mathbf{f} + \mathbf{f}^T S^{(i)T} S^{(i)}\mathbf{f}\right) \tag{11}$$

$$= ct. - 2\left(\sum_{i=1}^{I} x_i^T S^{(i)}\right)\mathbf{f} + \mathbf{f}^T\left(\sum_{i=1}^{I} S^{(i)T} S^{(i)}\right)\mathbf{f} \tag{12}$$

$$=: c - 2b^T\mathbf{f} + \mathbf{f}^T A\mathbf{f} \tag{13}$$

$$\tag{14}$$

We remark that this quadratic form has a special structure: since matrix $S^{(i)}$ is a block-row, whose blocks are each Toeplitz-Block-Toeplitz matrices[3] it follows that matrices $S^{(i)T}S^{(i)}$ will be Toeplitz mosaic matrices of identical block sizes. Consequently, matrix A will be a symmetric, positive semidefinite Toeplitz mosaic matrix. We can thus reduce the original problem to a structured least squares problem.

Structured Least Squares. The advantage of working with structured matrices comes mainly from the fact that the number of parameters is much smaller than the actual dimension of the matrix and from the existence of fast and superfast algorithms that exploit the displacement rank of many such types of matrices [12], [15].

First, we would like to point out that [11] presents an algebraic characterization of pseudoinverses of Toeplitz and Hankel mosaic matrices, which generalizes the well-known Gohberg-Semencul inversion formula for Toeplitz matrices, by using a general notion of Bezoutian to represent such matrices. The effect is that fast and superfast algorithms can be employed to compute the pseudoinverses, and consequently to solve the structured least-square problems. We shall only give here one example of such a result; a more detailed description and analysis of this idea will be the subject of a future paper.

Definition 1: A (q, p)-mosaic matrix B is said to be a generalized Toeplitz (q, p, r)-Bezoutian if its generating function admits the representation

$$\widehat{B}(\lambda, \mu) = \frac{1}{1 - \lambda\mu}\widehat{U}(\lambda)\widehat{V}(\mu)^T. \tag{15}$$

where $\widehat{U}(\lambda)$ is a $q \times (p+q+r)$ and $\widehat{V}(\mu)$ is a $p \times (p+q+r)$ matrix polynomial.

Then, the following theorem provides a characterization of the pseudoinverse of a Toeplitz-mosaic matrix, which means that actually computing the pseudoinverse can be performed efficiently, via several convolutions (in our case $p = q = K$, the number of kernels).

Theorem 1: [11] The Moore-Penrose inverse of a $(p, q)-$Toeplitz mosaic matrix is a Toeplitz $(q, p, q+p)-$Bezoutian.

A different, but equally attractive solution to the structured problem above is suggested by the approaches in [29], [30]. Namely, they solve Toeplitz least squares problems by translating it into an interpolation problem and using a superfast method to handle this one. Finding generators for the displacement of our particular Toeplitz-mosaic matrix is rather straightforward thanks to the algorithm in the appendix of [14]; this helps us reduce our own problem to a similar interpolation problem, which therefore can be solved efficiently.

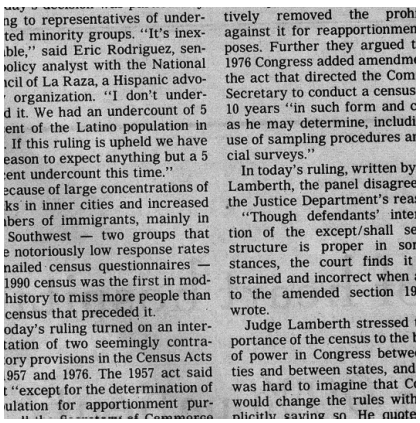
V. Experimental Results

In this section, we shall present the results of applying the above method to fairly different categories of images, to illustrate the importance of the signal class on the computed dictionary.

Natural Images. We first apply the method presented here to images from the Kyoto natural image database [8]. The dictionary started out as a set of 20 random 10×10 patches, and evolved as some of the kernels grew or shrank. A subset of kernels is displayed in Figure 1d; the learned kernels display

[3]We find it is useful to point out that the 1D correspondent is a block-row matrix with Toeplitz blocks (also known as Toeplitz-striped matrix).

(a) Natural image example.



ng to representatives of under-
ted minority groups. "It's inex-
ible," said Eric Rodriguez, sen-
olicy analyst with the National
ncil of La Raza, a Hispanic advo-
organization. "I don't under-
d it. We had an undercount of 5
ent of the Latino population in
. If this ruling is upheld we have
eason to expect anything but a 5
ent undercount this time."
ecause of large concentrations of
ks in inner cities and increased
bers of immigrants, mainly in
Southwest — two groups that
e notoriously low response rates
nailed census questionnaires —
1990 census was the first in mod-
history to miss more people than
census that preceded it.
oday's ruling turned on an inter-
tation of two seemingly contra-
ory provisions in the Census Acts
957 and 1976. The 1957 act said
"except for the determination of
ulation for apportionment pur-

tively removed the proh
against it for reapportionmen
poses. Further they argued t
1976 Congress added amendm
the act that directed the Com
Secretary to conduct a census
10 years "in such form and c
as he may determine, includi
use of sampling procedures a
cial surveys."

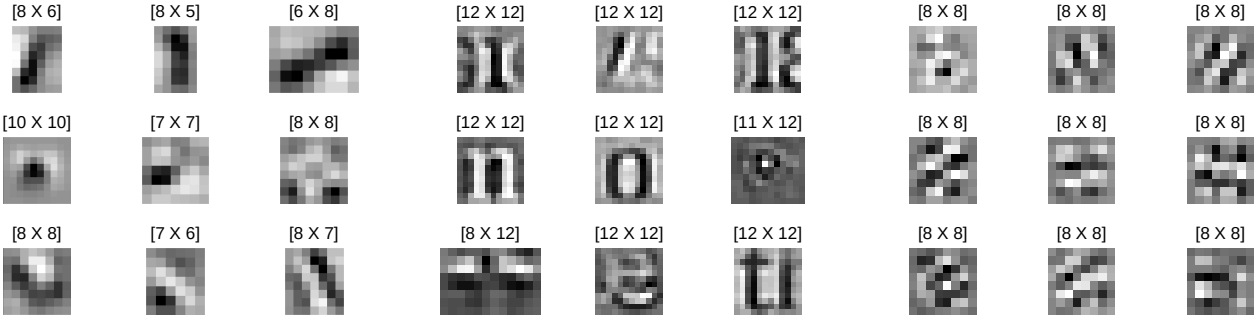
In today's ruling, written by
Lamberth, the panel disagree
the Justice Department's rea

"Though defendants' inte
tion of the except/shall se
structure is proper in so
stances, the court finds it
strained and incorrect when
to the amended section 19
wrote.

Judge Lamberth stressed
portance of the census to the
of power in Congress betwe
ties and between states, and
was hard to imagine that C
would change the rules with
plicitly saying so. He quote

(b) Newspaper image example

(c) Fingerprint image example



(d) Natural image kernels (after 300 iterations) (e) Newspaper image kernels (after 55 iterations) (f) Fingerprint image kernels (after 12 iterations)

Fig. 1. Results of adapting the kernels for three categories of images. First row: Image examples from each class. Second row: Subset of kernels adapted to that class. Kernels are up-scaled for a better visualization; actual pixel size is displayed above each kernel subplot.

the expected aspects for such a dataset (namely edges, ridges, cross patterns) but they also include other shapes (*e.g.*, "round" edges). One interesting aspect is that the kernels in the final dictionary (here, after 300 iterations) do not seem extremely sensitive to the starting point.

Newspaper Images. A different type of signals, with significantly distinct statistical structure is the class of scanned newspaper images. Particularities of this class mainly involve a more reduced set of predominant orientations. Figure 1e exhibits kernels adapted to this class, after 55 iterations and starting from random. As can be easily observed, kernels tend to capture mainly printed symbols; if a large enough set of kernels is used (*e.g.*, 40), they tend to stabilize to individual characters, or pairs thereof.

Fingerprint Images. Finally, we chose to apply our method on another highly distinct class, namely fingerprint images. Figure 1f displays a set of kernels learned from images in the Cross Match Verifier 300 sample fingerprint database [21]. Although initialized randomly, after only 12 iterations the 25 kernels already localize in frequency and orientation, although not also in space (easily explained by the structure of the signals).

The presented results have been learned form a set of training images ranging from 10 (newspaper) to 50 (natural images). The dictionary size was hand-picked to the reported values in order to avoid redundancy (*e.g.*, several kernels being copies or shifted versions of each other). We currently work on a mechanism to automatically control the number of "sufficient" kernels.

VI. Conclusion

We proposed an approach to deriving adaptive shift-invariant image representation. This method is computationally very efficient and enables eliminates kernel size constraints: they can have arbitrary lengths, which can lead to a multiscale dictionary. In the kernel update step, we focused on what we believe to be an under-explored family of algorithms, which exploit the structure of the least-squares optimization problem.

References

[1] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM J. Imaging Sci.*, vol. 1, no. 3, pp. 228–247, 2008.

[2] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Proc.*, vol. 8, no. 12, pp. 1688–1701, 1999.

[3] E. Candès and D. Donoho, "Ridgelets: a key to higher-dimensional intermittency?" *Phil. Trans. R. Soc. Lond. A.*, vol. 357, pp. 2495–2509, 1999.

[4] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying, "Fast discrete curvelet transforms," *Multisc. Model. Simul.*, vol. 5, pp. 861–899, 2005.

[5] S. Chen, D. Donoho, and M. Saunders, "Atomic decompositions by basis pursuit," *SIAM J. of Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1998.

[6] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *J. of Constr. Approx.*, pp. 57–98, 1997.

[7] M. Do and M. Vetterli, "Contourlets," in *Beyond Wavelets*, G. Welland, Ed. Academic Press, 2003.

[8] E. Doi, T. Inui, T.-W. Lee, T. Wachtler, and T. J. Sejnowski, "Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes," *Neural Comp.*, vol. 15, pp. 397–417, 2003, http://www.cnbc.cmu.edu/cplab/data_kyoto.html.

[9] D. Donoho, M. Vetterli, R. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Th.*, vol. 44, pp. 2435–2476, Oct. 1998.

[10] R. Grosse, R. Raina, H. Kwong, and A. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. UAI*, 2007.

[11] G. Heinig, "Generalized Inverses of Hankel and Toeplitz Matrices," *Lin. Alg. Appl.*, vol. 216, pp. 43–59, 1995.

[12] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*. Birkhäuser, 1984.

[13] P. Jost, P. Vandergheynst, S. Lesage, and R. Gribonval, "MoTIF: an efficient algorithm for learning translation invariant dictionaries," in *IEEE ICASSP*, 2006.

[14] T. Kailath and J. Chun, "Generalized Displacement Structure for Block-Toeplitz,Toeplitz-Block, and Toeplitz-Derived Matrices," *SIAM J. Matrix Anal. Appl.*, vol. 15, no. 1, pp. 114–128, 1994.

[15] T. Kailath and A. H. Sayed, Eds., *Fast reliable algorithms for matrices with structure*. SIAM, 1999.

[16] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit made tractable," in *Proc. ICASSP*, vol. 3, 2006, pp. 496–499.

[17] B. Mailhé, S. Lesage, R. Gribonval, and F. Bimbot, "Shift-invariant dictionary learning for sparse representations: extending K-SVD," in *Proc. EUSIPCO*, 2008.

[18] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, no. 2, pp. 3397–3415, 1993.

[19] G. Monaci, F. Sommer, and P. Vandergheynst, "Learning sparse generative models of audiovisual signals," in *EUSIPCO'08*, 2008.

[20] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[21] "Cross Match Verifier 300 sample fingerprint database," http://neurotechnology.com/download/CrossMatch_Sample_DB.zip, Neurotechnology.

[22] B. Olshausen, "Sparse codes and spikes," in *Probabilistic Models of the Brain: Perception and Neural Function*, R. Rao, B. Olshausen, and M. Lewicki, Eds., 2002, pp. 257–272.

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

[24] B. A. Olshausen, P. Sallee, and M. S. Lewicki, "Learning sparse images codes using a wavelet pyramid architecture," in *Advances in Neural Information Processing Systems*, vol. 12. MIT Press, 2000.

[25] P. Sallee, "Statistical Methods for Image and Signal Processing," Ph.D. dissertation, University of California, Davis, 2004.

[26] E. Smith and M. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computation*, vol. 17, no. 1, pp. 19–45, 2005.

[27] ——, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, 2006.

[28] D. Taubman and M. Marcellin, Eds., *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Springer, 2001.

[29] M. Van Barel, G. Heinig, and P. Kravanja, "An algorithm based on orthogonal polynomial vectors for Toeplitz least squares problems," in *Numerical analysis and its applications (NAA 2000)*, 2001, pp. 27–34.

[30] ——, "A superfast method for solving Toeplitz linear least squares problems," *Lin. Alg. Appl.*, vol. 366, pp. 441–457, 2003.