MORPHOLOGICAL DIVERSITY AND SPARSE IMAGE DENOISING

M.J. Fadili[a], *J.-L. Starck*[b] *and L. Boubchir*[a]

[a] GREYC UMR CNRS 6072
14050 Caen France

[b] DAPNIA/SEDI-SAP CEA-Saclay
91191 Gif-sur-Yvette France

ABSTRACT

Overcomplete representations are attracting interest in image processing theory, particularly due to their potential to generate *sparse* representations of data based on their *morphological diversity*. We here consider a scenario of image denoising using an *overcomplete* dictionary of sparse linear transforms. Rather than using the basic approach where the denoised image is obtained by simple averaging of denoised estimates provided by each sparse transform, we here develop an elegant bayesian framework to optimally combine the individual estimates. Our derivation of the optimally combined denoiser relies on a scale mixture of gaussian (SMG) prior on the coefficients in each representation transform. Exploiting this prior, we design a bayesian ℓ_2-risk (mean field) nonlinear estimator and we derive a closed-form for its expression when the SMG specializes to the Bessel K form prior. Experimental results are carried out to show the striking profits gained from exploiting sparsity of data and their morphological diversity.

Index Terms— Sparsity, Morphological diversity, Bayesian combined denoising.

1. INTRODUCTION

Recently, researchers spanning a diverse range of viewpoints have advocated the use of overcomplete signal/image representations (see e.g. [1, 2, 3]). Generally speaking, they suppose we have a an image vector $\mathbf{x} \in \mathbb{R}^n$, and a collection of vectors $(\varphi)_{\gamma \in \Gamma}$, Card $\Gamma = L$, with $L \geq n$, such that the image $\mathbf{x}$ can be written as the superposition of these elementary atoms $\mathbf{x} = \sum_{\gamma \in \Gamma} \alpha_\gamma \varphi_\gamma = \Phi \boldsymbol{\alpha}$.

Popular examples of Γ include: frequency (Fourier), scale-translation (wavelets), scale-translation-frequency (wavelet packets), translation-duration-frequency (cosine packets), scale-translation-angle (e.g. curvelets, bandlets, contourlets, etc.). The dictionary Φ is the $n \times L$ matrix whose columns are the generating atoms $\{\varphi_\gamma\}_{\gamma \in \Gamma}$, which are supposed to be normalized to a unit ℓ_2-norm. The forward transform is defined via a non-necessarily square full rank matrix $\mathbf{T} = \Phi^T \in \mathbb{R}^{L \times n}$, with $L \geq N$. When $L > n$ the dictionary is said to be redundant or overcomplete. Such representations differ from the more traditional basis representation because they offer a wider range of generating elements; potentially, this wider range allows more flexibility in signal representation and adaptativity to its morphological content, and hence more effectiveness at tasks like signal compression and restoration.

Owing to recent advances in modern harmonic analysis, many redundant systems, like the undecimated wavelet transform or curvelet pyramids, were shown to be very effective in sparsely representing images. By sparsity, we mean that we are seeking a good representation of $\mathbf{x}$ with only very few non-zero coefficients, i.e. $\|\boldsymbol{\alpha}\|_0 \ll n$. In most practical situations, the dictionary is built by taking union of one or several (sufficiently incoherent) transforms, generally each corresponding to an orthogonal basis or a tight frame. Choosing an appropriate dictionary is a key step towards a good sparse representation, hence recovery. A core idea here is the concept of morphological diversity, as initiated in [3]. When the transforms are amalgamated in one dictionary, they have to be chosen such that each leads to sparse representations over the parts of the images it is serving. Thus, to represent efficiently isotropic structures in an image, a qualifying choice is the wavelet transform [4]. The curvelet system [5, 6] is a very good candidate for representing piecewise smooth (C^2) images away from C^2 contours. The ridgelet transform [7, 8] has been shown to be very effective for representing global lines in an image. The local DCT [4] is well suited to represent locally stationary textures. These transforms are also computationally tractable particularly in large-scale applications. The associated implicit fast analysis and synthesis operators have typically complexities of $O(n)$ (e.g. orthogonal or bi-orthogonal wavelet transform) or $O(n \log n)$ (e.g. ridgelet, curvelet, local DCT transforms). Another desirable requirement that the merged transforms have to satisfy is that when a transform sparsely represents a part in the image, it yields non-sparse representations on the other content type.

Let's now consider the problem of recovering the sparsest representation possible in an overcomplete dictionary Φ, in presence of an additive gaussian white noise with variance σ_ε^2:

$$\mathbf{y} = \Phi \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \qquad (1)$$

The sparsest representation is then the solution to the optimization problem:

$$P_0 : \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \text{ s.t. } \|\mathbf{y} - \Phi \boldsymbol{\alpha}\|_2 \leq \delta \qquad (2)$$

This is an NP-hard combinatorial optimization problem. Au-

1-4244-0728-1/07/$20.00 ©2007 IEEE

thors in [2] proposed a convexified and relaxed form:

$$P_1 : \min_{\boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{y} - \Phi\boldsymbol{\alpha}\|_2 + \lambda \|\boldsymbol{\alpha}\|_1 \qquad (3)$$

(P_1) is a linear programming problem known as BPDN. It solves (P_0) under appropriate conditions [2]. Very recently, considerable attention has focused on theoretical and practical issues related to solving the underdetermined sparse solution problem (P_1) (see e.g. [9, 10, 11] and many others).

(P_1) can also be seen as the MAP estimator with a Laplacian prior on the coefficients. The core of our proposal is (i) to replace the ℓ_1-norm prior with a flexible family of sparsity promoting priors based on the SMG, and (ii) to replace the MAP estimator by the ℓ_2-risk mean-field (MF) estimator; that is we substitute integration (MF) for optimization (MAP). Rather than using the basic approach where the denoised image is obtained by simple averaging of denoised estimates in each sparse transform, we here develop a bayesian framework to optimally combine the individual estimates. Our derivation relies on a scale mixture of gaussians (SMG) prior on the coefficients in each representation transform. More specifically, we use a special instance of the SMG, namely the BKF prior, for which we have derived a closed-form for its MF estimator in a previous contribution [12]. Experimental results are carried out to show the effectiveness and performance of our method.

2. STATISTICAL PRIOR

In the bayesian approach a prior distribution is imposed on the representation coefficients in order to capture the sparseness of the linear expansion. The following section is intended to provide an introduction to SMG family suitable to characterize the sparse representation coefficient densities which have been already observed to be sharply peaked and heavily tailed. The SMG model have been already used in [13], and some of its instances have also been proposed in different works (see e.g. [14, 12]).

2.1. Scale Mixture of Gaussians (SMG)

Definition 1 ([15]) *Let X be a random variable (RV) with real-valued realizations. Under the SMG, there exist two independent RVs $U \geq 0$ and $Z \sim \mathcal{N}(0, 1)$ such that:*

$$X \overset{d}{=} Z\sqrt{U} \qquad (4)$$

Property 1 *SMG is a subset of the elliptically symmetric distributions. The pdf f_X exists at 0 if only if $\mathbb{E}[U^{-1/2}] < +\infty$. The pdf f_X is unimodal, symmetric around the mode and differentiable almost everywhere. Moreover, if f_U the pdf of U is differentiable, then:*

$$f_U(u) = \sqrt{\frac{\pi}{2}} u^{-3/2} \zeta\left((2u)^{-1}\right) \qquad (5)$$

where $\zeta(.)$ is the inverse Laplace transform of the pdf $f_X\left(\sqrt{.}\right)$.

The following lemma establishes that such a representation is adapted as a sparsity-promoting prior:

Lemma 1

(i) The RV X has a SMG representation if and only if the k^{th} derivatives of $f_X(\sqrt{y})$ have alternating sign, i.e.:

$$\left(-\frac{d}{dy}\right)^k f_X\left(\sqrt{y}\right) \geq 0 \;\; \forall y > 0 \qquad (6)$$

(ii) If $U \geq 0$ is random, then kurtosis$(X) > 0 \Longrightarrow$ the symmetric distribution of X is necessarily sharply peaked (leptokurtic) with heavy tails.

Proof: The first statement is due to [15]. The second one is straightforward by marginalizing with respect to the mixing RV U.

As for sparse representations, the empirical coefficient pdfs were observed to be symmetric around 0, leptokurtic and heavy tailed, these pdfs have their 1st and 2nd derivatives of alternating signs on $\mathbb{R}^+$. Consequently, Lemma 1 states that the SMG family fulfills all necessary requirements to capture the sparsity of decompositions and is then legitimate as a prior for the coefficients. Note also that a key advantage of SMG is that it transfers all desirable properties of the gaussian distribution through the mixing RV.

2.2. The BKF prior

The Bessel K form (BFK) prior [12] is a particular instance of the SMG family where the mixing RV U in Eq.4 is Gamma distributed with a shape parameter β and a scale parameter c. Our interest in this distribution relies on two facts: (i) the hyperparameters (β, c) associated to the BKF can be easily estimated even in presence of noise using either a cumulant or an EM-based estimator, (ii) the posterior-conditional mean (or mean-field) estimator associated the ℓ_2-bayesian risk has a closed form expression. See [12] for details.

3. BAYESIAN COMBINED DENOISING

Here, we describe our approach to optimally combine individual estimates obtained with each transform separately. This is accomplished in a bayesian framework where the optimal weights are derived. The overall strategy is sketched in Fig.1.

For the sake of simplicity, the dictionary is now supposed to be a union of M (sufficiently incoherent) bases $\{\Phi_m\}_{m=1,\ldots,M}$ (that is $L = Mn$). It is easy to see that the (merged) estimate at spatial location s can be written as a function of the individual estimates as follows:

$$\hat{\mathbf{x}}(s) = \sum_{m=1}^{M} \hat{\mathbf{x}}_m(s) P\left(\Phi_m | \mathbf{y}(s); \mathbf{U}_\gamma^m\right) \qquad (7)$$

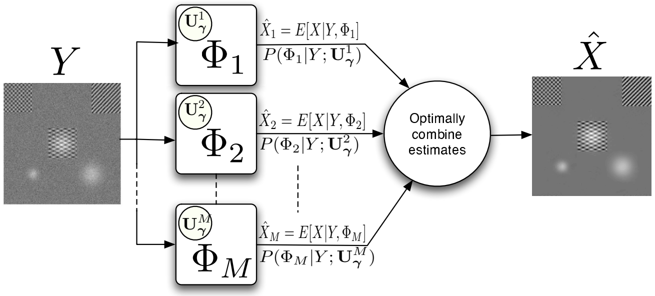
Y
$\mathbf{U}_\gamma^1$
Φ_1
$\hat{X}_1 = E[X|Y,\Phi_1]$
$P(\Phi_1|Y;\mathbf{U}_\gamma^1)$
$\mathbf{U}_\gamma^2$
Φ_2
$\hat{X}_2 = E[X|Y,\Phi_2]$
$P(\Phi_2|Y;\mathbf{U}_\gamma^2)$
$\mathbf{U}_\gamma^M$
Φ_M
$\hat{X}_M = E[X|Y,\Phi_M]$
$P(\Phi_M|Y;\mathbf{U}_\gamma^M)$
Optimally combine estimates
$\hat{X}$

Fig. 1. The flowchart of bayesian combined denoising.

where $\hat{\mathbf{x}}_m$ is the MF-based estimate of the image $\mathbf{x}$ obtained in the transform mth transform, and:

$$P\left(\Phi_m|\mathbf{y}(s);\mathbf{U}_\gamma^m\right) \propto \\ \int\cdots\int_0^{+\infty} P\left(\mathbf{y}(s)|\Phi_m,\mathbf{U}_\gamma^m\right) P\left(\mathbf{U}_\gamma^m|\Phi_m\right) P(\Phi_m)d\mathbf{U}_\gamma^m \tag{8}$$

Assume now that (i) the SMG prior is independent of the transform, and (ii) all transforms are equiprobable[1]. Then, using properties of the SMG, we obtain:

$$P\left(\Phi_m|\mathbf{y}(s),\mathbf{U}_\gamma^m\right) \propto \\ \int\cdots\int_0^{+\infty} \phi(\mathbf{y}(s);\mu_m(s),\tau_m(s)+\sigma_\varepsilon^2) f_{\mathbf{U}_\gamma^m}(\mathbf{u}_\gamma^m)\, du_1^m \ldots du_{|\Gamma_m|}^m \tag{9}$$

where $\phi(\mathbf{y};\mu,\sigma^2)$ stands for the normal pdf with mean μ and variance σ^2, μ_m is the low-pass component in the transform Φ_m that one may want to keep intact (vague prior). $f_{\mathbf{U}_\gamma^m}(\mathbf{u}_\gamma^m)$ is the joint pdf of the SMG prior mixing variables in the mth transform. $\tau_m(s)$ is the variance in the spatial domain:

$$\tau_m(s) = \sum_{\gamma_m\in\Gamma_m} \left|\varphi_{\gamma_m}(s)\right|^2 u_\gamma^m \tag{10}$$

Eq.9 is computationally intractable since it necessitates numerical integration to compute in practice. Nonetheless, this difficulty can be alleviated by assuming that the SMG priors have mixing RVs $\left(U_\gamma^m\right)_{m=1,\ldots,M,\gamma\in\Gamma_m}$ that are subband-independent and rapidly decreasing pdfs (point mass at the mode). Hence, Eq.9 simplifies to:

$$P(\Phi_m|\mathbf{y}(s),\mathbf{U}_\gamma^m) = \frac{\phi(\mathbf{y}(s);\mu_m(s),\tau_m(s)+\sigma_\varepsilon^2)}{\sum_{m'=1}^M \phi(\mathbf{y}(s);\mu_{m'}(s),\tau_{m'}(s)+\sigma_\varepsilon^2)} \tag{11}$$

where in the expression of $\tau_m(s)$, the mixing RV realization u_γ^m is replaced with the mode $\hat{u}_\gamma^m$. For example, in the case of the BKF prior, the mode is easily expressed in terms of the prior hyperparameters $\hat{u}_\gamma^m = \max\left((\beta_\gamma^m - 1)c_\gamma^m, 0\right)$.

Let's now turn to the expression of τ_m. To be computed, this will necessitate to have the atoms φ_{γ_m} available. But, as stated in the introduction, the dictionaries $(\Phi_m)_m$ are never constructed explicitly which would be otherwise computationally prohibitive (implicit fast analysis and reconstruction operators are used instead). Fortunately, for many usual transform bases (e.g. DCT, Haar basis), τ_m is exactly:

$$\tau_m = \sum_{\gamma_m} \hat{u}_\gamma^m \tag{12}$$

For other bases (wavelet transform with other wavelets than Haar), this expression is not exact but can be shown to be relevant to a good approximation.

4. EXPERIMENTAL RESULTS

The performance of the combined approach has been assessed on several 2D datasets, from which we here illustrate two examples. The first one is a synthetic image containing textured areas and gaussians. The dictionary naturally contained the wavelet transform (with Symmlet 4 QMF) and a local DCT (actually a local cosine packet basis at a fixed depth). Fig.2 shows the denoising results using ad hoc averaging of individual estimates and the combined adaptive strategy. The wavelet transform alone performs well on the smooth parts, while the local DCT alone is better in the textured areas. The combined approach performs well in both parts (a gain of 2dB was observed compared to single transform denoising). The combined approach also outperforms the simple averaging method. Furthermore, one may notice how the conditional probability map of the first transform (wavelet) reflects the spatial distribution of the gaussians. These results are confirmed when applied to the more complicated Barbara image. The dictionary contained the same transforms as before. A gain of 1dB was achieved for this image compared to a single transform denoising.

5. CONCLUSION

In this paper, a bayesian MF estimator exploiting sparsity in several transforms is proposed. The denoiser optimally combines the transforms in the dictionary instead of ad hoc averaging. As each transform is intended to sparsely represent certain parts of the image its is serving, our approach associates the advantages of all these representations, and thus, will perform well over all the image. A byproduct of the method is a (posterior) conditional probability map attached to each transform that may be useful for other purposes (e.g. separation). The performance of the approach clearly demonstrates the benefit of overcomplete over single transform denoising. This also confirms the striking profits gained from exploiting sparsity of data and their morphological diversity.

6. REFERENCES

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

[1] Alternatively, a prior probability map of the transforms that has been provided by a learning step can be easily fed into the above expression.

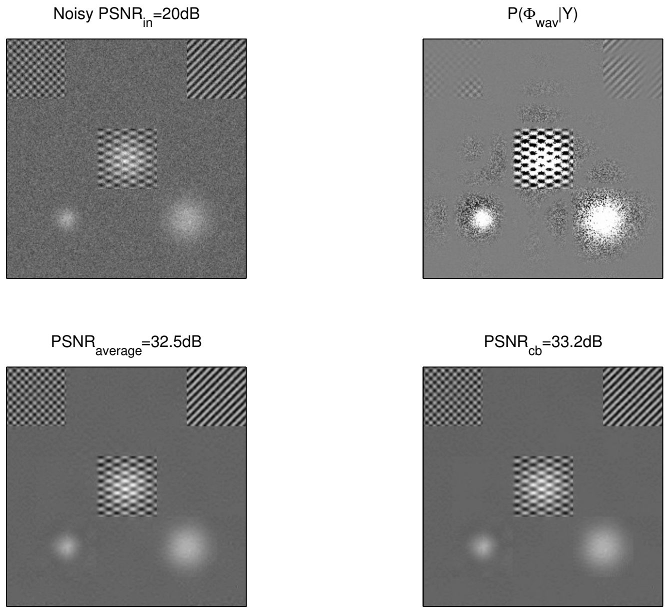

Fig. 2. Illustration on a synthetic example (texture+gaussian image with PSNR$_{in}$ = 20dB). From top to bottom and left to right: noisy, posterior probability map of the wavelet transform, combined estimate by averaging, optimally combined estimate.

[2] S.S. Chen, D.L. Donoho, and M.A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[3] J.-L. Starck, M. Elad, and D. Donoho, “Image decomposition via the combination of sparse representatntions and variational approach,” *IEEE Trans. Im. Proc.*, vol. 14, no. 10, pp. 1570–1582, 2005.

[4] S. G. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, 1998.

[5] E. J. Candès and D. L. Donoho, “Curvelets – a surprisingly effective nonadaptive representation for objects with edges,” in *Curve and Surface Fitting: Saint-Malo 1999*, A. Cohen, C. Rabut, and L.L. Schumaker, Eds., Nashville, TN, 1999, Vanderbilt University Press.

[6] E.J. Candès, L. Demanet, D. Donoho, and L. Ying, “Fast discrete curvelet transforms,” Tech. Rep., CalTech, Applied and Computational Mathematics, 2005, To appear in SIAM Multiscale Model. Simul.

[7] E.J. Candès and D.L. Donoho, “Ridgelets: the key to high dimensional intermittency?,” *Philosophical Transactions of the Royal Society of London A*, vol. 357, pp. 2495–2509, 1999.

[8] J. L. Starck, E. Candes, and D. L. Donoho, “The curvelet transform for image denoising,” *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 670–684, 2002.

[9] D.L. Donoho and M. Elad, “Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization,” *Proc. Nat. Aca. Sci.*, vol. 100, pp. 2197–2202, 2003.

[10] A.M. Bruckstein and M. Elad, “A generalized uncertainty principle and sparse representation in pairs of $\mathbf{r}^n$ bases,” *IEEE Transactions on Information Theory*, vol. 48, pp. 2558–2567, 2002.

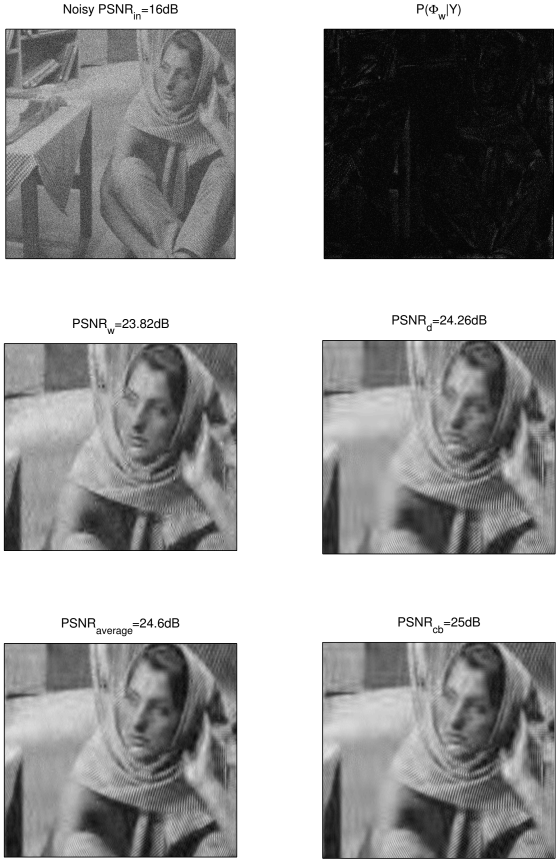

Fig. 3. Illustration on the Barbara image with PSNR$_{in}$ = 16dB) ($\sigma_\varepsilon = 40$). From top to bottom and left to right: noisy, ”wavelet” probability map, estimate in wavelets, estimate in local DCT, combined by averaging, optimally combined estimate.

[11] R. Gribonval and M. Nielsen, “Sparse representations in unions of bases,” *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[12] J. Fadili and L. Boubchir, “Analytical form for a bayesian wavelet estimator of images using the bassel k forms densities,” *IEEE Trans. Image Processing*, vol. 14, no. 2, pp. 231–240, 2005.

[13] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixture of gaussians in the wavelet domain,” *IEEE Transaction on Image Processing*, vol. 12, no. 11, pp. 1338–1351, November 2003.

[14] P. Moulin and J. Liu, “Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors,” *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[15] D. F. Andrews and C. L. Mallows, “Scale mixtures of normality,” *Journal of the Royal Statistical Society, Series B 36*, pp. 99–102, 1974.