Image Denoising — The SURE-LET Methodology

Thierry Blu

Department of Electronic Engineering
The Chinese University of Hong Kong

APSIPA Annual Summit and Conference 2010

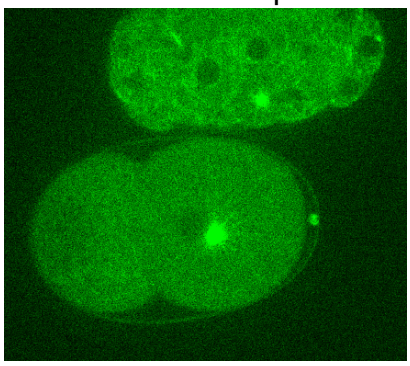
Tutorial prepared in collaboration with **Florian Luisier** *(SISL, Harvard Univ.)*

Outline

Outline

Noise in Images: Noise Sources

Noise: a random, undesirable, and often unavoidable perturbation.



Two main sources:

- Random nature of photon emission and detection;
- Imperfection of the electronic devices (photosensors, A/D converter,...).

Tremendous impact on image visualization and analysis (segmentation, tracking, recognition,...).

Noise in Images: Measurement Model

- Usual acquisition devices provide signals[1]

$$\mathbf{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}}$$

  that are corrupted with noise.

- Frequent modeling using an **additive white Gaussian noise (AWGN)** hypothesis

$$\underbrace{\mathbf{y}}_{\text{noisy signal}} = \underbrace{\mathbf{x}}_{\text{original signal}} + \underbrace{\mathbf{b}}_{\text{noise}}$$

  where $\mathscr{E}\{\mathbf{b}\} = \mathbf{0}$ and $\mathscr{E}\{\mathbf{b}\mathbf{b}^{\mathrm{T}}\} = \sigma^2\mathbf{Id}$.

- **Signal denoising** consists in finding a "good" candidate $\hat{\mathbf{x}}$ of $\mathbf{x}$ using **the noisy signal $\mathbf{y}$ only**; i.e., find the algorithm $\mathbf{F}$ such that

$$\hat{\mathbf{x}} = \mathbf{F}(\mathbf{y})$$

[1] Images are represented as *vectors*, using lexicographic ordering.

An Abundant Literature

Many approaches available, based on:

1. **Explicit hypotheses on the signal:**
   - Statistics-based: wavelet-domain (Bayesian) inference Donoho *et al.* 1994, Simoncelli *et al.* 1996, Abramovich *et al.* 1998, Vidakovic *et al.* 1998;
   - Regularization: Total Variation (TV) Osher *et al.* 1992;
   - PDE: anisotropic diffusion Perona *et al.* 1990;
2. **Heuristics:**
   - Filtering: Bilateral Filter Tommasi *et al.* 1998;
   - Patch-based: Non-Local Means Buades *et al.* 2005;
   - Any combination of approaches 1 when the hypotheses are not satisfied/checked.

NOTE:

- Some approaches can be either applied in the *signal-domain* or in a *transform-domain*.
- Most approaches involve several *nonlinear* parameters which are often set *empirically*.

Prior-Based Statistical Approaches

In the prior-based statistical approaches the signal to restore is considered as the realization of a *random* variable.

Various possible objectives to optimize:

- Maximum a posteriori (MAP)
- Minimum mean-squared error (MMSE)

All these methods assume that the following are explicitly given:

- The statistical relation (likelihood) between the measurements and the signal to restore:

$$\mathscr{P}\{\mathbf{y}|\mathbf{x}\} = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\|\mathbf{y}-\mathbf{x}\|^2}{2\sigma^2}\right)$$

- The probability density function (pdf) of the original signal $\mathscr{P}\{\mathbf{x}\}$.

Highly sensitive to the modeling of the pdf of the signal to restore.

Maximum a Posteriori

The MAP consists in choosing the estimate $\hat{\mathbf{x}}$ that maximizes the *posterior probability density*

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathscr{P}\{\mathbf{x}|\mathbf{y}\} = \arg\max_{\mathbf{x}} \mathscr{P}\{\mathbf{y}|\mathbf{x}\} \cdot \mathscr{P}\{\mathbf{x}\}$$

Optimal detector: Given noisy measurements of a signal $\mathbf{x}$ having a finite number of values x_1, x_2, ..., x_K occurring with probabilities p_1, p_2,..., p_K, the MAP minimizes the error probability

$$\mathscr{P}\{\hat{\mathbf{x}} \neq \mathbf{x}\}$$

NOTE: Description of the prior $\mathscr{P}\{\mathbf{x}\}$ may require many nonlinear parameters.

For signals with large or infinite number of levels, the probabilistic optimality of the MAP becomes irrelevant $\rightsquigarrow$ MMSE instead.

Linear MMSE: Wiener

The Wiener "filter" consists in finding the linear[2] estimate, $\hat{\mathbf{x}} = \hat{\mathbf{A}}\mathbf{y}$, that minimizes the *Mean-Squared Error* (MSE)

$$\underbrace{\mathscr{E}\left\{\frac{1}{N}\|\hat{\mathbf{A}}\mathbf{y}-\mathbf{x}\|^2\right\}}_{\text{MSE between } \hat{\mathbf{x}} \text{ and } \mathbf{x}} = \min_{\mathbf{A}} \mathscr{E}\left\{\frac{1}{N}\|\mathbf{A}\mathbf{y}-\mathbf{x}\|^2\right\}$$

Solution: Requires only the knowledge of the covariance matrix $\mathbf{\Gamma_x} = \mathscr{E}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ of the original signal

$$\mathbf{x} = \mathbf{\Gamma_x}\left(\mathbf{\Gamma_x} + \sigma^2\mathbf{Id}\right)^{-1}\mathbf{y}$$

NOTE: Although very popular, linear processing is not well-adapted to the processing of transient signals.

[2] if $\mathscr{E}\{\mathbf{x}\} = \mathbf{0}$ — an affine estimate is used, otherwise.

Nonlinear MMSE: Bayesian Least Squares

Problem: Find the optimal processing $\mathbf{F}(\cdot)$ that yields the estimate $\hat{\mathbf{x}} = \mathbf{F}(\mathbf{y})$ such that

$$\mathscr{E}\left\{\frac{1}{N}\|\mathbf{F}(\mathbf{y})-\mathbf{x}\|^2\right\} \text{ is minimized.}$$

Solution: The posterior expectation (conditional mean):

$$\hat{\mathbf{x}} = \mathscr{E}\{\mathbf{x}|\mathbf{y}\} = \int \mathbf{x}\,\mathscr{P}\{\mathbf{x}|\mathbf{y}\}\,\mathrm{d}^N\mathbf{x} \overset{\text{Bayes}}{=} \frac{1}{\mathscr{P}\{\mathbf{y}\}}\int \mathbf{x}\,\mathscr{P}\{\mathbf{y}|\mathbf{x}\}\cdot\mathscr{P}\{\mathbf{x}\}\,\mathrm{d}^N\mathbf{x}$$

where $\mathscr{P}\{\mathbf{y}\} = \int \mathscr{P}\{\mathbf{y}|\mathbf{x}\}\cdot\mathscr{P}\{\mathbf{x}\}\,\mathrm{d}^N\mathbf{x}$ is the marginal pdf of $\mathbf{y}$.

NOTE: The above integrals often need to be computed numerically.

The Bayesian MMSE algorithm requires the knowledge of the *pdf of the unknown signal* $\rightsquigarrow$ **Choice of prior ?**

Nonlinear MMSE: One Step Further

Problem: Find the optimal processing $\mathbf{F}(\cdot)$ that yields the estimate $\hat{\mathbf{x}} = \mathbf{F}(\mathbf{y})$ such that

$$\mathscr{E}\left\{\frac{1}{N}\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\|^2\right\} \text{ is minimized.}$$

Solution: In the case of AWGN, the posterior expectation $\hat{\mathbf{x}} = \mathscr{E}\{\mathbf{x}|\mathbf{y}\}$ can be simplified to (Stein 1981, Raphan & Simoncelli 2007):

$$\hat{\mathbf{x}} = \mathbf{y} + \sigma^2 \nabla \log \mathscr{P}\{\mathbf{y}\}$$

NOTE: Because $\mathscr{P}\{\mathbf{y}\} = \overbrace{\int \mathscr{P}\{\mathbf{y}|\mathbf{x}\} \cdot \mathscr{P}\{\mathbf{x}\}\, \mathrm{d}^N\mathbf{x}}^{\text{convolution with a Gaussian}}$, the optimal MSE processing is infinitely differentiable.

The optimal algorithm only requires the knowledge of the *pdf of the observed noisy signal* $\rightsquigarrow$ **No prior information is needed !**

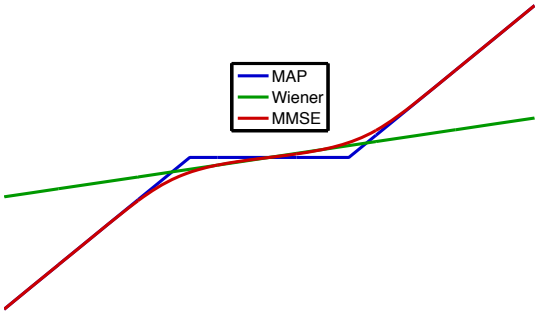
Examples

Assuming a Laplace prior, $\mathscr{P}\{\mathbf{x}\} = \prod_{n=1}^{N} \frac{\lambda}{2} \mathrm{e}^{-\lambda|x_n|}$, these statistical approaches yield a pointwise thresholding involving $T = \lambda\sigma^2$:

$$\text{MAP} \quad \hat{x}_n = \mathrm{soft}_T(y_n)$$

$$\text{Wiener} \quad \hat{x}_n = \frac{y_n}{1 + \frac{T^2}{2\sigma^2}}$$

$$\text{MMSE} \quad \hat{x}_n = y_n - T\frac{\mathrm{e}^{-\lambda y_n}\,\mathrm{erfc}\left(\frac{-y_n+T}{\sigma\sqrt{2}}\right) - \mathrm{e}^{\lambda y_n}\,\mathrm{erfc}\left(\frac{y_n+T}{\sigma\sqrt{2}}\right)}{\mathrm{e}^{-\lambda y_n}\,\mathrm{erfc}\left(\frac{-y_n+T}{\sigma\sqrt{2}}\right) + \mathrm{e}^{\lambda y_n}\,\mathrm{erfc}\left(\frac{y_n+T}{\sigma\sqrt{2}}\right)}$$



Regularization Approaches

The signal estimate $\hat{\mathbf{x}}$ is selected as the minimizer of a (convex) regularized cost-functional

$$J(\mathbf{x},\mathbf{y}) = \underbrace{\Psi(\mathbf{x},\mathbf{y})}_{\text{data-fidelity term}} + \lambda \underbrace{\Phi(\mathbf{x})}_{\text{penalty}}$$

Typical choice of data-fidelity term:

$$\Psi(\mathbf{x},\mathbf{y}) = \|\mathbf{y}-\mathbf{x}\|^2 \propto \text{ negative log-likelihood (AWGN)}$$

Typical choices of penalty:

- Tikhonov (smoothness prior): $\Phi(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|^2$;
- Sparsity prior: $\Phi(\mathbf{x}) = \|\mathbf{x}\|_{\ell_0} \rightsquigarrow \Phi(\mathbf{x}) = \|\mathbf{x}\|_{\ell_1}$;
- TV (edge prior): $\Phi(\mathbf{x}) = \| \, |\nabla \mathbf{x}| \, \|_{\ell_1}$.

NOTE: Depending on the choice of data-fidelity and penalty terms, $J(\mathbf{x},\mathbf{y})$ can be re-interpreted as a *statistical prior* and its optimization equivalent to a MAP.

No explicit distance minimization between original and denoised signal.

Minimizing $\mathscr{E}\left\{\|\mathbf{F}(\mathbf{y})-\mathbf{x}\|^2\right\}$ yields an algorithm $\mathbf{F}: \mathbf{y} \mapsto \hat{\mathbf{x}}$ that *depends on the probability of* $\mathbf{y}$ *alone*: $\mathbf{F}(\mathbf{y}) = \mathbf{y} + \sigma^2 \nabla \log \mathscr{P}\{\mathbf{y}\}$.

Problem: we have only one realization of the noisy image $\mathbf{y}$.
Solution: *estimate* $\mathscr{E}\left\{\|\mathbf{F}(\mathbf{y})-\mathbf{x}\|^2\right\}$ from $\mathbf{y}$, instead of $\mathscr{P}\{\mathbf{y}\}$.

MSE estimation

Consider the random variable[a]

$$\text{SURE}(\mathbf{y}) = \frac{1}{N}\|\mathbf{F}(\mathbf{y})-\mathbf{y}\|^2 + \frac{2\sigma^2}{N}\text{div}\{\mathbf{F}(\mathbf{y})\} - \sigma^2$$

Under the *additive white Gaussian noise* hypothesis, this random variable is an *unbiased estimate of the MSE* *Stein et al. 1981*

$$\mathscr{E}\{\text{SURE}(\mathbf{y})\} = \mathscr{E}\left\{\|\mathbf{F}(\mathbf{y})-\mathbf{x}\|^2/N\right\}$$

[a] Divergence operator: $\text{div}\{\mathbf{F}(\mathbf{y})\} \stackrel{\text{def}}{=} \sum_k \frac{\partial F_k(\mathbf{y})}{\partial y_k}$.

The original signal $\mathbf{x}$ may, or may not be random.
No assumptions on $\mathbf{x}$ are needed.

A simple proof

On the one hand (remember that $\mathbf{y} = \mathbf{x} + \mathbf{b}$)

$$\mathscr{E}\left\{\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\|^2\right\} = \mathscr{E}\left\{\|\mathbf{F}(\mathbf{y})\|^2\right\} - 2\underbrace{\mathscr{E}\left\{\mathbf{x}^{\mathrm{T}}\mathbf{F}(\mathbf{y})\right\}}_{\mathscr{E}\{(\mathbf{y}-\mathbf{b})^{\mathrm{T}}\mathbf{F}(\mathbf{y})\}} + \underbrace{\|\mathbf{x}\|^2}_{\mathscr{E}\{\|\mathbf{y}\|^2\}-N\sigma^2}$$
$$= \mathscr{E}\left\{\|\mathbf{F}(\mathbf{y}) - \mathbf{y}\|^2\right\} + 2\mathscr{E}\left\{\mathbf{b}^{\mathrm{T}}\mathbf{F}(\mathbf{y})\right\} - N\sigma^2$$

and on the other hand (*Stein's Lemma*)

$$\mathscr{E}\left\{\mathbf{b}^{\mathrm{T}}\mathbf{F}(\mathbf{y})\right\} = \int \underbrace{\mathscr{P}\{\mathbf{b}\}\,\mathbf{b}^{\mathrm{T}}}_{-\sigma^2\nabla\mathscr{P}\{\mathbf{b}\}^{\mathrm{T}}}\mathbf{F}(\mathbf{x}+\mathbf{b})\,\mathrm{d}^N\mathbf{b} \qquad \text{(Gaussian pdf)}$$
$$= \int \sigma^2\,\mathscr{P}\{\mathbf{b}\}\,\mathrm{div}\{\mathbf{F}(\mathbf{x}+\mathbf{b})\}\,\mathrm{d}^N\mathbf{b} \qquad \text{(by parts)}$$
$$= \mathscr{E}\left\{\sigma^2\mathrm{div}\{\mathbf{F}(\mathbf{y})\}\right\}$$

Equivalence SURE-MSE

SURE$(\mathbf{y})$ has a small variance (law of large numbers: $\propto 1/N$), which implies SURE$(\mathbf{y}) \approx \mathscr{E}\{\text{SURE}(\mathbf{y})\}$. Hence

$$\frac{1}{N}\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\|^2 \approx \text{SURE}(\mathbf{y})$$

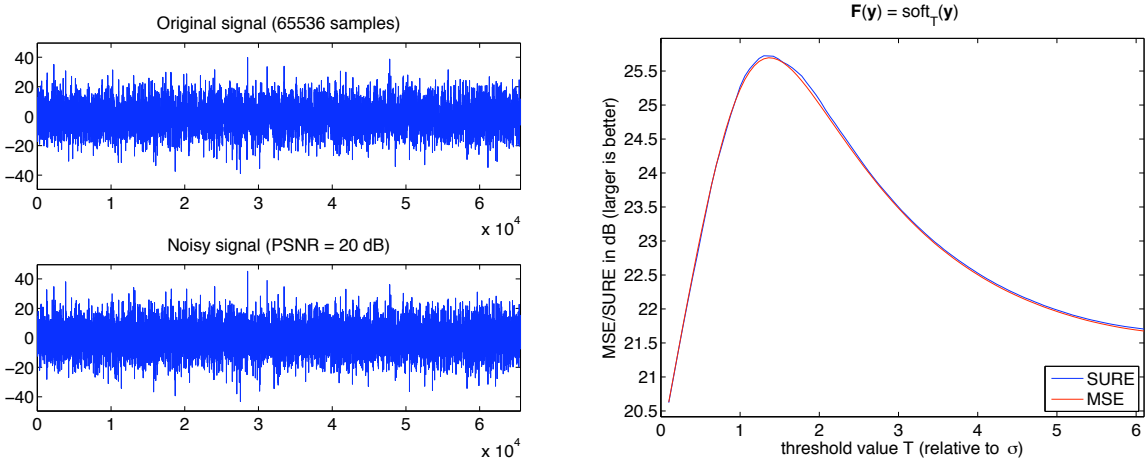
NOTE: The SURE–MSE match worsens when $\mathbf{F}(\mathbf{y})$ is less *regular*; some boundedness of $\mathrm{div}\{\mathbf{F}(\mathbf{y})\}$ is needed $\rightsquigarrow$ hard-threshold excluded.

Example Donoho 1995: SURE soft-threshold

$$\text{SURE}_{\text{soft}} = \frac{1}{N}\Bigg(\overbrace{\sum_{|y_n|\le T} y_n^2 + \sum_{|y_n|\ge T} T^2}^{\|\mathbf{F}(\mathbf{y})-\mathbf{y}\|^2} + 2\sigma^2\overbrace{\sum_{|y_n|\ge T} 1}^{\mathrm{div}\{\mathbf{F}(\mathbf{y})\}}\Bigg) - \sigma^2$$

Closeness between SURE and MSE

Processing a noisy signal (left) with several lengths, using several different pointwise thresholding functions



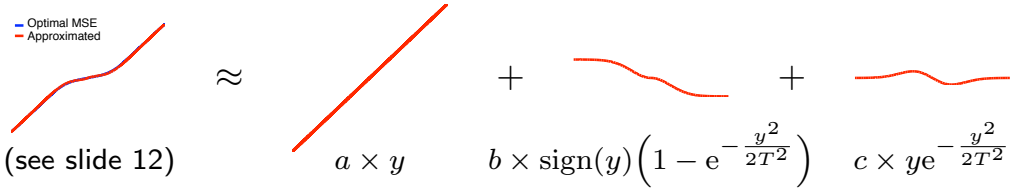
NOTE: The use of the SURE (instead of the MSE) is particularly justified for large data sizes (e.g., images).

Approximation of processings

Functions can often be efficiently approximated onto adapted bases.

Examples of bases: wavelets ($\mathbf{L}^2$ functions), sinc kernels (bandlimited functions), radial basis functions (scattered points interpolation), etc.

The MMSE result $\mathbf{F}(\mathbf{y}) = \mathbf{y} + \sigma^2 \nabla \log \mathscr{P}\{\mathbf{y}\}$ indicates that the optimal processing is *slowly varying*. It can thus, in principle, be represented on a basis of few functions — e.g., the identity and spline/Gaussian functions.



(see slide 12) $\quad a \times y \quad b \times \mathrm{sign}(y)\left(1 - \mathrm{e}^{-\frac{y^2}{2T^2}}\right) \quad c \times y\mathrm{e}^{-\frac{y^2}{2T^2}}$

Linear Expansion of Thresholds

An approximation of the optimal denoising process as a (finite) linear combination of elementary processes

$$\mathbf{F}(\mathbf{y}) = \sum_{k=1}^{K} a_k \mathbf{F}_k(\mathbf{y})$$

The approximation is all the better as the order, K, is larger.

The linear space approximation will prove particularly useful when combined with a quadratic objective functional (e.g., MSE or SURE), as the optimization boils down to solving a *linear system of equations*.

The idea of LET is that a genuine *approximation* of the optimal processing can be sufficient, while having useful *linear* properties.

Choosing the LET basis

Based on Wiener theory, *homogenous* (Gaussian, zero-mean) images are optimally denoised by *linear transformations*.

By *segmenting/partitioning* a non-homogenous image into homogenous zones, the "optimal" denoising process can thus be expressed as a sum of linear processes within each zone

$$\mathbf{F}(\mathbf{y}) = \sum_{\text{zones}} \overbrace{\gamma_k(\mathbf{y})}^{\text{indicator function of zone } k} \mathbf{A}_k \mathbf{y}$$

Hence, the choice of a LET basis essentially amounts to choosing a "good" (MSE-wise) segmentation algorithm.

Choosing the LET basis

Example: A simple threshold tends to segment a signal into large values, and small values. A possible choice[3] for the indicator function of the small values is

$$\gamma(y) = \mathrm{e}^{-\frac{y^2}{2T^2}}$$

Then, a possible LET function is of the form

$$F(y) = \underbrace{\gamma(y) \times ay}_{\text{small } y} + \underbrace{\big(1-\gamma(y)\big) \times by}_{\text{large } y}$$

The coefficients a and b characterize the linear behavior of the processing in each zone.

NOTE: A practical choice for T is $\sqrt{6}\,\sigma$ (noise), which can be related to a significance level in a statistical test.

[3] for a $\tanh$-based threshold, see Pesquet *et al.* 1997

Recapitulation of the SURE-LET approach

1. Instead of finding an *approximation of the signal* $\mathbf{x}$, find an *approximation of the processing* $\mathbf{F}(\mathbf{y})$ that transforms $\mathbf{y}$ into $\hat{\mathbf{x}}$;
2. Instead of minimizing the MSE between $\hat{\mathbf{x}}$ and $\mathbf{x}$, minimize an (unbiased) *estimate* of this MSE, based on $\mathbf{y}$ alone (SURE);
3. Express $\mathbf{F}(\mathbf{y})$ as a linear decomposition (LET) $\sum_k a_k \mathbf{F}_k(\mathbf{y})$ of basis processings $\mathbf{F}_k(\mathbf{y}) \rightsquigarrow$ linear system of equations (fast, unique).

NOTE: The number K of elementary processings is chosen very small (usually, $K < 200$), compared to the number of pixels N.
$\rightsquigarrow$ faster algorithm, and better agreement between MSE and SURE.

The SURE minimization

By restricting $\mathbf{F}(\mathbf{y})$ to be of the LET form $\sum_k a_k \mathbf{F}_k(\mathbf{y})$, the becomes a *quadratic* expression, in function of the a_k's. Its minimization yields, for all $k = 1, 2, \ldots, K$

$$\sum_{l=1}^{K} \mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{F}_l(\mathbf{y})\, a_l = \mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{y} - \sigma^2 \mathrm{div}\left\{\mathbf{F}_k(\mathbf{y})\right\}$$

Finally, by stacking the LET coefficients in $\mathbf{a} = [a_1, a_2, \ldots, a_K]^{\mathrm{T}}$, we get

$$\mathbf{a} = \mathbf{M}^{-1}\mathbf{c} \quad \text{where} \quad \left| \begin{array}{l} \mathbf{M} = \left[\mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{F}_l(\mathbf{y})\right]_{1 \le k,l \le K} \\ \mathbf{c} = \left[\mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{y} - \sigma^2 \mathrm{div}\left\{\mathbf{F}_k(\mathbf{y})\right\}\right]_{1 \le k \le K} \end{array} \right.$$

NOTE: When $\mathbf{M}$ is non-invertible, it means that one LET basis element depends linearly on the other $\mathbf{F}_k \rightsquigarrow$ decrease the LET-order to $K - 1$.

The Oracle minimization

The same LET optimization, by minimizing the MSE $\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\|^2$ instead of the SURE yields, for all $k = 1, 2, \ldots, K$

$$\sum_{l=1}^{K} \mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{F}_l(\mathbf{y})\, a_l = \mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{x}$$

This also boils down to solving a linear system of equations

$$\mathbf{a} = \mathbf{M}^{-1}\mathbf{c}' \quad \text{where} \quad \left| \begin{array}{l} \mathbf{M} = \left[\mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{F}_l(\mathbf{y})\right]_{1 \le k,l \le K} \\ \mathbf{c}' = \left[\mathbf{F}_k(\mathbf{y})^{\mathrm{T}} \mathbf{x}\right]_{1 \le k \le K} \end{array} \right.$$

NOTE: The Oracle computation allows to choose elementary LET processings $\mathbf{F}_k$ that are likely to yield more efficient denoising results.

A strategy for evaluating algorithms

How to evaluate the potential of an algorithm, that usually involves a number of non-linear parameters?

- Approximate the resulting algorithm as a LET; i.e., transfer the non-linear degrees of freedom to linear parameters;
- Probe the efficiency of the algorithm through Oracle minimization.

Example: If the algorithm $\mathbf{F}(\mathbf{y};\lambda)$ depends on *one* non-linear parameter, λ, approximate it using *two* (or more) LETs

$$\mathbf{F}(\mathbf{y};\lambda) = a_1\mathbf{F}(\mathbf{y};\lambda_1) + a_2\mathbf{F}(\mathbf{y};\lambda_2)$$

where λ_1, λ_2 are fixed: $\left[\lambda_1, \lambda_2\right]$ is the expected range of values for λ.

Monte-Carlo divergence estimation

The computation of the *divergence* term in the SURE may be impractical when N is large: a direct application of the formula

$$\operatorname{div}\{\mathbf{F}(\mathbf{y})\} = \sum_{n=1}^{N} \frac{\partial F_n(\mathbf{y})}{\partial y_n}$$

may prove too much CPU intensive.

An alternative is to use a consequence of Stein's Lemma

$$\operatorname{div}\{\mathbf{F}(\mathbf{y})\} \approx \mathbf{b}_0^{\mathrm{T}} \frac{\mathbf{F}(\mathbf{y}+\varepsilon\mathbf{b}_0) - \mathbf{F}(\mathbf{y})}{\varepsilon} \quad \text{(law of large numbers)}$$

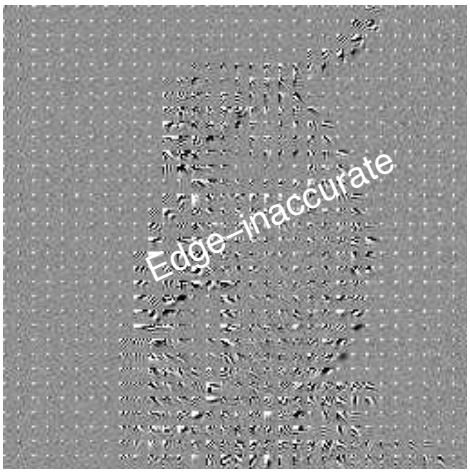
where $\mathbf{b}_0$ is a normalized (unit-variance, zero-mean) Gaussian white noise. ε is some small value compared to the level of noise; typ., $\varepsilon = \sigma/100$.

NOTE: Particularly useful when $\mathbf{F}(\mathbf{y})$ is not obtained explicitely, but through a "black-box" algorithm like TV regularization Ramani *et al.* 2008.

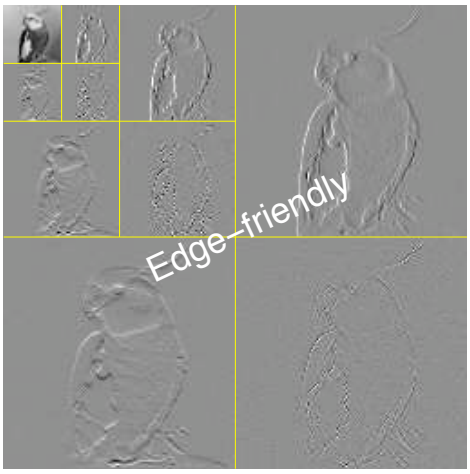
Linear transformations

In order to exploit their strong local correlations, it is advantageous to represent the pixels in another domain: *Discrete Cosine Transform* (DCT), *Block DCT*, *Wavelet Transform*, etc.

BDCT transformed image (block–size = 8)



Wavelet decomposition (3 iterations)



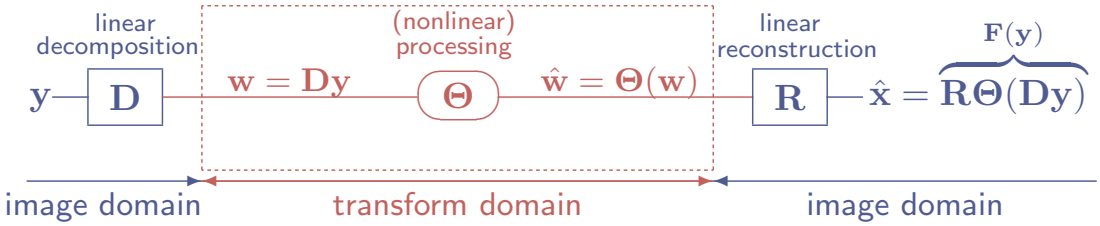
Most generally, a linear transformation maps an image $\mathbf{y}$ onto another image $\mathbf{w}$ through a matrix multiplication $\mathbf{Dy}$. It is assumed that the transformation can be inverted using a matrix $\mathbf{R}$.

Desirable properties (not all of them can be satisfied at once):

- Perfect reconstruction: $\mathbf{RD} = \mathbf{Id}$;
- $\mathbf{D}$ yields a sparse/decorrelated image representation;
- Shift, scale, rotation invariance;
- Orthonormality.

Example: *undecimated* wavelet transforms/BDCT are shift-invariant, but are not orthogonal.

Processing images expressed in a *sparse representation* considerably increases denoising efficiency.



Graphical overview: transform-domain thresholding

SURE-LET methodology: specify a LET basis $\mathbf{F}_k(\mathbf{y})$ as follows

$$\boldsymbol{\Theta}(\mathbf{w}) = \sum_{k=1}^{K} a_k \boldsymbol{\Theta}_k(\mathbf{w}) \;\rightsquigarrow\; \mathbf{F}_k(\mathbf{y}) = \mathbf{R}\boldsymbol{\Theta}_k(\mathbf{D}\mathbf{y})$$

Potential issue: efficient computation[4] of the SURE (essentially the $\operatorname{div}\{\mathbf{F}_k\}$ term) for this type of processing $\rightsquigarrow$ Monte-Carlo technique.

[4]However, *exact* expression in a number of practical cases (periodic extensions).

Orthonormality

A decomposition is orthonormal iff $\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{Id}$. Properties:

- The reconstruction is given by $\mathbf{R} = \mathbf{D}^{\mathrm{T}}$;
- Preservation of the energies: $\|\mathbf{w}\| = \|\mathbf{y}\|$ and $\|\hat{\mathbf{x}} - \mathbf{x}\| = \|\hat{\mathbf{w}} - \mathbf{D}\mathbf{x}\|$;
- Statistical *independence* of the transformed coefficients;

NOTE: an orthonormal decomposition is automatically *non-redundant*.

If $\mathbf{w}_j = \mathbf{D}_j\mathbf{y}$ for $j = 1, 2, \ldots, J$ where $\mathbf{D} = [\mathbf{D}_1; \mathbf{D}_2; \ldots; \mathbf{D}_J]$, then the unbiased estimate of $\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ can be written in the *transformed domain*

$$\mathrm{SURE}(\mathbf{y}) = \frac{1}{N}\left(\sum_{j=1}^{J} \|\boldsymbol{\Theta}_j(\mathbf{w}_j) - \mathbf{w}_j\|^2 + 2\sigma^2 \operatorname{div}\{\boldsymbol{\Theta}_j(\mathbf{w}_j)\}\right) - \sigma^2$$

where $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1; \boldsymbol{\Theta}_2; \ldots; \boldsymbol{\Theta}_J]$.

Optimizing the denoising process $\mathbf{F}(\mathbf{y})$ is equivalent to denoising *separately* the denoising processes $\boldsymbol{\Theta}_j$ in the transformed domain.

Simple wavelet thresholding

Choice of an orthonormal wavelet transform[5] (e.g., symlet 8). Then, the processing in subband j is a simple thresholding $\hat{w}_{j,n} = \theta_j(w_{j,n})$ for each of the coordinates $n = 1, 2, \ldots, N_j$ of $\mathbf{w}_j$, and

$$\text{SURE}_j(\mathbf{w}_j) = \frac{1}{N_j}\left(\sum_{n=1}^{N_j} \left|\theta_j(w_{j,n}) - w_{j,n}\right|^2 + 2\sigma^2\theta_j'(w_{j,n})\right) - \sigma^2$$

SURE-LET simple threshold

A two-parameter zone-selection function

$$\theta_j(w) = a_j w + b_j w \mathrm{e}^{-\frac{w^2}{12\sigma^2}}$$

where a_j and b_j are obtained by minimizing $\text{SURE}_j(\mathbf{w}_j)$.

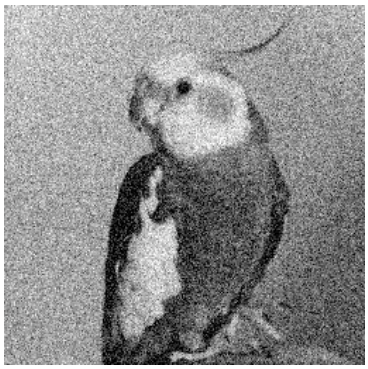
NOTE: SureShrink Donoho 1995 makes the choice $\theta_j(w) = \text{soft}_{T_j}(w)$ and minimizes $\text{SURE}_j(\mathbf{w}_j)$ for T_j.

[5] However, any (non-wavelet) *orthonormal* transform can be used.

In details, a_j, b_j solve the following linear system of equations

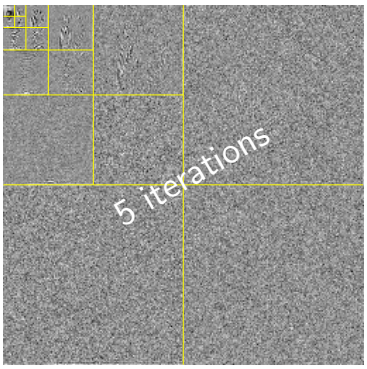
$$\frac{\partial \text{SURE}_j}{\partial a_j} = 0 \rightsquigarrow \sum_{n=1}^{N_j} a_j w_{j,n}^2 + b_j w_{j,n}^2 \mathrm{e}^{-\frac{w_{j,n}^2}{12\sigma^2}} = -N_j\sigma^2 + \sum_{n=1}^{N_j} w_{j,n}^2$$

$$\frac{\partial \text{SURE}_j}{\partial b_j} = 0 \rightsquigarrow \sum_{n=1}^{N_j} a_j w_{j,n}^2 \mathrm{e}^{-\frac{w_{j,n}^2}{12\sigma^2}} + b_j w_{j,n}^2 \mathrm{e}^{-\frac{w_{j,n}^2}{6\sigma^2}} = \sum_{n=1}^{N_j} \left(\frac{7}{6} w_{j,n}^2 - \sigma^2\right) \mathrm{e}^{-\frac{w_{j,n}^2}{12\sigma^2}}$$

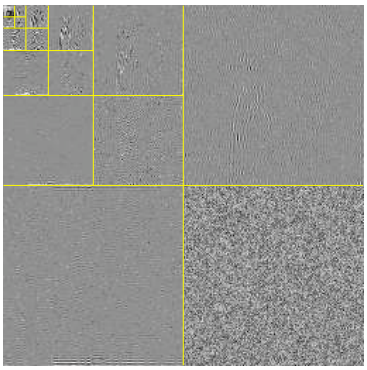


Noisy: PSNR $= 18$ dB

wavelet decomposition →



↓ simple thresholding

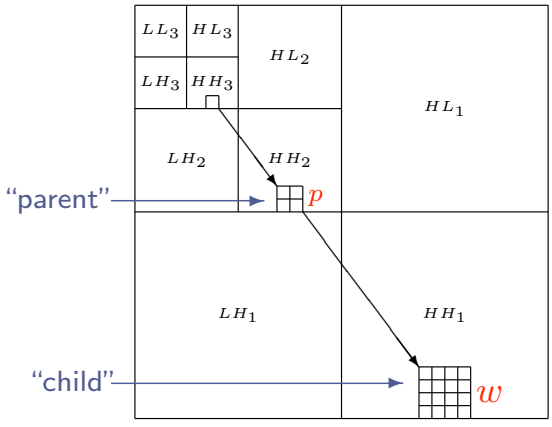


← wavelet reconstruction

Denoised: PSNR $= 29.06$ dB (SureShrink: PSNR $= 28.73$ dB)

InterScale wavelet thresholding

The relative locality of the DWT implies that there may be a *spatial correlation* between different wavelet scales: three potential *tree*-structures — LH, HH and HL



Interscale thresholding consists in expressing the denoised estimate as

$$\hat{w}_{j,n} = \theta_j(w_{j,n}, p_{j,n})$$

InterScale wavelet thresholding

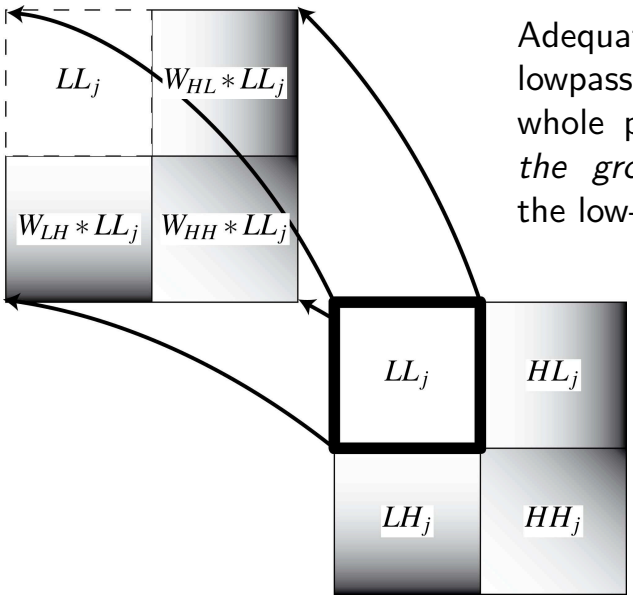
Principle: separate the parent into *large* and *small* coefficients, and within each zone so defined, apply a pointwise thresholding function:

$$\theta_j(w,p) = \underbrace{\mathrm{e}^{-\frac{p^2}{12\sigma^2}}\left(a_j w + b_j w \mathrm{e}^{-\frac{w^2}{12\sigma^2}}\right)}_{\text{small parents}} + \underbrace{(1-\mathrm{e}^{-\frac{p^2}{12\sigma^2}})\left(a'_j w + b'_j w \mathrm{e}^{-\frac{w^2}{12\sigma^2}}\right)}_{\text{large parents}}$$

NOTE: DWT is orthogonal, hence w and p are *statistically independent* $\rightsquigarrow$ same SURE formula as for the simple threshold case.

PROBLEM: the wavelet coefficients are not exactly aligned from band to band (filtering and downsampling effect). How to obtain a parent aligned exactly with its child?

Parent/child alignment: Group-Delay Compensation



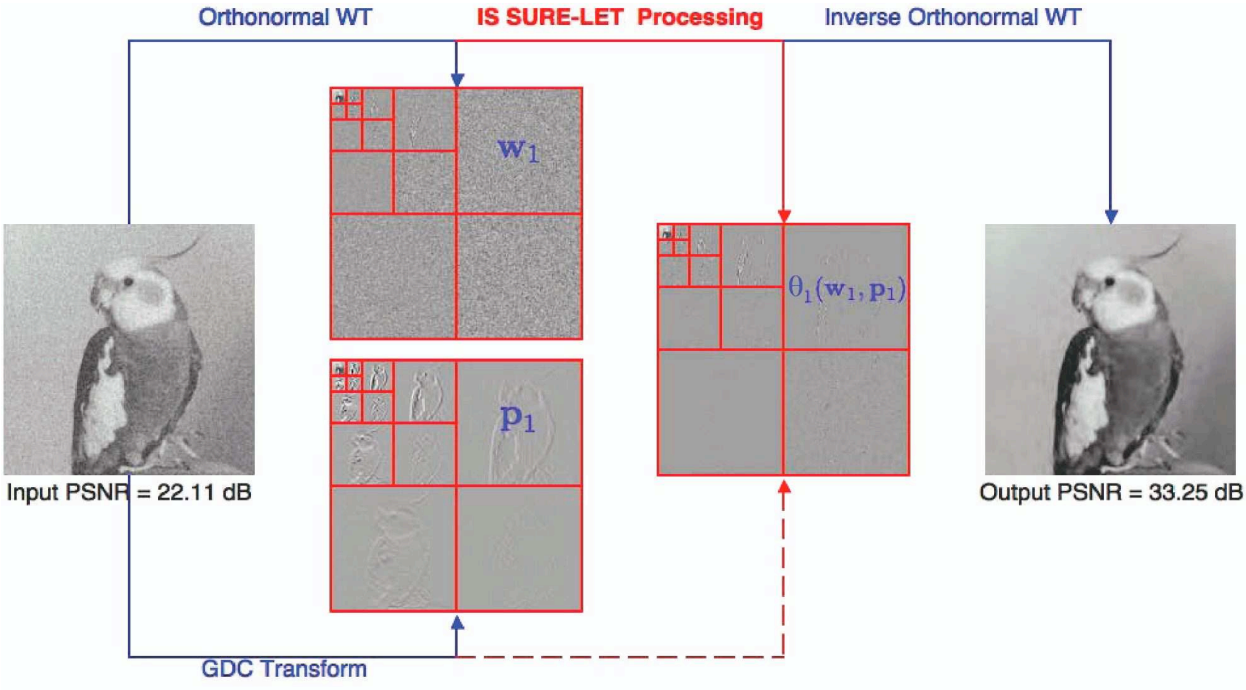
Adequate high-pass filtering of the lowpass LL_j — which contains the whole parent tree: W *compensates the group-delay* difference between the low-pass and the high-pass band.

GDC filter formula

$W(z^2) = (1+z^{-2})G(z^{-1})G(-z^{-1})$

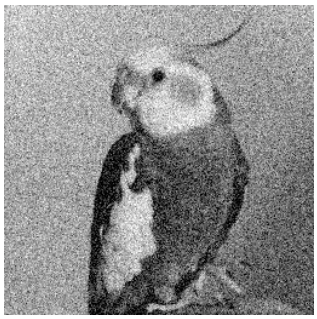
where $G(z)$ = wavelet filter.

Overview of the interscale SURE-LET denoising



Example of result

Noisy



PSNR=18 dB

SureShrink

PSNR=28.73 dB

SURE-LET interscale

PSNR=30.18 dB

Best non-redundant transform-domain algorithm.

Extension to multichannel denoising

Direct generalization by replacing:

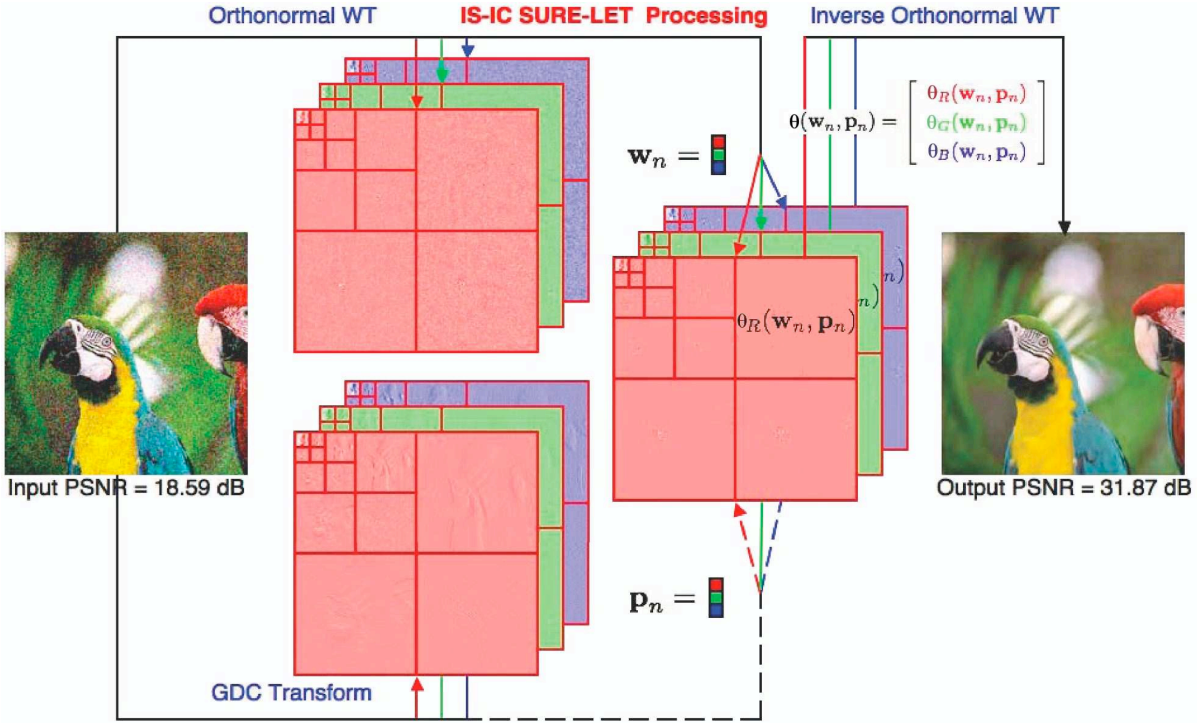
- *scalar*-valued by *vector*-valued wavelet coefficients;
- *scalar*-valued by *matrix*-valued LET parameters.

Assuming $\mathbf{Q}$ = covariance matrix of the noise, and $\gamma(x) = \exp(-x/12)$

$$\boldsymbol{\theta}_j(\mathbf{w},\mathbf{p}) = \underbrace{\gamma(\mathbf{p}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{p})\gamma(\mathbf{w}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{w})}_{\text{small parents and small coefficients}}\mathbf{a}_{1,j}^{\mathrm{T}}\mathbf{w} + \underbrace{\big(1-\gamma(\mathbf{p}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{p})\big)\gamma(\mathbf{w}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{w})}_{\text{large parents and small coefficients}}\mathbf{a}_{2,j}^{\mathrm{T}}\mathbf{w} + \underbrace{\gamma(\mathbf{p}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{p})\big(1-\gamma(\mathbf{w}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{w})\big)}_{\text{small parents and large coefficients}}\mathbf{a}_{3,j}^{\mathrm{T}}\mathbf{w} + \underbrace{\big(1-\gamma(\mathbf{p}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{p})\big)\big(1-\gamma(\mathbf{w}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{w})\big)}_{\text{large parents and large coefficients}}\mathbf{a}_{4,j}^{\mathrm{T}}\mathbf{w}$$

NOTE: Automatically selects the best color space (color images).

Overview of the Multichannel SURE-LET denoising



Undecimated wavelet denoising

Limitations of non-redundant transformations

- High sensitivity to shifts $\rightsquigarrow$ inconsistent reconstruction of edges
- Low design flexibility $\rightsquigarrow$ poor directional sensitivity

Solution: increase the redundancy

Shifts: Cycle-Spinning Coifman 1995, Undecimated DWT Guo 1995;
Rotations: Steerable Pyramid Simoncelli 1995, Complex DWT Kingsbury 1998;
Edges: Curvelets Candès 2002; etc...

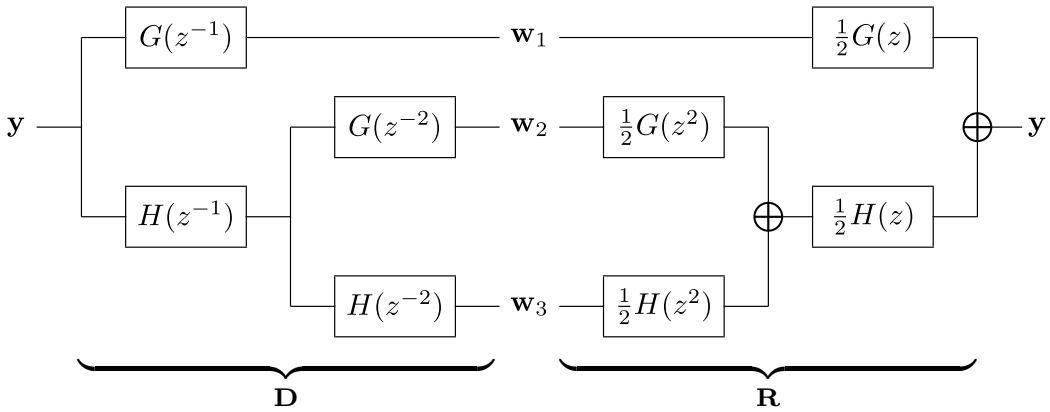
Redundancy vs orthonormality

Although it is still possible to have $\mathbf{R} = \mathbf{D}^{\mathrm{T}}$ (*tight frame*)

- $\mathbf{RD} = \mathbf{Id}$ but $\mathbf{DR} \neq \mathbf{Id}$
- Energies: $\|\mathbf{w}\| = \|\mathbf{y}\|$ (if tight frame) but $\|\hat{\mathbf{x}} - \mathbf{x}\| \neq \|\hat{\mathbf{w}} - \mathbf{Dx}\|$;
- Statistical *dependence* of the transformed coefficients;

In addition, redundancy brings about a higher computational cost.

Two iterations of a 1D UDWT



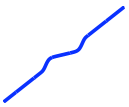
Perfect reconstruction condition: $\mathbf{RD} = \mathbf{Id}$

NOTE: same lowpass and highpass filters, $H(z)$ and $G(z)$, as in the non-redundant WT case.

Undecimated simple wavelet thresholding

Hard-like[6] thresholding rule

In each wavelet subband j, the noisy coefficients are thresholded using



$$\theta_j(w) = a_j w + b_j w\left(1 - \mathrm{e}^{-\left(\frac{w}{3\sigma}\right)^8}\right)$$

where (a_j, b_j) change from subband to subband — i.e., two parameters per subband.

The optimal set of parameters $\{a_j, b_j\}$ is then found by minimizing the global image-domain SURE.

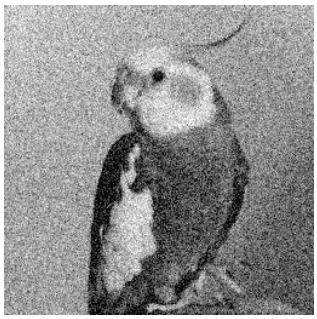
NOTE: Contrary to the nonredundant case, it is not possible to optimize the SURE separately in each subband.

[6]Hard threshold cannot be optimized using SURE, for not being differentiable.

Undecimated pointwise wavelet thresholding

Undecimated discrete symlet 8 transform

Noisy



PSNR=18 dB

SureShrink

PSNR=28.73 dB

SURE-LET

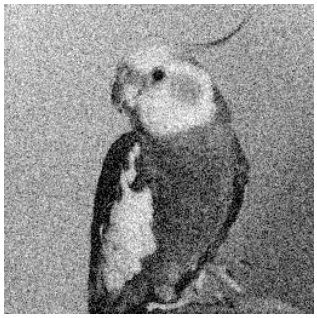
PSNR=31.15 dB

NOTE: Surprisingly, it is the simplest wavelet type (Haar) that works best. Smallest support?

Undecimated pointwise wavelet thresholding

Undecimated discrete Haar wavelet transform

Noisy



PSNR=18 dB

SureShrink

PSNR=28.73 dB

SURE-LET

PSNR=31.91 dB

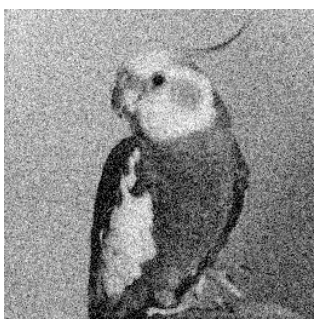
NOTE: Surprisingly, it is the simplest wavelet type (Haar) that works best. Shortest support?

Extensions

- **Multivariate** wavelet thresholding: taking into account both *interscale* and *local* wavelet dependencies;
- Thresholding (possibly multivariate) in a **dictionary** of transforms.
- **Multiframe** video denoising: involving motion compensation;

Orthonormal discrete symlet 8 transform

Noisy



PSNR=18 dB

SURE-LET interscale

PSNR=30.18 dB

SURE-LET multivariate

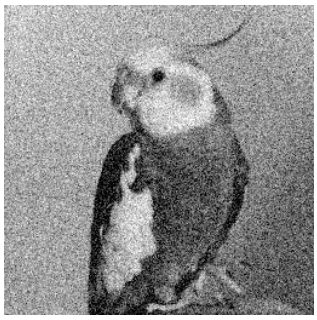
PSNR=30.65 dB

Extensions

- **Multivariate** wavelet thresholding: taking into account both *interscale* and *local* wavelet dependencies;
- Thresholding (possibly multivariate) in a **dictionary** of transforms.
- **Multiframe** video denoising: involving motion compensation;

Undecimated discrete Haar wavelet transform

Noisy



PSNR=18 dB

SURE-LET

PSNR=31.91 dB

SURE-LET multivariate

PSNR=32.22 dB

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Transform domain denoising
Orthogonal Representations/Transformations
Redundant Representations/Transformations
Noise Variance Estimation

Extensions

- **Multivariate** wavelet thresholding: taking into account both *interscale* and *local* wavelet dependencies;
- Thresholding (possibly multivariate) in a **dictionary** of transforms.
- **Multiframe** video denoising: involving motion compensation;

Dictionary of two transforms (UWT Haar & 12×12-BDCT)

Noisy

PSNR=18 dB

SURE-LET UWT Haar

PSNR=25.90 dB

SURE-LET Multivariate Dictionary

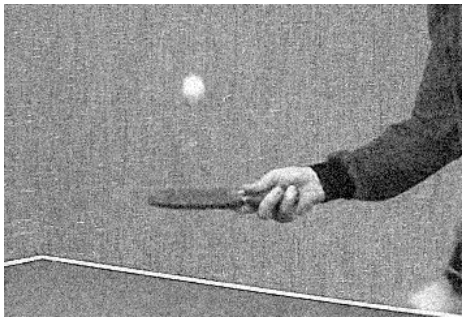
PSNR=28.80 dB

Extensions

- **Multivariate** wavelet thresholding: taking into account both *interscale* and *local* wavelet dependencies;
- Thresholding (possibly multivariate) in a **dictionary** of transforms.
- **Multiframe** video denoising: involving motion compensation;

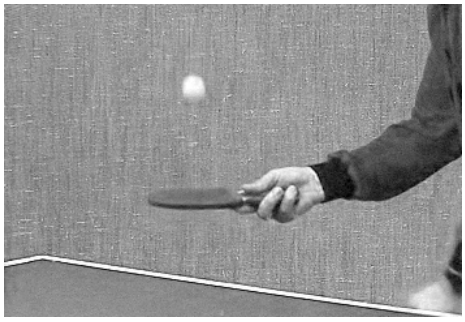
Orthonormal discrete symlet 8 transform

Noisy



PSNR=22.11 dB

Multiframe SURE-LET



PSNR=30.85 dB

Noise Variance Estimation

The most popular approach for estimating the variance σ^2 of the AWGN for wavelet-based denoising algorithms: **MAD estimator** Donoho 1995

$$\hat{\sigma} = 1.4826 \operatorname{med}\left\{|\mathbf{y} - \operatorname{med}\{\mathbf{y}\}|\right\}, \; y_n \in HH$$

+ Simple and accurate for relatively high levels of noise;
− Inaccurate for moderate to low levels of noise.

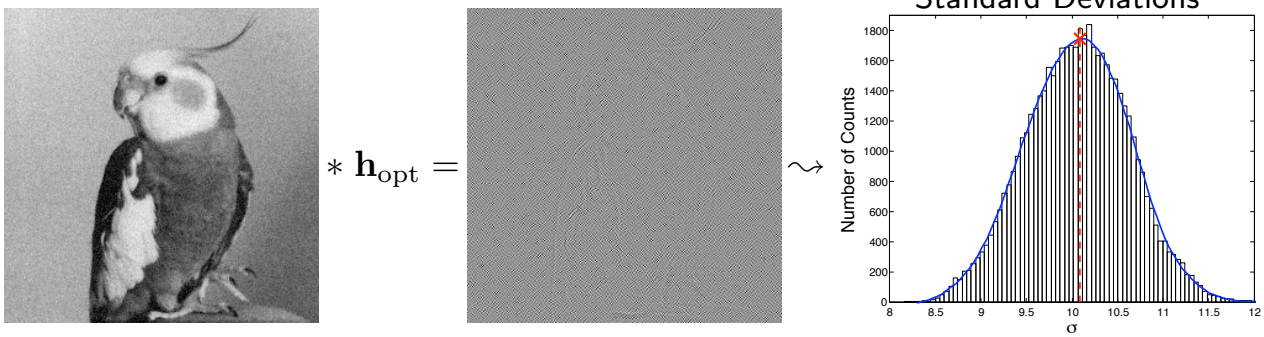
Proposed approach: **Eigenfilter-based design** Vaidyanathan *et al.* 1987

1. Find $\mathbf{h}_{\text{opt}} = \arg\min_{\mathbf{h}\in\mathbb{R}^M} \|\mathbf{h} * \mathbf{y}\|^2$ subject to $\|\mathbf{h}\|^2 = 1$
   $\rightsquigarrow$ Eigenvector corresponding to the smallest eigenvalue of the autocorrelation matrix $\mathbf{\Gamma}_{\mathbf{y}} = \left[\sum_{n=1}^{N} y_{n-i} y_{n-j}\right]_{1\le i,j\le M}$
2. Noise variance robustly estimated from the filtered residual $(\mathbf{h}_{\text{opt}} * \mathbf{y})$, as the mode of the smoothed histogram of the local noise variances computed inside blocks of given size (typically, 25×25).

Noise Variance Estimation

Overview of the Proposed Approach

Noisy Input: $\sigma = 10$ | Residual | Distribution of the Local Standard Deviations

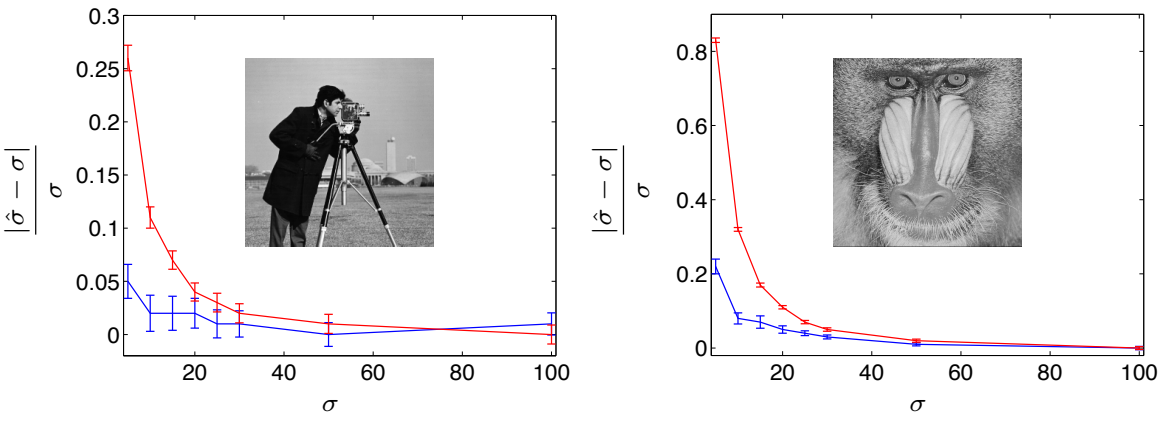


Estimated σ: $\hat{\sigma} = 10.09$

Noise Variance Estimation

Performance of the Proposed Approach

Cameraman | **Mandrill**



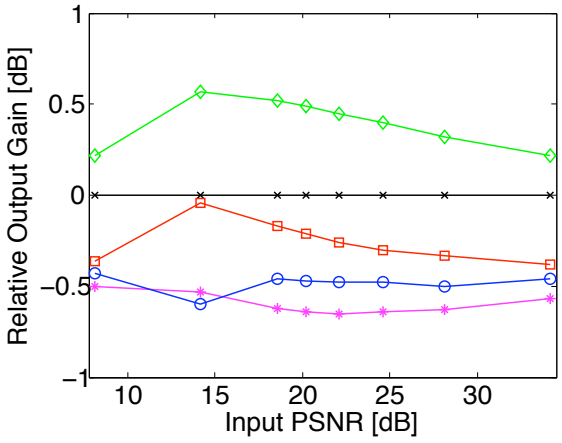
Proposed Approach MAD Estimator

Protocol for Fair Comparisons

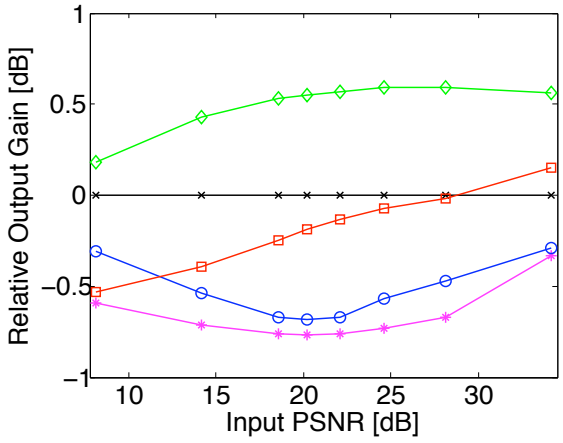
- Denoising of a representative set of standard grayscale/color images and video sequences, corrupted by simulated AWGN at 8 different powers $\sigma \in [5, 10, 15, 20, 25, 30, 50, 100]$ (assumed to be known).
- PSNR results averaged over 10 different noise realizations for each noise standard deviation.
- Parameters of each method set according to the values given in the corresponding referred papers or optimized in the MMSE sense (if not explicitly provided).

The Non-Redundant Case: PSNR Comparisons

Peppers 256×256 



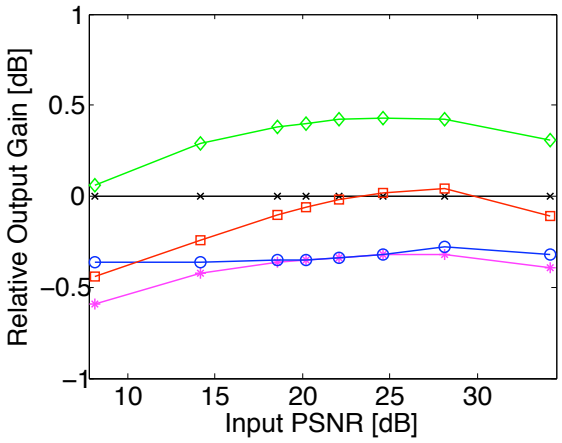
House 256×256 



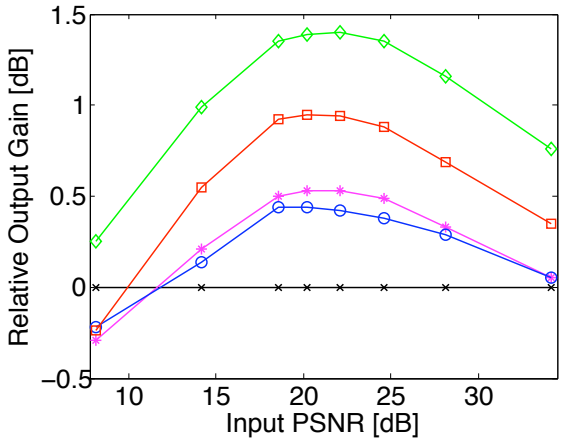
Interscale SURE-LET (baseline)
Multivariate SURE-LET
BiShrink Sendur & Selesnick 2002
ProbShrink Pižurica *et al.* 2006
BLS-GSM Portilla *et al.* 2003

The Non-Redundant Case: PSNR Comparisons

Lena 512×512 



Barbara 512×512 



Interscale SURE-LET (baseline)
Multivariate SURE-LET

BiShrink Sendur & Selesnick 2002
ProbShrink Pižurica *et al.* 2006
BLS-GSM Portilla *et al.* 2003

The Non-Redundant Case: Visual Comparisons

Original

Average SSIM[1]: 1.000

Noisy

Average SSIM: 0.284

[1] Structural Similarity Index Map Wang, Bovik, Sheikh & Simoncelli 2004

The Non-Redundant Case: Visual Comparisons

Original

Average SSIM: 1.000

Multivariate SURE-LET

Average SSIM: 0.894

[1] Structural Similarity Index Map Wang, Bovik, Sheikh & Simoncelli 2004

The Non-Redundant Case: Visual Comparisons

BiShrink

Average SSIM: 0.877

Multivariate SURE-LET

Average SSIM: 0.894

[1] Structural Similarity Index Map Wang, Bovik, Sheikh & Simoncelli 2004

The Non-Redundant Case: Visual Comparisons

ProbShrink

Average SSIM: 0.882

Multivariate SURE-LET

Average SSIM: 0.894

[1] Structural Similarity Index Map Wang, Bovik, Sheikh & Simoncelli 2004

The Non-Redundant Case: Visual Comparisons

BLS-GSM

Average SSIM: 0.888

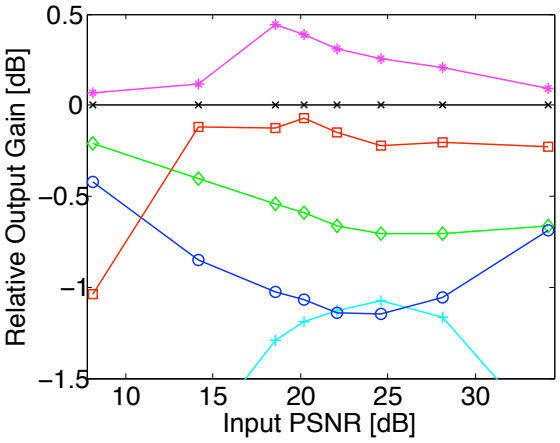
Multivariate SURE-LET

Average SSIM: 0.894

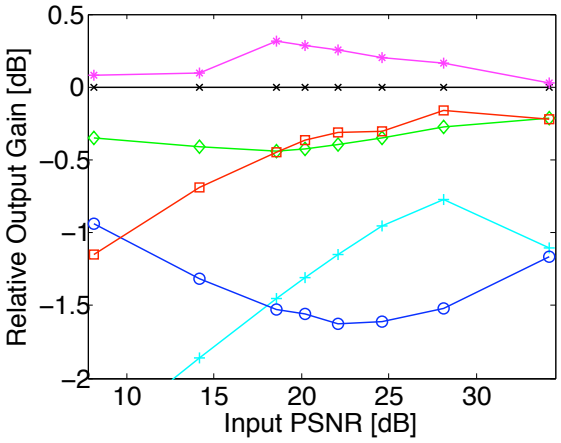
[1] Structural Similarity Index Map Wang, Bovik, Sheikh & Simoncelli 2004

The Redundant Case: PSNR Comparisons

Peppers 256×256 



Coco 256×256 

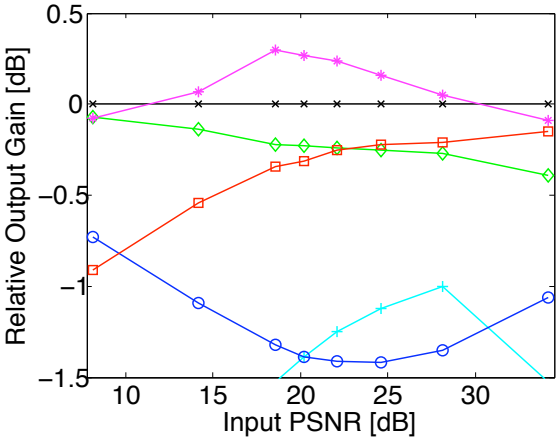


Multivariate SURE-LET (baseline)
BM3D Dabov *et al.* 2007
NLmeans Buades *et al.* 2005
Fast TV Chambolle 2004
BLS-GSM Portilla *et al.* 2003
K-SVD Elad & Aharon 2006

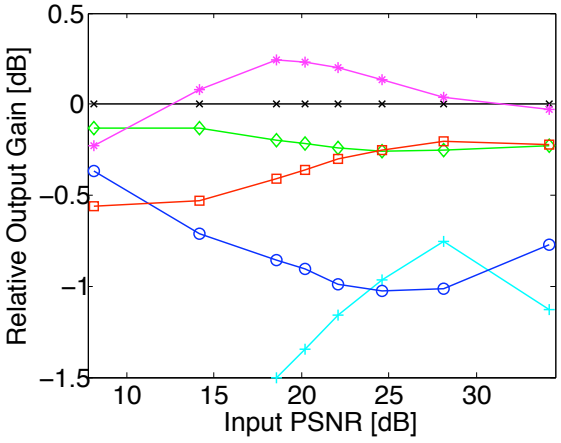
Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithms
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

The Redundant Case: PSNR Comparisons

Boat 512×512 



Goldhill 512×512 



Multivariate SURE-LET (baseline)
BM3D Dabov *et al.* 2007
NLmeans Buades *et al.* 2005
Fast TV Chambolle 2004
BLS-GSM Portilla *et al.* 2003
K-SVD Elad & Aharon 2006

The Redundant Case: Visual Comparisons

Original

Average SSIM: 1.000

Noisy

Average SSIM: 0.263

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithms
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

The Redundant Case: Visual Comparisons

Original

Average SSIM: 1.000

Multivariate SURE-LET

Average SSIM: 0.739

The Redundant Case: Visual Comparisons

NLmeans

Average SSIM: 0.662

Multivariate SURE-LET

Average SSIM: 0.739

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

The Redundant Case: Visual Comparisons

Fast TV

Average SSIM: 0.704

Multivariate SURE-LET

Average SSIM: 0.739

The Redundant Case: Visual Comparisons

BLS-GSM

Average SSIM: 0.732

Multivariate SURE-LET

Average SSIM: 0.739

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

The Redundant Case: Visual Comparisons

K-SVD

Average SSIM: 0.711

Multivariate SURE-LET

Average SSIM: 0.739

The Redundant Case: Visual Comparisons

BM3D

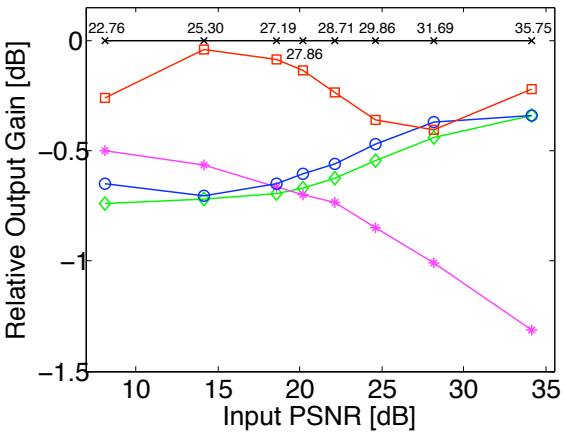
Average SSIM: 0.754

Multivariate SURE-LET

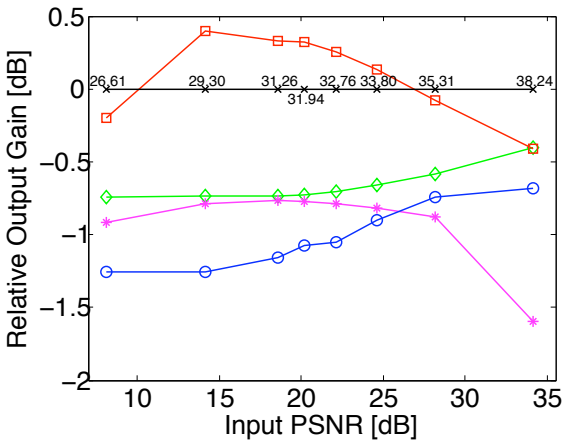
Average SSIM: 0.739

Color Images: PSNR Comparisons

Girl 256×256 



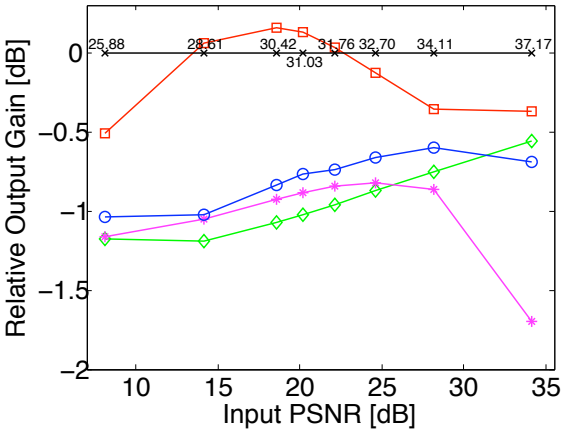
Lena 512×512 



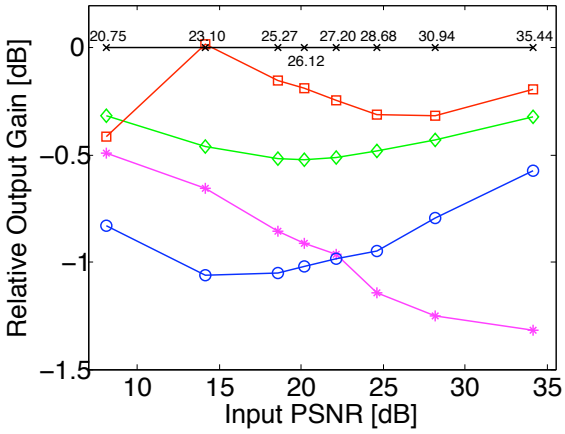
Multichannel SURE-LET (baseline)
Non-redundant multichannel SURE-LET
ProbShrink-YUV Pižurica *et al.* 2005
ProbShrink-MB Pižurica *et al.* 2006
CBM3D Dabov *et al.* 2007

Color Images: PSNR Comparisons

Peppers 512×512 



Mandrill 512×512 



Multichannel SURE-LET (baseline)
Non-redundant multichannel SURE-LET
ProbShrink-YUV Pižurica *et al.* 2005
ProbShrink-MB Pižurica *et al.* 2006
CBM3D Dabov *et al.* 2007

Color Images: Visual Comparisons

Original

Average SSIM: 1.000

Noisy

Average SSIM: 0.221

Color Images: Visual Comparisons

Original

Average SSIM: 1.000

Multichannel SURE-LET

Average SSIM: 0.872

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

Color Images: Visual Comparisons

ProbShrink-MB

Average SSIM: 0.825

Multichannel SURE-LET

Average SSIM: 0.872

Color Images: Visual Comparisons

ProbShrink-YUV

Average SSIM: 0.841

Multichannel SURE-LET

Average SSIM: 0.872

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Grayscale Image Denoising
Color Image Denoising
Video Denoising

Color Images: Visual Comparisons

CBM3D

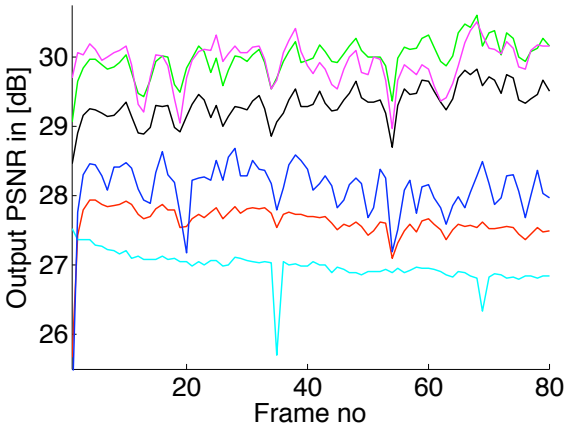
Average SSIM: 0.882

Multichannel SURE-LET

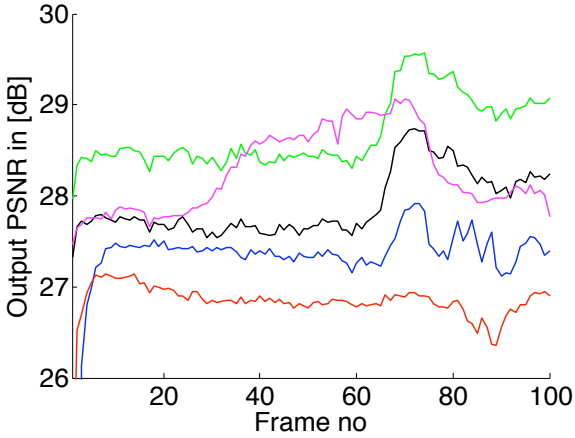
Average SSIM: 0.872

Frame-by-Frame PSNR Comparisons

Flowers at PSNR $= 24.61$ dB



Bus at PSNR $= 20.17$ dB



Multiframe SURE-LET (OWT)
Multiframe SURE-LET ($\mathrm{CS} = 5$)
SEQWT Pižurica *et al.* 2004
WRSTF Zlokolica *et al.* 2006
Real-time WRSTF Jovanov *et al.* 2009
VBM3D Dabov *et al.* 2007

Visual Comparison

Noisy Input

PSNR $= 20.17$ dB

Multiframe SURE-LET ($\mathrm{CS} = 5$)

PSNR $= 31.62$ dB

More Realistic Measurement Model

Most light intensity measurements $\mathbf{y} = [y_1 \dots y_N]^{\mathrm{T}}$ are more accurately modeled as a vector $\mathbf{z}$ of **independent Poisson random variables** degraded by **independent AWGN** $\mathbf{b}$:

$$\mathbf{y} = \mathbf{z} + \mathbf{b}, \text{ where } \mathbf{z} \sim \mathcal{P}(\mathbf{x}) \text{ and } \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Id})$$

This model accounts for:

- Random nature of photon emission/detection
  $\rightsquigarrow$ **signal-dependent** degradation;
- Thermal instabilities of the electronic devices
  $\rightsquigarrow$ **signal-independent** noise.

Only few denoising algorithms consider this hybrid measurement model.

Two Main Approaches for Poisson Intensity Estimation

- **Variance-stabilizing transform (VST):**
  Design a transform T such that $\mathrm{T}(y) - \mathrm{T}(x) \underset{x \to +\infty}{\xrightarrow{\mathcal{D}}} \mathcal{N}(0,1)$
  - Anscombe and its extension to Poisson-Gaussian noise
    Murtagh *et al.* 1995;
  - Haar-Fisz Fryzlewicz & Nason 2004;
  - Multiscale VST Jansen 2006, Fadili *et al.* 2008.
- **Direct handling of Poisson statistics:**
  Almost exclusively in a Bayesian framework
  - Multiscale Bayesian model Nowak *et al.* 1999;
  - Hypothesis testing Kolaczyk 1999, Fadili *et al.* 2007;
  - Penalized likelihood Sardy *et al.* 2004, Willett & Nowak 2007.

Potential of purely data-driven, prior-free
MMSE techniques remains under-exploited.

PURE: Poisson-Gaussian Unbiased Risk Estimate

Let $\mathbf{y} = \mathbf{z} + \mathbf{b}$ with $\mathbf{z} \sim \mathcal{P}(\mathbf{x})$ independent of $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Id})$. Let $\mathbf{f}(\mathbf{y}) = [f_n(\mathbf{y})]_{1 \leq n \leq N}$ such that $\mathscr{E}\{|\partial f_n(\mathbf{y})/\partial y_n|\} < +\infty$. Then,

$$\text{PURE} = \frac{1}{N}\left(\|\mathbf{f}(\mathbf{y})\|^2 - 2\mathbf{y}^{\mathrm{T}}\mathbf{f}^{-}(\mathbf{y}) + 2\sigma^2 \mathrm{div}\left\{\mathbf{f}^{-}(\mathbf{y})\right\}\right) + \frac{1}{N}\left(\|\mathbf{y}\|^2 - \mathbf{1}^{\mathrm{T}}\mathbf{y}\right) - \sigma^2$$

is an **unbiased estimate of the expected MSE**; i.e.,

$$\mathscr{E}\{\text{PURE}\} = \frac{1}{N}\mathscr{E}\left\{\|\mathbf{f}(\mathbf{y}) - \mathbf{x}\|^2\right\}$$

Notation: $\mathbf{f}^{-}(\mathbf{y}) = [f_n(\mathbf{y} - \mathbf{e}_n)]_{1 \leq n \leq N}$, where $(\mathbf{e}_n)_{1 \leq n \leq N}$ is the canonical basis of $\mathbb{R}^N$.

PURE: Poisson-Gaussian Unbiased Risk Estimate

Sketch of proof: Need to estimate

$$\mathscr{E}\left\{\|\mathbf{f}(\mathbf{y}) - \mathbf{x}\|^2\right\} = \sum_n \left(\mathscr{E}\left\{f_n^2(\mathbf{y})\right\} - 2\underbrace{\mathscr{E}\left\{x_n f_n(\mathbf{y})\right\}}_{1} + \underbrace{x_n^2}_{2}\right)$$

1. Poisson's Lemma Hudson 1978, Tsui & Press 1982:

$$\begin{aligned}\mathscr{E}\left\{x_n f_n(\mathbf{y})\right\} &= \mathscr{E}\left\{x_n f_n(\mathbf{z} + \mathbf{b})\right\} \\ &= \mathscr{E}\left\{z_n f_n(\mathbf{z} + \mathbf{b} - \mathbf{e}_n)\right\}\end{aligned}$$

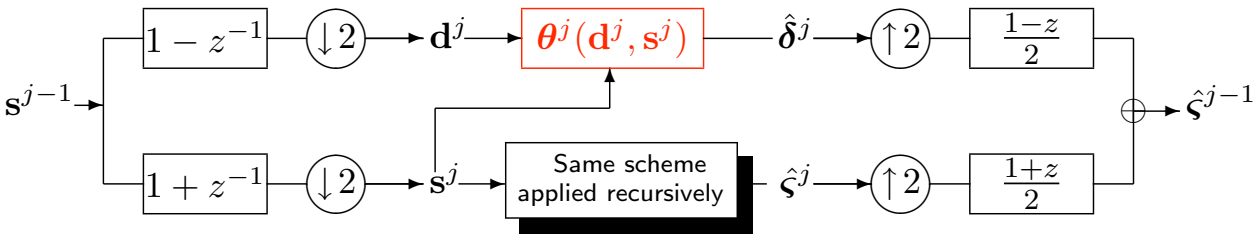
Stein's Lemma Stein 1981:

$$\begin{aligned}\mathscr{E}\left\{z_n f_n(\mathbf{z} + \mathbf{b} - \mathbf{e}_n)\right\} &= \mathscr{E}\left\{y_n f_n(\mathbf{y} - \mathbf{e}_n)\right\} - \mathscr{E}\left\{b_n f_n(\mathbf{z} + \mathbf{b} - \mathbf{e}_n)\right\} \\ &= \mathscr{E}\left\{y_n f_n(\mathbf{y} - \mathbf{e}_n)\right\} - \sigma^2\mathscr{E}\left\{\partial f_n(\mathbf{y} - \mathbf{e}_n)/\partial y_n\right\}\end{aligned}$$

2. Notice that: $x_n^2 = \mathscr{E}\left\{x_n y_n\right\} \overset{1}{=} \mathscr{E}\left\{y_n(y_n - 1)\right\} - \sigma^2$

The Unnormalized Haar Wavelet Transform

Denoising by **interscale** thresholding of the unnormalized **Haar-wavelet** coefficients: set $\mathbf{s}_0 = \mathbf{y}$, then for $j = 1, 2, \ldots, J$



Haar conservation properties:

- **Error energy:** $\text{MSE} = \frac{2^{-J}}{N}\|\hat{\boldsymbol{\varsigma}}^J - \boldsymbol{\varsigma}^J\|^2 + \sum_{j=1}^{J} \frac{2^{-j}}{N}\|\hat{\boldsymbol{\delta}}^j - \boldsymbol{\delta}^j\|^2$
- **Statistics:** $\mathbf{s}^j \sim \mathcal{P}(\boldsymbol{\varsigma}^j) + \mathcal{N}(\mathbf{0}, \sigma_j^2 \mathbf{Id})$, where $\sigma_j^2 = 2^j\sigma^2$

Allows independent processing of each wavelet subband.

Interscale Haar-Wavelet-Domain PURE

Let $\boldsymbol{\theta}(\mathbf{d}, \mathbf{s}) = \boldsymbol{\theta}^j(\mathbf{d}^j, \mathbf{s}^j)$ be an estimate of the noise-free wavelet coefficients $\boldsymbol{\delta} = \boldsymbol{\delta}^j$. Define $\boldsymbol{\theta}^+(\mathbf{d}, \mathbf{s})$ and $\boldsymbol{\theta}^-(\mathbf{d}, \mathbf{s})$ by

$$\begin{cases} \theta_n^+(\mathbf{d}, \mathbf{s}) = \theta_n(\mathbf{d} + \mathbf{e}_n, \mathbf{s} - \mathbf{e}_n) \\ \theta_n^-(\mathbf{d}, \mathbf{s}) = \theta_n(\mathbf{d} - \mathbf{e}_n, \mathbf{s} - \mathbf{e}_n) \end{cases}$$

Then the random variable[1]

$$\begin{aligned} \text{PURE}_j = \frac{1}{N_j}\Big(& \|\boldsymbol{\theta}(\mathbf{d}, \mathbf{s})\|^2 + \|\mathbf{d}\|^2 - \mathbf{1}^{\mathrm{T}}\mathbf{s} - N_j\sigma_j^2 \\ & - \mathbf{d}^{\mathrm{T}}\big(\boldsymbol{\theta}^-(\mathbf{d}, \mathbf{s}) + \boldsymbol{\theta}^+(\mathbf{d}, \mathbf{s})\big) - \mathbf{s}^{\mathrm{T}}\big(\boldsymbol{\theta}^-(\mathbf{d}, \mathbf{s}) - \boldsymbol{\theta}^+(\mathbf{d}, \mathbf{s})\big) \\ & + \sigma_j^2\left(\text{div}_{\mathbf{d}}\left\{\boldsymbol{\theta}^-(\mathbf{d}, \mathbf{s}) + \boldsymbol{\theta}^+(\mathbf{d}, \mathbf{s})\right\} + \text{div}_{\mathbf{s}}\left\{\boldsymbol{\theta}^-(\mathbf{d}, \mathbf{s}) - \boldsymbol{\theta}^+(\mathbf{d}, \mathbf{s})\right\}\right)\Big) \end{aligned}$$

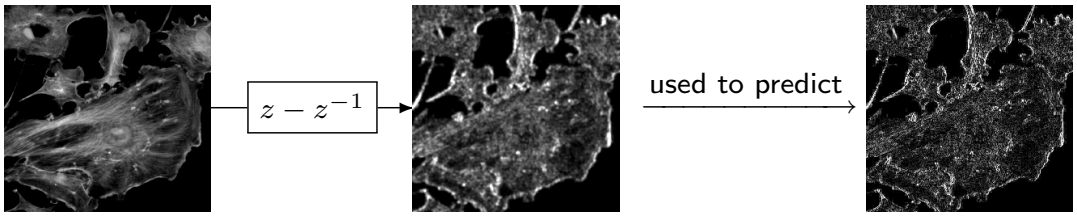
is an **unbiased estimate of the expected MSE** for the jth subband; i.e.,

$$\mathscr{E}\{\text{PURE}_j\} = \mathscr{E}\{\text{MSE}_j\}$$

[1] A similar result for pure Poisson noise can be found in Hirakawa *et al.* 2009.

Interscale Haar-Wavelet-Domain LET

Lowpass: $\mathbf{s}$ Smoothed interscale predictor: $\mathbf{p}$ Magnitude of the highpass



$$\theta_n(\mathbf{d}_n, \mathbf{s}_n) = \underbrace{\gamma_n(p_n^2)\gamma_n(d_n^2)}_{\text{small predictor and small coefficient}} a_1 d_n + \underbrace{\overline{\gamma}_n(p_n^2)\gamma_n(d_n^2)}_{\text{large predictor and small coefficient}} a_2 d_n +$$

$$\underbrace{\gamma_n(p_n^2)\overline{\gamma}_n(d_n^2)}_{\text{small predictor and large coefficient}} a_3 d_n + \underbrace{\overline{\gamma}_n(p_n^2)\overline{\gamma}_n(d_n^2)}_{\text{large predictor and large coefficient}} a_4 d_n +$$

$$\underbrace{\gamma_n(p_n^2) a_5 \tilde{d}_n + \overline{\gamma}_n(p_n^2) a_6 \tilde{d}_n}_{\text{sign consistency enhancement}}$$

where $\gamma_n(x) = \mathrm{e}^{-\frac{|x|}{12(|s_n|+\sigma^2)}}$ and $\overline{\gamma}_n(x) = 1 - \gamma_n(x)$.

PURE for Arbitrary Nonlinear Processing

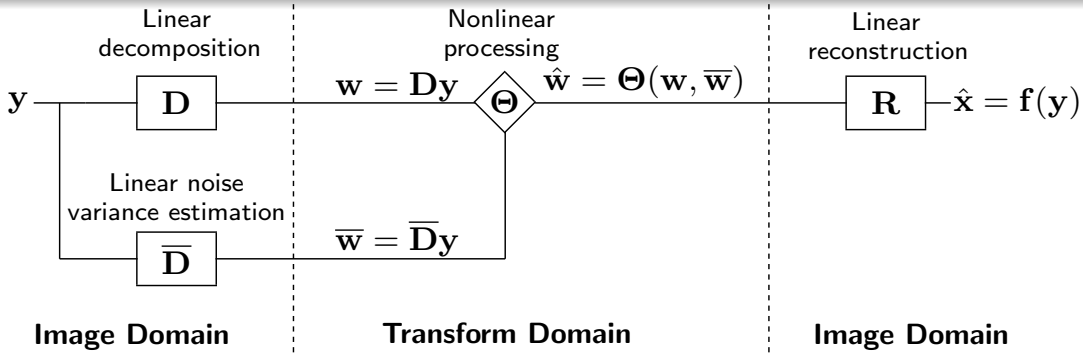
Problem: PURE is time-consuming to compute for an arbitrary nonlinear processing due to the term: $\mathbf{f}^-(\mathbf{y}) = [f_n(\mathbf{y} - \mathbf{e}_n)]_{1 \le n \le N}$.

Solution: First-order Taylor series approximation of $\mathbf{f}^-(\mathbf{y})$ given by $\mathbf{f}^-(\mathbf{y}) \simeq \mathbf{f}(\mathbf{y}) - \partial\mathbf{f}(\mathbf{y})$, where $\partial\mathbf{f}(\mathbf{y}) = [\frac{\partial f_n(\mathbf{y})}{\partial y_n}]_{1 \le n \le N}$.

Consequently, provided that each f_n varies slowly, **PURE is well-approximated** by

$$\widehat{\mathrm{PURE}} = \frac{1}{N}\left(\|\mathbf{f}(\mathbf{y})\|^2 - 2\mathbf{y}^{\mathrm{T}}(\mathbf{f}(\mathbf{y}) - \partial\mathbf{f}(\mathbf{y})) + 2\sigma^2 \mathrm{div}\left\{\mathbf{f}(\mathbf{y}) - \partial\mathbf{f}(\mathbf{y})\right\}\right) + \frac{1}{N}(\|\mathbf{y}\|^2 - \mathbf{1}^{\mathrm{T}}\mathbf{y}) - \sigma^2$$

PURE for Arbitrary Transform-Domain Processing

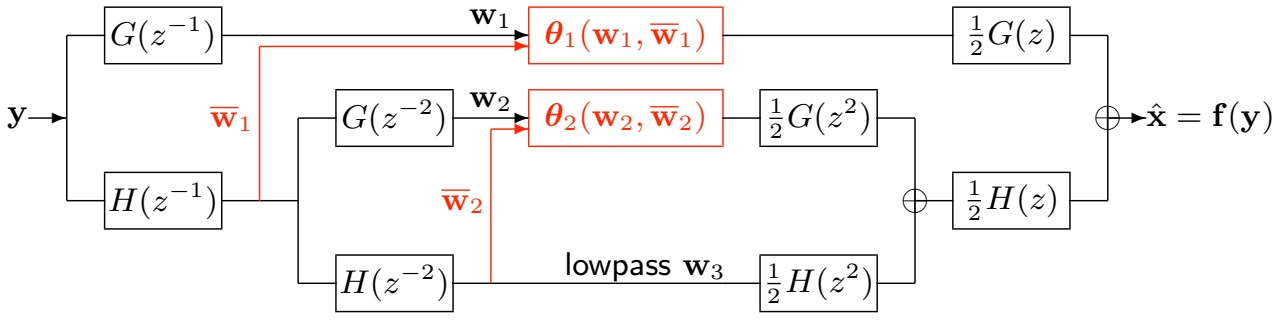


For *pointwise* processing $\boldsymbol{\Theta}(\mathbf{w}, \overline{\mathbf{w}}) = [\theta_l(w_l, \overline{w}_l)]_{1 \le l \le L}$, $\widehat{\text{PURE}}$ becomes:

$$\begin{aligned}\widehat{\text{PURE}} \;=\; & \frac{1}{N}\|\mathbf{f}(\mathbf{y}) - \mathbf{y}\|^2 + \frac{2}{N}\Big(\boldsymbol{\Theta}_1(\mathbf{w}, \overline{\mathbf{w}})^{\mathrm{T}}(\mathbf{D} \bullet \mathbf{R}^{\mathrm{T}})\mathbf{y} + \boldsymbol{\Theta}_2(\mathbf{w}, \overline{\mathbf{w}})^{\mathrm{T}}(\overline{\mathbf{D}} \bullet \mathbf{R}^{\mathrm{T}})\mathbf{y}\Big) + \\ & \frac{2\sigma^2}{N}\Big(\operatorname{diag}\{\mathbf{D}\mathbf{R}\}^{\mathrm{T}}\boldsymbol{\Theta}_1(\mathbf{w}, \overline{\mathbf{w}}) + \operatorname{diag}\{\overline{\mathbf{D}}\mathbf{R}\}^{\mathrm{T}}\boldsymbol{\Theta}_2(\mathbf{w}, \overline{\mathbf{w}})\Big) - \\ & \frac{2\sigma^2}{N}\Big(\operatorname{diag}\{(\mathbf{D} \bullet \mathbf{D})\mathbf{R}\}^{\mathrm{T}}\boldsymbol{\Theta}_{11}(\mathbf{w}, \overline{\mathbf{w}}) - \operatorname{diag}\{(\overline{\mathbf{D}} \bullet \overline{\mathbf{D}})\mathbf{R}\}^{\mathrm{T}}\boldsymbol{\Theta}_{22}(\mathbf{w}, \overline{\mathbf{w}}) - \\ & 2\operatorname{diag}\{(\mathbf{D} \bullet \overline{\mathbf{D}})\mathbf{R}\}^{\mathrm{T}}\boldsymbol{\Theta}_{12}(\mathbf{w}, \overline{\mathbf{w}})\Big) - \frac{1}{N}\mathbf{1}^{\mathrm{T}}\mathbf{y} - \sigma^2\end{aligned}$$

Example: Undecimated Haar Thresholding

Undecimated Haar filterbank: $H(z) = \frac{1}{\sqrt{2}}(1 + z)$ and $G(z) = \frac{1}{\sqrt{2}}(1 - z)$



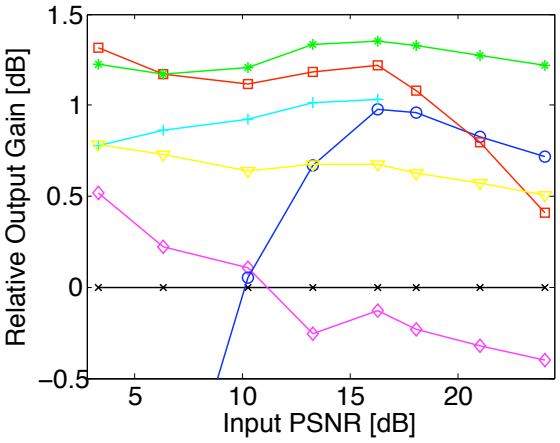
Subband-adaptive thresholding function:

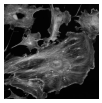
$$\theta_j(w, \overline{w}) = a_{j,1} \cdot w + a_{j,2} \cdot w \exp\left(-\left(\frac{w}{3t_j(\overline{w})}\right)^8\right)$$

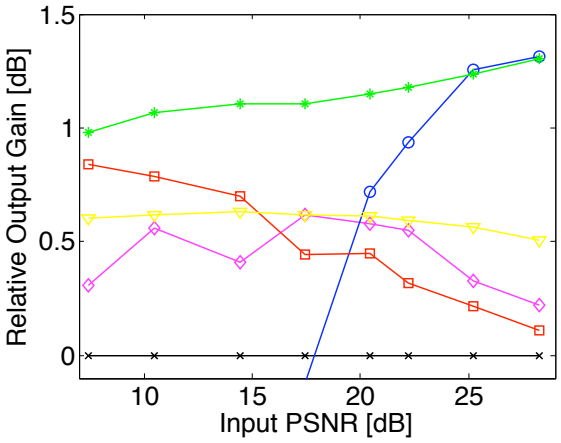
with signal-dependent threshold: $t_j(\overline{w}) = \sqrt{2^{-j/2}|\overline{w}| + \sigma^2}$

Some PSNR Comparisons

Cameraman 256×256 



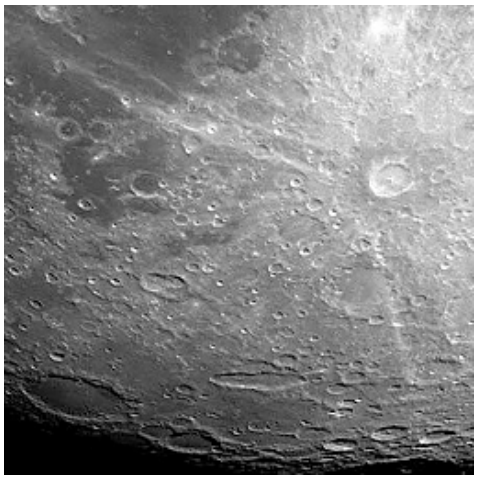
Cells 512×512 



Haar PURE-LET (baseline)
Haar PURE-LET (5 cycle-spins)
Redundant PURE-LET
Haar-Fisz Fryzlewicz & Nason 2004
Anscombe+BLS-GSM Portilla *et al.* 2003
Platelet Willett & Nowak 2007
PH-HMT Lefkimmiatis *et al.* 2009

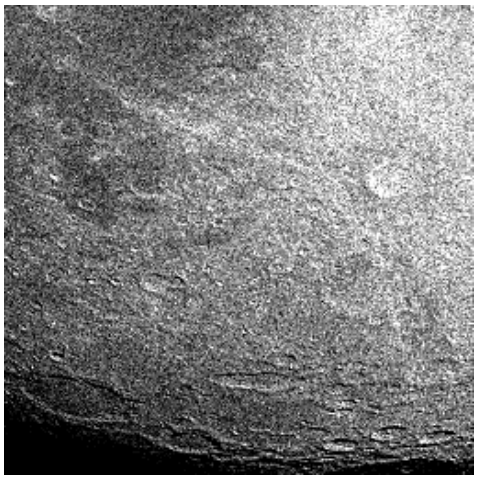
Some Visual Comparisons

Original



Average SSIM: 1.000

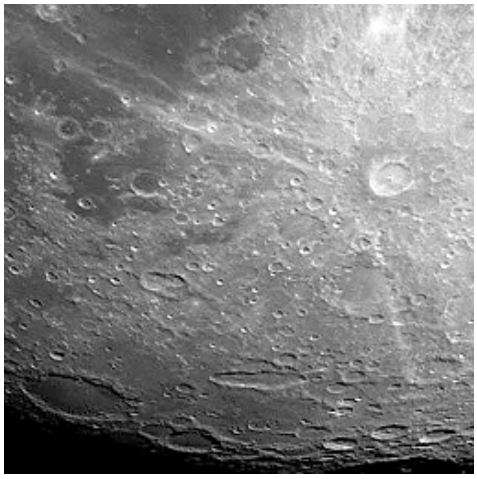
Noisy



Average SSIM: 0.385

Some Visual Comparisons

Original



Average SSIM: 1.000

Redundant PURE-LET

Average SSIM: 0.543

Some Visual Comparisons

Haar-Fisz

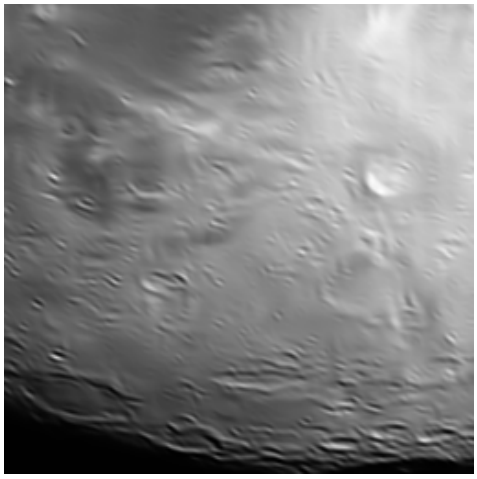
Average SSIM: 0.445

Redundant PURE-LET

Average SSIM: 0.543

Some Visual Comparisons

Anscombe+BLS-GSM



Average SSIM: 0.432

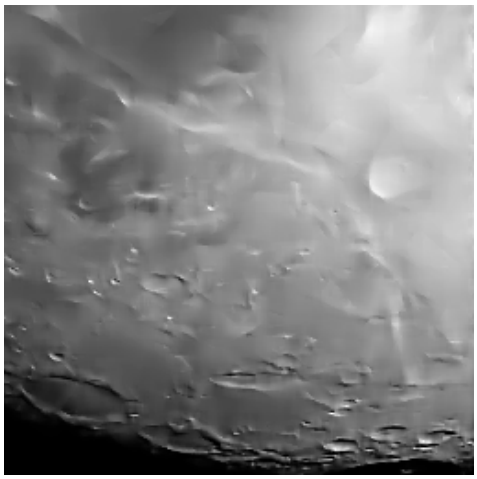
Redundant PURE-LET

Average SSIM: 0.543

Image Denoising Methods
The SURE-LET Methodology
SURE-LET Algorithmics
Algorithm Comparisons
Extension to Poisson-Gaussian Denoising
Poisson-Gaussian MSE Estimate
Interscale Haar-Wavelet Algorithm
Redundant Algorithm
Some Comparisons
Fluorescence Microscopy Results

Some Visual Comparisons

Platelet



Average SSIM: 0.420

Redundant PURE-LET

Average SSIM: 0.543

Some Visual Comparisons

Haar PURE-LET

Average SSIM: 0.520

Redundant PURE-LET

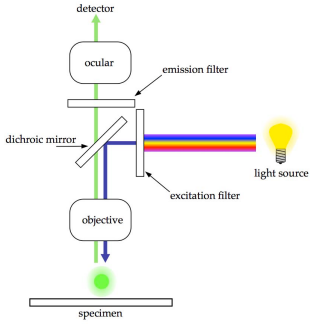
Average SSIM: 0.543

Fluorescence Microscopy

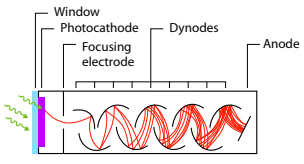
A fluorescence microscope is an imaging system that performs:

- Excitation of fluorescent constituents of a specimen;
- Focusing/filtering of the fluorescent light emitted from the specimen;
- Amplification/quantification of the light received at the ocular.

Combined with protein tagging (e.g., with GFP), fluorescence microscopy allows to image selected fine structures of living cells.



Optical description



Photomultiplier tube

Noise in Fluorescence Microscopy

Three main sources:

- **Photon-counting noise:** major source of noise due to the random nature of photon emission/detection (signal-dependent);
- **Measurement noise:** thermal instabilities of the various electronic devices (signal-independent);
- **Other:** autofluorescence and bleaching (reduced by short exposure and low fluorophore concentration).

$\rightsquigarrow$ **Measurement model:** scaled Poisson rdv degraded by AWGN

$$y \sim \alpha \mathcal{P}(x) + \mathcal{N}(\mu, \sigma^2)$$

α: detector gain μ: detector offset σ^2: AWGN variance

Noise Parameters Estimation

Affine relationship between sample-mean and sample-variance:

$$\left.\begin{array}{rcccl} \mu_y & \stackrel{\text{def}}{=} & \mathscr{E}\{y\} & = & \alpha x + \mu \\ \sigma_y^2 & \stackrel{\text{def}}{=} & \mathrm{Var}\{y\} & = & \alpha^2 x + \sigma^2 \end{array}\right\} \rightarrow \sigma_y^2 = \alpha \mu_y + \underbrace{\sigma^2 - \alpha\mu}_{\beta}$$

Simple estimation procedure: (similar to Lee 1989, Boulanger *et al.* 2007)

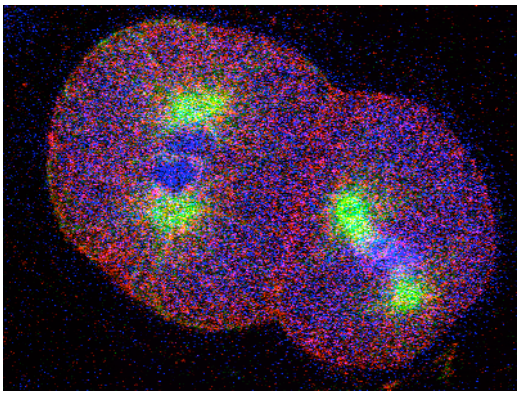
1. Compute μ_y and σ_y^2 in many small regions of the noisy image.
2. Perform a robust linear regression on the set of points (μ_y, σ_y^2).
3. Identify α as the slope of the fitted line and β as the ordinate at $\mu_y = 0$.
4. σ^2 and μ can be estimated independently in signal-free regions and cross-checked with β.

Experiments: 2D Sample

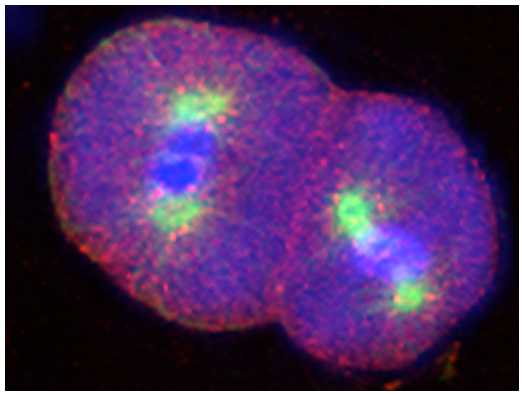
Specifications:

- 512×512 image acquired on a confocal microscope at the Imaging Center of the IGBMC, France;
- *C. elegans* embryo labeled with 3 fluorescent dyes;
- Each channel has been processed independently.

Raw Data



UWT PURE-LET

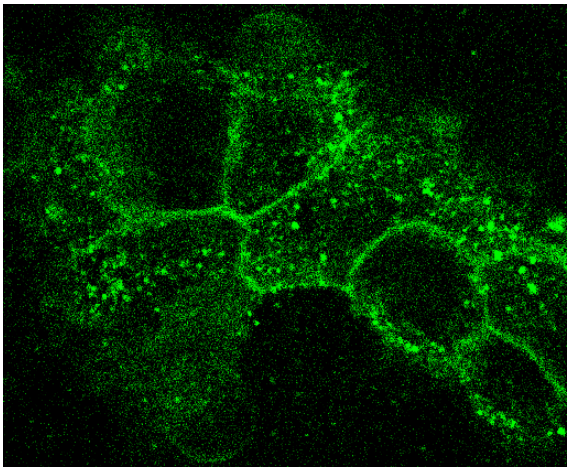


Experiments: 3D Sample

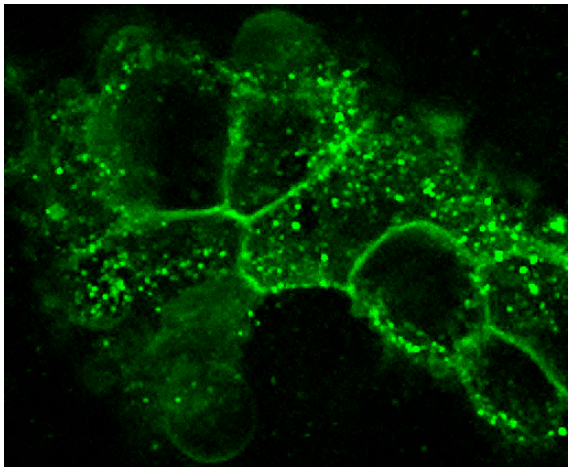
Specifications:

- $1024 \times 1024 \times 64$ volume of confocal microscopy images;
- Fibroblast cells labeled with *DiO* and 100nm fluorescent beads;
- Voxel resolution: $0.09 \times 0.09 \times 0.37 \mu\text{m}^3$.

Raw Data



Multislice Haar PURE-LET

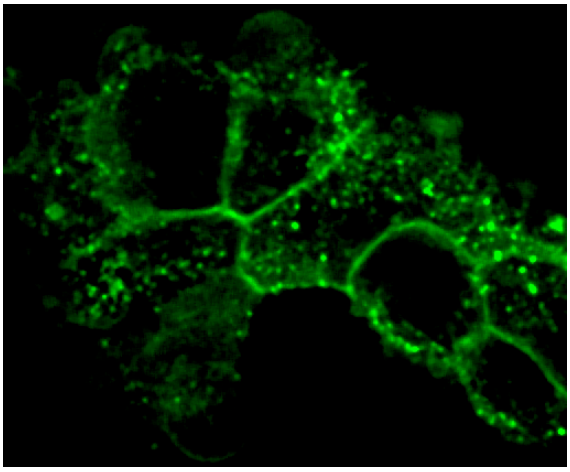


Experiments: 3D Sample

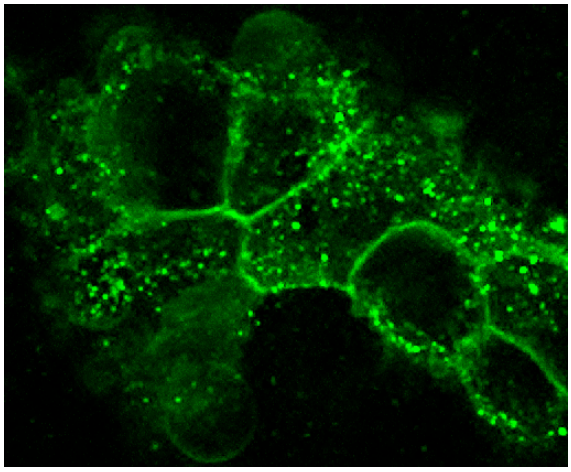
Specifications:

- $1024 \times 1024 \times 64$ volume of confocal microscopy images;
- Fibroblast cells labeled with *DiO* and 100nm fluorescent beads;
- Voxel resolution: $0.09 \times 0.09 \times 0.37 \mu\text{m}^3$.

3D Median Filter



Multislice Haar PURE-LET

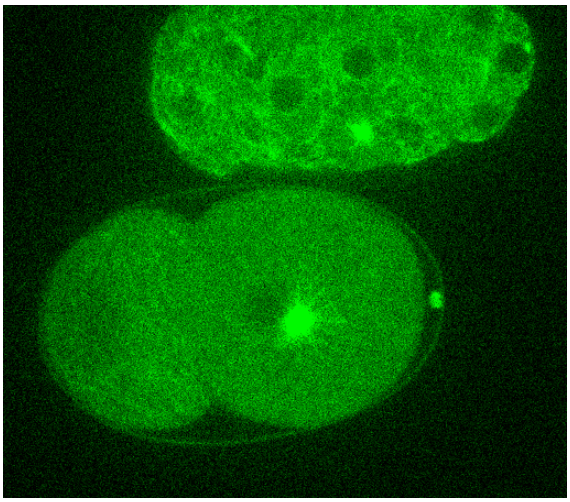


Experiments: 2D Timelapse Sequence

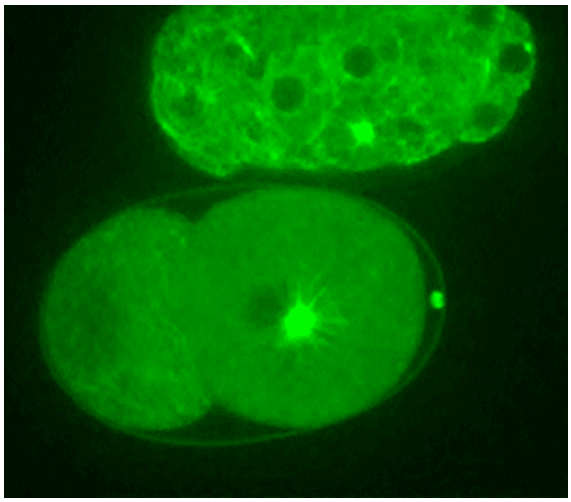
Specifications:

- $448 \times 512 \times 100$ image sequence of confocal microscopy images;
- *C. elegans* embryos labeled with GFP;

Raw Data



Multiframe Haar PURE-LET



Conclusion

Presentation of a **generic methodology** for building signal/image denoising algorithms.

Advantages:

- Does not require hypotheses on the signal, only on the noise (SURE/PURE);
- No parameters to tune;
- Fast, non-iterative (SURE/PURE + LET);
- Natural construction of multivariate/redundant thresholding rules.

Although they involve only simple thresholding operations in a transformed domain (single step, no training, no block-matching, no direction/edge detection), the proposed algorithms reach the state of the art in image/video denoising.

Main References

Luisier *et al.*, "Image Denoising in Mixed Poisson-Gaussian Noise", *IEEE Transactions on Image Processing*. To appear (2010).

Luisier *et al.*, "Fast Interscale Wavelet Denoising of Poisson-Corrupted Images", *Signal Processing*, Vol. 90 (2), pp. 415-427, February 2010.

Luisier *et al.*, "SURE-LET for Orthonormal Wavelet-Domain Video Denoising", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 20 (6), pp. 913-919, June 2010.

Luisier *et al.*, "SURE-LET Multichannel Image Denoising: Interscale Orthonormal Wavelet Thresholding", *IEEE Transactions on Image Processing*, Vol. 17 (4), pp. 482-492, April 2008.

Blu *et al.*, "The SURE-LET Approach to Image Denoising", *IEEE Transactions on Image Processing*, Vol. 16 (11), pp. 2778-2786, November 2007.

Luisier *et al.*, "A New SURE Approach to Image Denoising: Interscale Orthonormal Wavelet Thresholding", *IEEE Transactions on Image Processing*, Vol. 16 (3), pp. 593-606, March 2007. **Young Author Best Paper Award 2009**.

Internet links

- **Authors**: thierry.blu@m4x.org and florian.luisier@a3.epfl.ch
- **Papers**: www.ee.cuhk.edu.hk/~tblu/ and bigwww.epfl.ch/
- **Demos**: bigwww.epfl.ch/
  Orthonormal grayscale and color image denoising
- **Software**: bigwww.epfl.ch/
  Matlab implementations of SURE-LET algorithms
  PURE-LET denoising plugin for ImageJ

A New SURE Approach to Image Denoising: Interscale Orthonormal Wavelet Thresholding

Florian Luisier, Thierry Blu, *Senior Member, IEEE*, and Michael Unser, *Fellow, IEEE*

***Abstract*—This paper introduces a new approach to orthonormal wavelet image denoising. Instead of postulating a statistical model for the wavelet coefficients, we directly parametrize the denoising process as a sum of elementary nonlinear processes with unknown weights. We then minimize an estimate of the mean square error between the clean image and the denoised one. The key point is that we have at our disposal a very accurate, statistically unbiased, MSE estimate—Stein's unbiased risk estimate—that depends on the noisy image alone, not on the clean one. Like the MSE, this estimate is quadratic in the unknown weights, and its minimization amounts to solving a linear system of equations. The existence of this *a priori* estimate makes it unnecessary to devise a specific statistical model for the wavelet coefficients. Instead, and contrary to the custom in the literature, these coefficients are not considered random anymore. We describe an interscale orthonormal wavelet thresholding algorithm based on this new approach and show its near-optimal performance—both regarding quality and CPU requirement—by comparing it with the results of three state-of-the-art nonredundant denoising algorithms on a large set of test images. An interesting fallout of this study is the development of a new, group-delay-based, parent–child prediction in a wavelet dyadic tree.**

***Index Terms*—Image denoising, interscale dependencies, orthonormal wavelet transform, Stein's unbiased risk estimate (SURE) minimization.**

I. Introduction

DURING acquisition and transmission, images are often corrupted by additive noise that can be modeled as Gaussian most of the time. The main aim of an image denoising algorithm is then to reduce the noise level, while preserving the image features. The multiresolution analysis performed by the wavelet transform has been shown to be a powerful tool to achieve these goals. Indeed, in the wavelet domain, the noise is uniformly spread throughout the coefficients, while most of the image information is concentrated in the few largest ones (sparsity of the wavelet representation).

The most straightforward way of distinguishing information from noise in the wavelet domain consists of thresholding the wavelet coefficients. Of the various thresholding strategies, *soft-thresholding* is the most popular and has been theoretically justified by Donoho and Johnstone [1]. These authors have shown that the shrinkage rule is near-optimal in the minimax sense and provided the expression of the optimal threshold value T—called the "universal threshold"—as a function of the noise power σ^2 when the number of samples N is large: $T = \sqrt{2\sigma^2 \log N}$. The use of the universal threshold to denoise images in the wavelet domain is known as *VisuShrink* [2].

Yet, despite its theoretical appeal, minimax is different from mean-squared error (MSE) as a measure of error. A lot of work has been done to propose alternative thresholding strategies that behave better in terms of MSE than *VisuShrink* [3]–[6]. Donoho and Johnstone themselves acknowledged this flaw and suggested to choose the optimal threshold value T by minimizing Stein's unbiased risk estimator (SURE) [7] when the data fail to be sparse enough for the minimax theory to be valid. This hybrid approach has been coined *SureShrink* by their authors [1]. Without challenging the soft-thresholding strategy, alternative threshold value selections have been proposed as well. One of the most popular was proposed by Chang *et al.*, who derived their threshold in a Bayesian framework, assuming a generalized Gaussian distribution for the wavelet coefficients. This solution to the wavelet denoising problem is known as *BayesShrink* [8] and has a better MSE performance than *SureShrink*.

Beyond the pointwise approach, more recent investigations have shown that substantially larger denoising gains can be obtained by considering the intra- and interscale correlations of the wavelet coefficients. In addition, increasing the redundancy of the wavelet transform is strongly beneficial to the denoising performances, a point to which we will come back later. We have selected three such techniques reflecting the state-of-the-art in wavelet denoising, against which we will compare our results.

- *Portilla et al.* [9]:[1] Their main idea is to model the neighborhoods of coefficients at adjacent positions and scales as a Gaussian scale mixture (GSM); the wavelet estimator is then a Bayes least squares (BLS). Their denoising method, consequently called *BLS-GSM*, is the most efficient up-to-date approach.
- *Pižurica et al.* [10]:[2] Assuming a generalized Laplacian prior for the noise-free data, their approach called *ProbShrink* is driven by the estimation of the probability that a given coefficient contains significant information—notion of "signal of interest."
- *Sendur et al.* [11], [12]:[3] Their method, called *BiShrink*, is based on new non-Gaussian bivariate distributions

Manuscript received February 28, 2006; revised September 14, 2006. This work was supported in part by the Center for Biomedical Imaging (CIBM) of the Geneva–Lausanne Universities and the EPFL, in part by the foundations Leenaards and Louis-Jeantet, and in part by the Swiss National Science Foundation under Grant 200020-109415. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Srdjan Stankovic.

The authors are with the Biomedical Imaging Group (BIG), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: florian.luisier@epfl.ch; thierry.blu@epfl.ch; michael.unser@epfl.ch).

Digital Object Identifier 10.1109/TIP.2007.891064

[1]Available at http://www.decsai.ugr.es/~javier/denoise/software/index.htm, with a 3 × 3 neighborhood as suggested by the authors.

[2]Available at http://www.telin.ugent.be/~sanja/, with a 3 × 3 neighborhood and a threshold value $T = \sigma$ as suggested by the authors.

[3]Available at http://www.taco.poly.edu/WaveletSoftware/denoise2.html, with a 7 × 7 neighborhood as suggested by the authors.

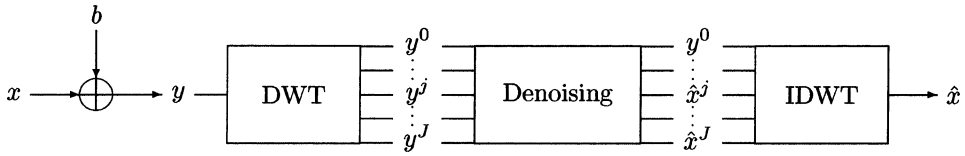


Fig. 1. Principle of wavelet denoising.

to model interscale dependencies. A nonlinear bivariate shrinkage function using the maximum *a posteriori* (MAP) estimator is then derived. In a second paper, these authors have extended their approach by taking into account the intrascale variability of wavelet coefficients.

These techniques have been devised for both redundant and nonredundant transforms.

Despite reports on the superior denoising performances of redundant transforms [13], [14], we will only consider critically sampled wavelet transforms in this paper. The rationale behind our choice is that, since there is no *added* information—only *repeated* information—in redundant transforms, we believe that, eventually, a nonredundant transform may match the performance of redundant ones. This would potentially be very promising since the major drawback of redundant transforms are their memory and CPU time requirements which limits their routine use for very large images and, above all, usual volumes of data.

More than a specific denoising algorithm, this paper is about a powerful new method for optimizing *beforehand*—unaware of the clean image—the performance of a denoising method. Here, we want in particular to promote Stein's unbiased risk estimate (SURE) which is nothing less than an *a priori* estimation of the MSE resulting from an arbitrary processing of noisy data. This estimate turns out to be more accurate as more data are available, which is the case of images. Wavelet denoising methods routinely involve a statistical description of the coefficient distribution [15], an estimation of the—always nonlinear—statistical parameters and then, a search for the best denoising algorithm for this type of statistics. In contrast, by taking advantage of Stein's MSE estimate, our method goes directly to the last step, without caring for the statistical description: in short, we do not make any explicit hypotheses on the clean image. In fact, we do not consider it as a random process at all; the randomness in our formulation follows from the Gaussian white noise alone.

Our approach consists, thus, in parametrizing the denoising method and choosing the parameters that minimize this MSE estimate. Previous techniques using the SURE required the minimization of complicated expressions for few nonlinear parameters [16], [17] or the use of parallel block iterative convex programming [18]. What makes our approach more tractable and efficient, is precisely the parametrizing method: a *linear combination* of nonlinear denoising functions—thresholding functions. Because of this "linear" choice, the minimization of the MSE estimate merely amounts to solving a linear system of equations, whose size is the number of weights in the linear combination. Obviously, the number of parameters, or degrees of freedom, is not a challenge and highly complicated thresholding behaviors can be obtained this way. In the context of image denoising, a univariate linear parametrization combined with an implicit SURE minimization was already evoked in [19] (*sigmoidal filtering*).

Because of the particular simplicity of Stein's estimate for pointwise denoising functions, we will not exploit the full potential of the theory in this paper and will only consider interscale pointwise thresholding in the orthonormal wavelet transform. This excludes any intrascale considerations. Yet, we will show that our denoising method performs better than the nonredundant versions of the state-of-the-art methods [9], [10], [12] on almost all tested images, to the noteworthy exception of *Barbara*, which may require intrascale processing. Without any optimization attempts in our implementation, the comparison of computation times already show how economical our method is.

The paper is organized as follows. In Section II, we expose the SURE theory for functions of one or several statistically independent variables, and sketch the principles of our parametrization strategy. In Section III, we show how these principles can be exploited to build an efficient pointwise thresholding function that outperforms all known pointwise techniques. In Section IV, we extend the approach to a thresholding function that involves coarser scale parents as well. On this occasion, we develop a new formula to build a parent coefficient out of parent subbands, and, finally, we compare our denoising method to the best available nonredundant techniques (Section V). Both the competitiveness and robustness of our method validate our new approach as an attractive solution for image denoising.

II. Theoretical Elements

A. Problem Setting

Wavelet denoising consists of three main stages (see Fig. 1).

i) Perform a discrete wavelet transform (DWT) to the noisy data $y = (y_n)_{n\in[1,N]}$ which is the sum of the noise-free data $x = (x_n)_{n\in[1,N]}$ and the noise $b = (b_n)_{n\in[1,N]}$.

ii) Denoise J noisy wavelet subimages $y^j = x^j + b^j = (y_n^j)_{n\in[1,N_j]}$, $j \in [1, J]$ by computing J estimates $\hat{x}^j$ of the noise-free highpass subbands x^j.

iii) Reconstruct the denoised image by applying the inverse discrete wavelet transform (IDWT) on the processed highpass wavelet subimages $\hat{x}^j$ to obtain an estimate $\hat{x}$ of the noise-free data x.

One can make two important remarks that set the context in which we will develop our denoising method.

- We will only consider additive Gaussian white noise following a normal law defined by a zero mean and a known[4] σ^2 variance; i.e., $b \sim \mathcal{N}(0, \sigma^2)$.

[4]In practice, the noise standard deviation can be accurately estimated using a robust median estimator [1].

- We will only consider *orthonormal* wavelet transform; the consequences are as follows.
  - — The mean-square error (MSE) in the space domain is a weighted sum of the MSE of each individual subband

$$\underbrace{\left\langle |\hat{x}-x|^2 \right\rangle}_{\text{MSE}} = \sum_{j=0}^{J} \frac{N_j}{N} \underbrace{\left\langle |\hat{x}^j - x^j|^2 \right\rangle}_{\text{MSE}^j} \tag{1}$$

    where we have introduced the notation

$$\langle u \rangle = \frac{1}{N} \sum_{n=1}^{N} u_n \tag{2}$$

    for the statistical mean estimate.
  - — The noise remains white and Gaussian with same statistics in the orthonormal wavelet domain, i.e., $b^j \sim \mathcal{N}(0, \sigma^2)$.

This allows us to apply a new denoising function independently in every highpass subband, which means that our solution is subband-adaptive like most of the successful wavelet denoising approaches.

B. Stein's Unbiased MSE Estimate (SURE)

In denoising applications, the performance is often measured in terms of peak signal-to-noise ratio (PSNR), which can be defined as follows:[5]

$$\text{PSNR} = 10 \log_{10} \left(\frac{\max(x^2)}{\langle |\hat{x}-x|^2 \rangle} \right). \tag{3}$$

Since the noise is a random process, we introduce an expectation operator $\mathcal{E}\{\}$ to guess the potential results obtained after processing the noisy data y. Note that the noise-free data x is not modeled as a random process; thus, $\mathcal{E}\{x\} = x$.

The aim of image denoising is naturally to maximize the PSNR and, thus, to minimize the MSE defined in (1). In this paper, we choose to estimate each x^j by a pointwise function of y^j

$$\left(\hat{x}_n^j\right)_{n \in [1,N_j]} = \left(\theta^j \left(y_n^j\right)\right)_{n \in [1,N_j]}.$$

From now on, we will drop the subband index j since a new denoising function is independently applied in each individual subband. Our goal is to find a function θ that minimizes

$$\text{MSE} = \left\langle |\theta(y) - x|^2 \right\rangle = \left\langle \theta(y)^2 \right\rangle - 2 \left\langle x\theta(y) \right\rangle + \langle x^2 \rangle. \tag{4}$$

In practice, we only have access to the noisy signal $y = x + b$, and not to the original signal x. In (4), we, thus, need to remove the explicit dependence on x. Note that, since $\langle x^2 \rangle$ has no influence in the minimization process, we do not need to estimate it. The remaining problematic term is only $\langle x\theta(y) \rangle$. However, the following theorem, a version of which was proposed by Stein in [7], allows us to overcome this difficulty.

[5]For 8-bit images, usually $\max(x^2) = 255^2$.

Theorem 1: Let $\theta : \mathbb{R} \to \mathbb{R}$ be a (weakly) differentiable function that does not explode at infinity.[6] Then, the following random variable:

$$\begin{aligned} \epsilon &= \left\langle \theta(y)^2 - 2y\theta(y) + 2\sigma^2\theta'(y) \right\rangle + \langle x^2 \rangle \\ &= \underbrace{\frac{1}{N} \sum_{n=1}^{N} \left(\theta^2(y_n) - 2y_n\theta(y_n) + 2\sigma^2\theta'(y_n) \right)}_{\tilde{\epsilon}} + \langle x^2 \rangle \end{aligned} \tag{5}$$

is an unbiased estimator of the MSE, i.e.

$$\mathcal{E}\{\epsilon\} = \mathcal{E}\left\{ \left\langle |\theta(y) - x|^2 \right\rangle \right\}.$$

Proof: We can develop the square error between x_n and its estimate $\theta(y_n)$ as

$$\begin{aligned} \mathcal{E}\left\{ |\theta(y_n) - x_n|^2 \right\} &= \mathcal{E}\left\{\theta^2(y_n)\right\} - 2\mathcal{E}\left\{x_n\theta(y_n)\right\} + \mathcal{E}\left\{x_n^2\right\} \\ &= \mathcal{E}\left\{\theta^2(y_n)\right\} - 2\mathcal{E}\left\{y_n\theta(y_n)\right\} \\ &\quad + 2\mathcal{E}\left\{b_n\theta(y_n)\right\} + x_n^2 \end{aligned}$$

where each term is well-defined thanks to the hypothesis on θ.

We then use the fact that the Gaussian probability density $q(b_n)$ satisfies $b_n q(b_n) = -\sigma^2 q'(b_n)$ to evaluate $\mathcal{E}\{b_n\theta(y_n)\}$

$$\begin{aligned} \mathcal{E}\left\{b_n\theta(y_n)\right\} &= \int \theta(x_n + b_n) b_n q(b_n) db_n \\ &= -\sigma^2 \int \theta(x_n + b_n) q'(b_n) db_n \\ &= \sigma^2 \int \theta'(x_n + b_n) q(b_n) db_n \quad \text{(by parts)} \\ &= \sigma^2 \mathcal{E}\left\{\theta'(y_n)\right\}. \end{aligned} \tag{6}$$

Note that the integrated part $[\sigma^2\theta(x_n + b_n)q(b_n)]_{-\infty}^{+\infty}$ vanishes by hypothesis. This is known as Stein's Lemma [7] and leads to

$$\begin{aligned} \mathcal{E}\left\{ |\theta(y_n) - x_n|^2 \right\} &= \mathcal{E}\left\{\theta^2(y_n)\right\} - 2\mathcal{E}\left\{y_n\theta(y_n)\right\} \\ &\quad + 2\sigma^2\mathcal{E}\left\{\theta'(y_n)\right\} + x_n^2. \end{aligned}$$

Since the expectation of a sum is equal to the sum of the expectations, we can deduce that

$$\begin{aligned} \mathcal{E}\left\{ \left\langle |\theta(y) - x|^2 \right\rangle \right\} &= \mathcal{E}\left\{\left\langle \theta^2(y) \right\rangle\right\} - 2\mathcal{E}\left\{\left\langle y\theta(y) \right\rangle\right\} \\ &\quad + 2\sigma^2\mathcal{E}\left\{\left\langle \theta'(y) \right\rangle\right\} + \langle x^2 \rangle. \end{aligned}$$

■

As said before, there is no need to estimate $\langle x^2 \rangle$, since this term will disappear in the minimization. So, in practice, we will consider $\tilde{\epsilon}$ which is the only part of the MSE estimate that depends on the choice of the denoising function θ.

Note that Theorem 1 is still valid if $\theta(y)$ is replaced by a two-variable denoising function $\theta(y, z)$ where y is random, but independent[7] of y. In particular, in an orthonormal wavelet transform—which transforms Gaussian white noise into Gaussian

[6]Typically, such that $|\theta(z)| \leq \text{Const} \cdot \exp(az^2)$ for $a < (1/2\sigma^2)$.

[7]We recall that the randomness of $y = x + b$ only results from the Gaussian white noise b, because no statistical model is assumed on the noise-free data x.

white noise—z can be any wavelet coefficients other than y itself.

The result given by Theorem 1 becomes particularly interesting in image denoising applications, where the number of samples is large. Indeed, by the law of large numbers, the standard deviation of ϵ is small; i.e., the estimate ϵ is close to its expectation which is the MSE of the denoising procedure. As a result, we can use ϵ as if it were the true MSE. The next section shows how to use Theorem 1 efficiently.

C. SURE Approach to Image Denoising

Our denoising approach amounts to minimizing ϵ over a range of reasonable denoising functions θ. We claim that this will result in the minimization of the MSE over the same range of functions, up to a small random error inversely proportional to the square root of the number of samples. Before defining more precisely which denoising functions we consider reasonable, we can illustrate the search for the optimal value T by applying Theorem 1 when θ is the well-known *soft-thresholding* function defined by

$$\theta(y) = \operatorname{sign}(y)\left(|y| - T\right)_+ \tag{7}$$

where $(x)_+ = \max(x, 0)$.

By Theorem 1, the following expression has to be minimized over T

$$\tilde{\epsilon}(T) = \left\langle \left(2\sigma^2 + T^2 - y^2\right)\left(|y| - T\right)_+^0 \right\rangle. \tag{8}$$

The last expression has its minimum exactly for the same T as the following formula:

$$\operatorname{SURE}(T; y) = \sigma^2 - \frac{1}{N} \times \left(2\sigma^2 \cdot \#\left\{n : |y_n| \le T\right\} - \sum_{n=1}^{N} \min\left(|y_n|, T\right)^2\right) \tag{9}$$

which appears in [1].

The estimated optimal threshold value is then: $\tilde{T}_{\text{opt}} = \arg\min_T(\operatorname{SURE}(T; y)) = \arg\min_T(\tilde{\epsilon}(T))$.

We must notice here that the so-called *SureShrink* procedure developed by Donoho and Johnstone in [1] uses, in fact, a hybrid scheme between the SURE theory and the universal threshold (asymptotically optimal when the data exhibit a high level of sparsity). Their minimization of $\operatorname{SURE}(T; y)$ is, thus, restricted to $T \in [0; T_{\text{univ}}]$, where $T_{\text{univ}} = \sqrt{2\sigma^2 \log N}$ is the universal threshold. Our opinion, however, is that this restriction is unnecessary—and often suboptimal—in image denoising applications where quality is measured by a mean-square criterium. This is because, even though natural images have small wavelet coefficients, these are not vanishing as required by the strict sparsity results. It may even be argued that these small coefficients convey important texture information and should not, thus, be set to zero.

As we can verify in Fig. 3, the estimate of the theorem is statistically very reliable and robust, making it completely suitable for an accurate estimation of the optimal threshold.

The *soft-thresholding* function (see Fig. 2) exhibits two main drawbacks. First, it only depends on a single parameter T, and, thus, its shape is not very flexible; second, this dependency is not linear. The consequence of these two remarks is that the sensitivity of the *soft-thresholding* function with respect to the value of T is high, and that finding the optimal threshold requires a nonlinear search algorithm.

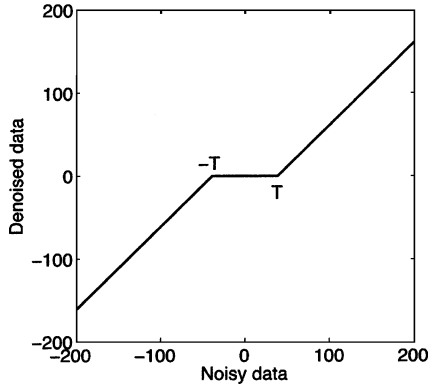


Fig. 2. *Soft-thresholding* function.

In order to mitigate this issue, we choose to build a denoising function that depends *linearly* on a set of parameters—degrees of freedom—which we will determine exactly by minimizing ϵ. The exact minimization is especially simple (linear) because the MSE estimate ϵ has a quadratic form, much like the true MSE. The key idea is, thus, to build a linearly parameterized denoising function of the form

$$\theta(y) = \sum_{k=1}^{K} a_k \varphi_k(y) \tag{10}$$

where K is the number of parameters.

If we introduce (10) into the estimate of the MSE given in Theorem 1 and perform differentiations over the a_k, we obtain for all $k \in [1; K]$

$$0 = \frac{1}{2}\frac{\partial \epsilon}{\partial a_k} = \left\langle \theta(y)\varphi_k(y) - y\varphi_k(y) + \sigma^2 \varphi_k'(y) \right\rangle$$

$$\Updownarrow$$

$$\sum_{l=1}^{K} \underbrace{\left\langle \varphi_k(y)\varphi_l(y) \right\rangle}_{M_{k,l}} a_l - \underbrace{\left\langle y\varphi_k(y) - \sigma^2 \varphi_k'(y) \right\rangle}_{c_k} = 0.$$

These equations can be summarized in matrix form as $\mathbf{Ma} = \mathbf{c}$, where $\mathbf{a} = [a_1 \ldots a_K]^T$ and $\mathbf{c} = [c_1 \ldots c_K]^T$ are vectors of size $K \times 1$, and $\mathbf{M} = [M_{k,l}]_{1 \le k,l \le K}$ is a matrix of size $K \times K$. This linear system is solved for $\mathbf{a}$ by

$$\mathbf{a} = \mathbf{M}^{-1}\mathbf{c} \tag{11}$$

which makes our approach very simple to implement. Note that, since we are only interested in the minimum of ϵ, we are ensured that there will always be a solution. When several solutions are admissible (e.g., when $\operatorname{rank}(\mathbf{M}) < K$) any one of them will be acceptable—in particular, the one provided by the pseudoinverse of $\mathbf{M}$. When this degeneracy occurs, we will conclude that the parameters a_k belong to some linear subspace and, thus, that some of them are useless (the function is "over-parameterized"). Of course, it is desirable to keep the number of degrees of freedom K as low as possible in order for the estimate ϵ to keep a small variance.

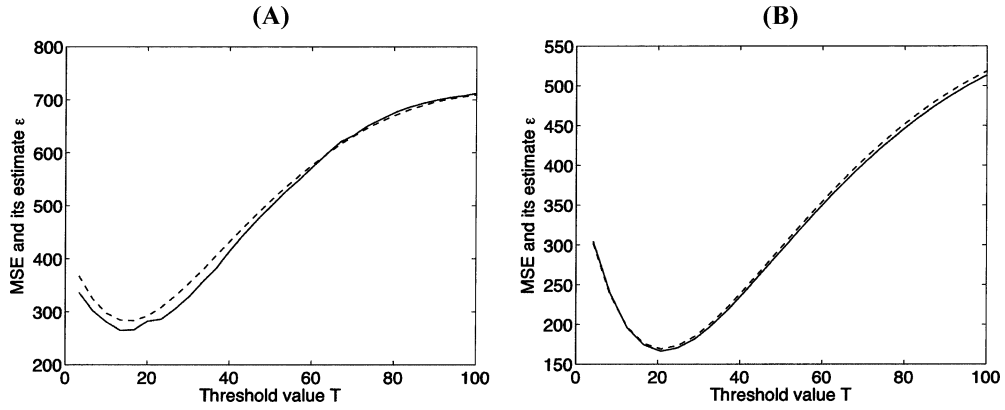


Fig. 3. Statistical accuracy of Theorem 1 illustrated with the *soft-threshold*: true MSE is in dashed lines, while its estimate ϵ is in solid line. (a) $N = 32 \times 32$ samples and $\sigma = 20$. (b) $N = 256 \times 256$ samples and $\sigma = 20$. The variance of the estimator decreases when the number of samples N increases, making Theorem 1 statistically reliable for image denoising applications.

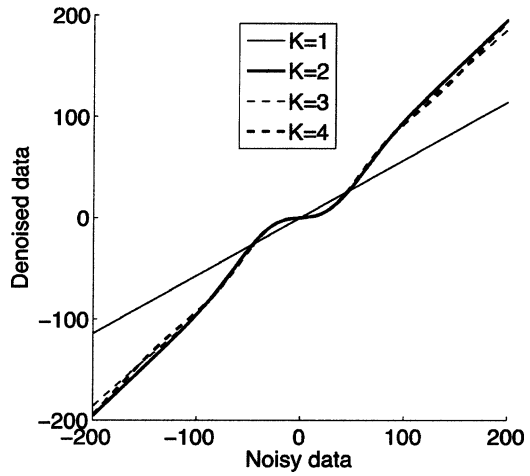


Fig. 4. Shape of our denoising function (12) in a particular subband, for various K and optimized a_k's and T.

III. Efficient SURE-Based Pointwise Thresholding

In the previous section, we have proposed a general form of denoising functions (10). The difficulty is now to choose suitable basis functions φ_k that will determine the shape of our denoising function. Therefore, we want the denoising function θ to satisfy the following properties:

- differentiability: required to apply Theorem 1;
- anti-symmetry: the wavelet coefficients are not expected to exhibit a sign preference;
- linear behavior for large coefficients: because $\theta(y)$ should asymptotically tend to y.

After trying several types of φ_k, we have found that all of them give quite similar results, when the above conditions are satisfied. We have, thus, decided to retain the following pointwise denoising function:

$$\theta(y) = \sum_{k=1}^{K} a_k y e^{-(k-1)\frac{y^2}{2T^2}}. \tag{12}$$

We choose derivatives of Gaussians (DOG) because they decay quite fast, which ensures a linear behavior close to the identity for large coefficients (see Fig. 4).

In addition to the linear coefficients, our denoising function contains two nonlinear dependencies: the number of terms K and the parameter T. We will see later that they can be fixed independently of the image.

If we consider only one parameter ($K = 1$), our denoising function simply becomes $\theta(y) = a_1 y$, which is the simplest linear pointwise denoising function. The direct minimization of the estimate ϵ provides

$$a_1 = 1 - \frac{\sigma^2}{\langle y^2 \rangle} \tag{13}$$

which is known as the James–Stein estimator [20].

Practical tests (with optimization over the parameter T, independently in each subband) on various images and with various noise levels have shown that, as soon as $K \geq 2$, the results become quite similar. It, thus, appears that it is sufficient to keep as few as $K = 2$ terms in (12). This is confirmed in Fig. 4, which shows that the shape of our denoising function is nearly insensitive to the variation of $K \geq 2$.

Moreover, the optimal value of the parameter T is closely linked to the standard deviation σ of the noise and in a lesser way to the number of parameters K. Its interpretation is quite similar as in the case of the *soft-threshold*: It manages the transition between low SNR to high SNR coefficients. In our case though, the variations of the minimal ϵ (over a_k) when T changes are quite small (see Fig. 5), because our denoising function is much more flexible than the *soft-threshold*. This sensitivity becomes even smaller as the number of parameters K increases. In fact, this indicates that some parameters are, in that case, useless.

To summarize, we have shown that both the number of terms K and the parameter T have only a minor influence on the quality of the denoising process. This indicates that these two parameters do not have to be optimized; instead, they can be fixed once for all, independently of the type of image. From a practical point of view, we suggest to use $K = 2$ terms and $T = \sqrt{6}\sigma$ (see Fig. 5), leading to the following pointwise thresholding function:

$$\theta_0(y; \mathbf{a}) = \left(a_1 + a_2 e^{-\frac{y^2}{12\sigma^2}} \right) y. \tag{14}$$

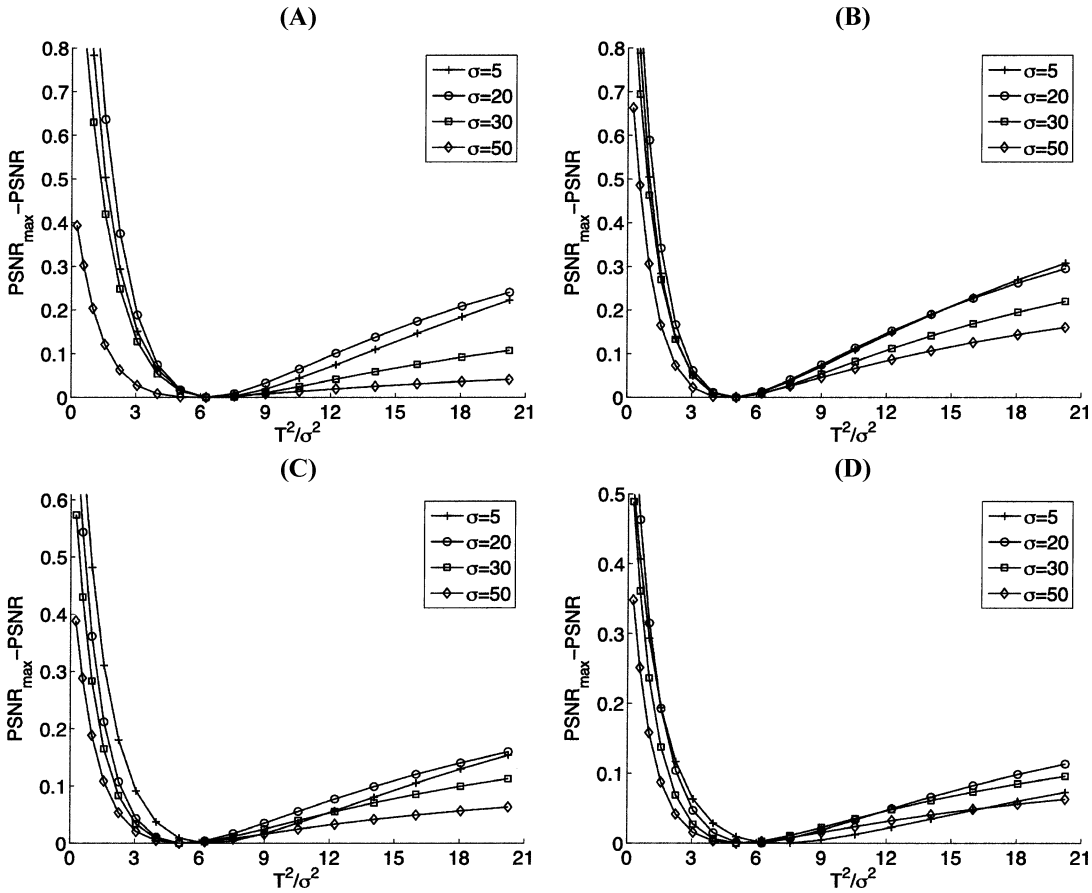


Fig. 5. Sensitivity of our denoising function (14) with respect to variations of T. (a) *Peppers* 256 × 256. (b) *MIT* 256 × 256. (c) *Lena* 512 × 512. (d) *Boat* 512 × 512. We can notice that for all images and for the whole range of input PSNR the maximum of the PSNR is reached for $(T^2/\sigma^2) \simeq 6$.

TABLE I
COMPARISON OF OUR SUM OF DOG (14) WITH THE ORACLE *SOFT-THRESHOLD* (NONREDUNDANT *SYM8*, FOUR ITERATIONS)

σ	5	10	20	30	50	5	10	20	30	50
Method	**Boat** 512 × 512					**Goldhill** 512 × 512				
OracleShrink	36.09	32.11	28.64	26.81	24.79	35.99	31.97	28.75	27.18	25.45
Sum of DOG (*Oracle*)	**36.35**	**32.37**	**28.85**	**27.03**	**25.01**	**36.21**	**32.25**	**28.99**	**27.42**	**25.67**
Sum of DOG (SURE)	**36.35**	**32.37**	**28.85**	27.02	25.00	**36.21**	**32.25**	**28.99**	27.41	25.66
Method	**Peppers** 256 × 256					**Bridge** 256 × 256				
OracleShrink	36.38	32.06	28.03	25.84	23.34	34.83	29.81	25.77	23.93	22.06
Sum of DOG (*Oracle*)	**36.67**	**32.36**	**28.28**	**25.97**	**23.47**	**34.89**	**30.00**	**26.10**	**24.29**	**22.40**
Sum of DOG (SURE)	**36.67**	32.35	28.67	25.95	23.45	**34.89**	**30.00**	26.09	24.28	22.39

Note: output PSNRs have been averaged over ten noise realizations.

Now, it is interesting to evaluate the efficiency of our denoising function (14) and the accuracy of our minimization process based on an estimate ϵ of the MSE. We propose to compare our results with the best results that can be reached by the popular *soft-threshold* with an optimal threshold choice (*OracleShrink*). Two main observations naturally come out of Table I.

i) SURE is a reliable estimate of the MSE, since the resulting average loss in PSNR is within 0.02 dB for all images.
ii) Our sum of DOG (14) gives better PSNRs than the optimal *soft-threshold*.

IV. EFFICIENT SURE-BASED INTERSCALE THRESHOLDING

The integration of interscale information has been shown to improve the denoising quality, both visually and in terms of PSNR [9], [11], [21]. However, the gain brought is often modest, especially considering the additional complications involved by this processing [9]. In this section, we reformulate the problem by first building a loose prediction y_p of wavelet coefficients y out of a suitably filtered version of the lowpass subband at the same scale, and then by including this predictor in an explicit pointwise denoising function. Apart from the

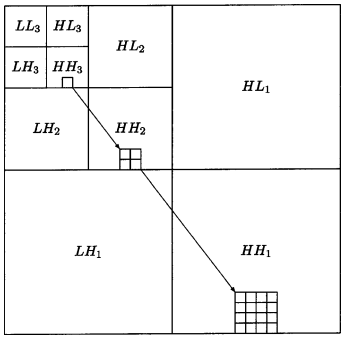


Fig. 6. Three stages of a fully decimated orthogonal wavelet transform and the so-called parent–child relationship.

specific denoising problem addressed in this paper, we believe more generally that other applications (e.g., compression, detection, segmentation) could benefit as well from the theory that leads to this predictor.

A. Building the Interscale Prediction

The wavelet coefficients that lie on the same dyadic tree (see Fig. 6) are well known to be large together in the neighborhood of image discontinuities. What can, thus, be predicted with reasonably good accuracy are the position of large wavelet coefficients out of parents at lower resolutions. However, getting the actual values of the finer resolution scale coefficients seem somewhat out of reach. This suggests that the best we can get out of between-scale correlations is a segmentation between regions of large and small coefficients. This comes back to the idea of signal of interest proposed by Pižurica *et al.* in [10].

In a critically sampled orthonormal wavelet decomposition, the parent subband is half the size of the child subband. The usual way of putting the two subbands in correspondence is simply to expand the parent by a factor two. Unfortunately, this approach does not take into account the potential—noninteger— shift caused by the filters of the DWT. We, thus, propose a more sophisticated solution, which addresses this issue and ensures the alignment of image features between the child and its parent.

Our idea comes from the following observation: Let LH_j and LL_j be, respectively, bandpass and lowpass outputs at iteration j of the filterbank. Then, if the *group delay*[8] between the bandpass and the lowpass filters are *equal*, no shift between the features of LH_j and LL_j will occur. Of course, depending on the amplitude response of the filters, some features may be attenuated, blurred, or enhanced, but their location will remain unchanged. When the group delays differ—which is the general case—we, thus, propose to filter the lowpass subband LL_j in order to *compensate for the group delay difference* with LH_j. This operation is depicted in Fig. 7(a): LL_j is filtered in the three bandpass "directions" by adequately designed filters W_{HL}, W_{HH}, and W_{LH}, providing aligned—i.e., group delay compensated—subbands with HL_j, HH_j, and LH_j.

Because the filters considered in this paper are separable, we only have to consider 1-D group delay compensation (GDC).

[8]For example, the frequency gradient of the phase response, with a minus sign.

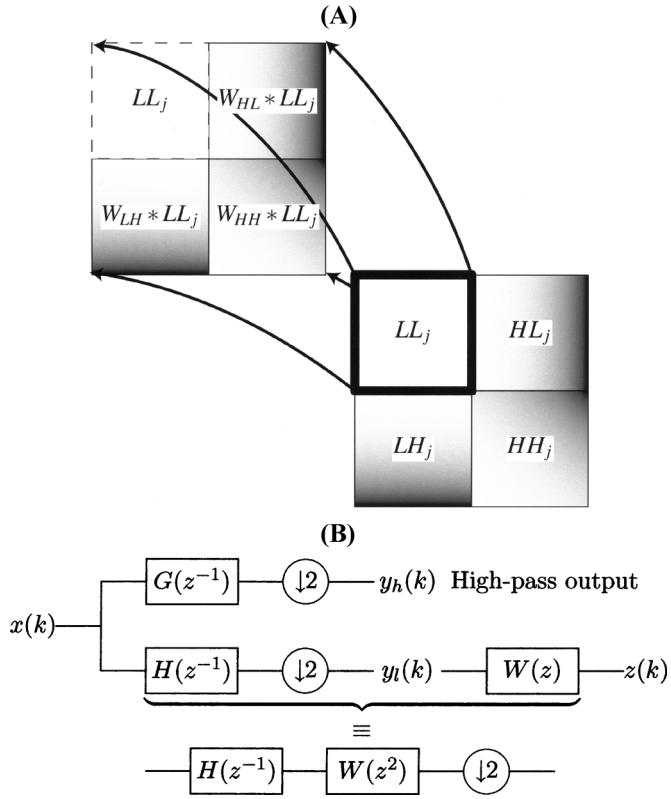


Fig. 7. One way of obtaining the whole parent information out of the lowpass subband: (a) 2-D illustration; (b) 1-D filterbank illustration.

Definition 1: We say that two filters $H(z)$ and $G(z)$ are group delay compensated if and only if the group delay of the quotient filter $H(z)/G(z)$ is zero identically; i.e., if and only if there exists a (anti-) symmetric filter $R(z) = \pm R(z^{-1})$ such that $H(z) = G(z)R(z)$.

The following result shows how to choose a GDC filter in a standard orthonormal filterbank.

Theorem 2: For the output of the dyadic orthonormal filterbank of Fig. 7(b) to be group delay compensated, it is necessary and sufficient that

$$W(z^2) = G(z^{-1})G(-z^{-1})(1 + \epsilon z^{-2})R(z^2) \tag{15}$$

where $\epsilon = \pm 1$ and $R(z) = R(z^{-1})$ is arbitrary.

Proof: Group delay compensation between the two filterbank branches is equivalent to [see Fig. 7(b)]

$$H(z^{-1})W(z^2) = G(z^{-1})R_1(z) \tag{16}$$

where $R_1(z) = \epsilon R_1(z^{-1})$ is an arbitrary symmetric $(\epsilon = 1)$ or anti-symmetric $(\epsilon = -1)$ filter.

Because the filters H and G are orthonormal, we have $H(z^{-1}) = zG(-z)$, and, thus, (16) can be rearranged as

$$W(z^2) = \frac{G(z^{-1})R_1(z)}{zG(-z)} = G(z^{-1})G(-z^{-1})\underbrace{\frac{z^{-1}R_1(z)}{G(-z)G(-z^{-1})}}_{R_2(z).} \tag{17}$$

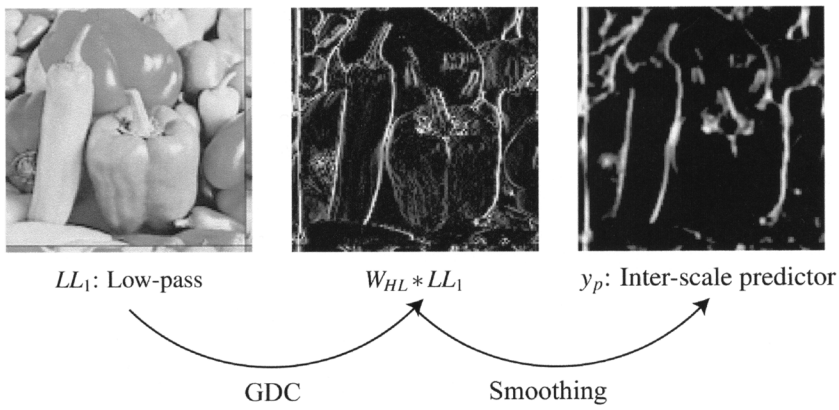


Fig. 8. Building an efficient interscale predictor, illustrated with a particular subband (HL_1) of the noise-free *Peppers* image.

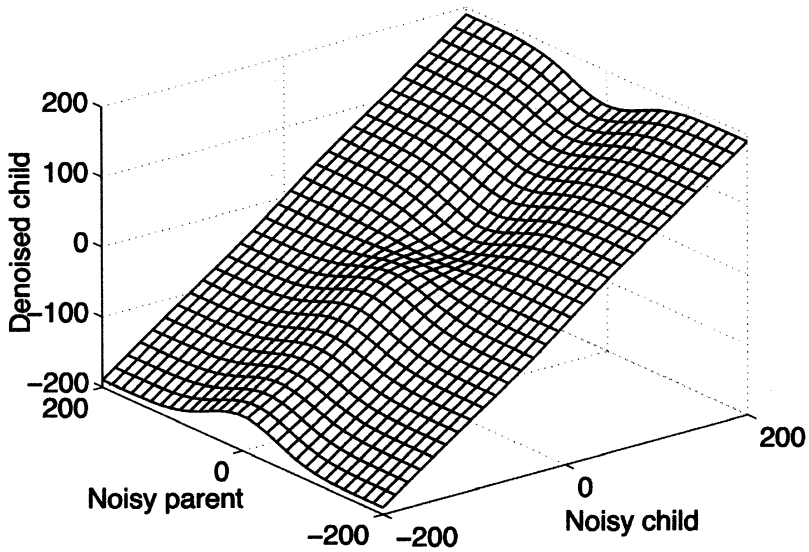


Fig. 9. Three-dimensional surface plot of a possible realization of our interscale thresholding function (21).

We observe that $R_2(z)$ is an even polynomial because both $G(z^{-1})G(-z^{-1})$ and $W(z^2)$ are. If we denote $R_2(z) = (1 + \epsilon z^{-2})R(z^2)$, then the symmetry of $R_1(z)$ implies that

$$\begin{aligned} R(z^{-2}) &= \frac{zR_1(z^{-1})}{(1+\epsilon z^2)G(-z)G(-z^{-1})} \\ &= \frac{\epsilon z R_1(z)}{(1+\epsilon z^2)G(-z)G(-z^{-1})} \\ &= \frac{z^{-1}R_1(z)}{(1+\epsilon z^{-2})G(-z)G(-z^{-1})} \\ &= R(z^2) \end{aligned}$$

i.e., $R(z)$ is and arbitrary zero-phase filter.

After substitution in (17), this finally leads us to the formulation (15), as an equivalent characterization of the group delay compensation in the filterbank of Fig. 7(b). ■

In addition to (15), the GDC filter $W(z)$ has to satisfy a few constraints:

- *energy preservation*, i.e., $\sum_{n\in\mathbb{Z}} w_n^2 = 1$, in order for the amplitude of the two outputs to be comparable;
- *highpass behavior*, in order for the filtered lowpass image to "look like" the bandpass target;
- *shortest possible response*, in order to minimize the enlargement of image features.

We can give a simple GDC filter in the case of symmetric filters. The shortest highpass $W(z)$ satisfying the GDC condition is in fact the simple gradient filter: $W(z) = z - 1$. If the symmetry is not centered at the origin but at a position n_0, then $W(z) = z^{-n_0}(z - 1)$. This type of solution is still adequate for near-symmetric filters such as the Daubechies *symlets* [22]. When the lowpass filter is not symmetric, we can simply take $R(z^2) = 1$ in (15).

Finally, in order to increase the homogeneity inside regions of similar magnitude coefficients, we apply a 2-D smoothing filter—a normalized Gaussian kernel $G(x) = (1/\sqrt{2\pi})e^{-(x^2/2)}$—onto the absolute value of the GDC output. In the rest of the paper, we will refer to the so-built interscale predictor by y_p.

B. Integrating the Interscale Predictor

Now that we have built the interscale predictor y_p, we have to suitably integrate it into our pointwise denoising function. As mentioned before, this interscale predictor does not tell us much about the actual value of its corresponding child wavelet

TABLE II
DENOISING PERFORMANCE IMPROVEMENT BROUGHT BY OUR INTERSCALE STRATEGY (NONREDUNDANT *SYM8*, FOUR ITERATIONS)

σ	5	10	20	30	50	100	5	10	20	30	50	100
Method	**Peppers** 256 × 256						**House** 256 × 256					
Expansion by 2	36.76	32.49	28.46	26.21	23.62	20.92	37.50	33.59	30.03	28.07	25.78	22.92
Proposed	**37.17**	**33.18**	**29.33**	**27.13**	**24.43**	**21.32**	**37.88**	**34.29**	**30.93**	**28.98**	**26.58**	**23.51**

Note: output PSNRs have been averaged over ten noise realizations.

TABLE III
COMPARISON OF SOME OF THE MOST EFFICIENT DENOISING METHODS (NONREDUNDANT *SYM8*, FOUR ITERATIONS)

σ	5	10	15	20	25	30	50	100	5	10	15	20	25	30	50	100
Input PSNR	34.15	28.13	24.61	22.11	20.17	18.59	14.15	8.13	34.15	28.13	24.61	22.11	20.17	18.59	14.15	8.13
Method	**Peppers** 256 × 256								**House** 256 × 256							
BayesShrink	35.83	31.49	29.30	27.85	26.72	25.73	23.17	20.73	36.91	32.92	30.81	29.42	28.44	27.66	25.49	22.87
BiShrink 7 × 7	36.61	32.55	30.25	28.66	27.47	26.51	23.89	20.80	37.54	33.60	31.56	30.16	29.07	28.20	25.83	22.84
ProbShrink 3 × 3	36.72	32.68	30.41	28.85	27.67	26.70	23.85	20.85	37.59	33.84	31.74	30.29	29.20	28.35	25.99	23.17
BLS-GSM 3 × 3	36.80	32.86	30.62	29.07	27.90	26.97	24.40	20.88	**38.01**	34.26	32.23	30.79	29.65	28.72	26.15	22.97
Our bivariate (21)	**37.17**	**33.18**	**30.91**	**29.33**	**28.12**	**27.13**	**24.43**	**21.32**	37.88	**34.29**	**32.32**	**30.93**	**29.86**	**28.98**	**26.58**	**23.51**
Best redundant	*37.09*	*33.33*	*31.26*	*29.84*	*28.74*	*27.84*	*25.30*	*21.98*	*38.43*	*35.04*	*33.23*	*31.91*	*30.87*	*30.01*	*27.62*	*24.53*
Method	**Al** 512 × 512								**Bridge** 256 × 256							
BayesShrink	37.77	34.17	32.10	30.67	29.63	28.84	26.67	23.84	34.81	29.80	27.30	25.75	24.69	23.90	22.04	19.99
BiShrink 7 × 7	38.01	34.50	32.57	31.23	30.21	29.39	27.09	24.01	34.94	29.93	27.38	25.81	24.75	23.97	22.11	19.97
ProbShrink 3 × 3	38.11	34.58	32.64	31.28	30.08	29.32	27.18	24.24	34.59	29.61	27.20	25.74	24.73	23.97	22.10	20.08
BLS-GSM 3 × 3	38.38	34.83	32.93	31.58	30.53	29.68	27.35	24.20	34.98	29.98	27.50	26.02	25.01	24.25	22.34	20.00
Our bivariate (21)	**38.43**	**34.90**	**32.97**	**31.64**	**30.64**	**29.84**	**27.61**	**24.56**	**35.06**	**30.22**	**27.84**	**26.36**	**25.33**	**24.56**	**22.60**	**20.35**
Best redundant	*38.90*	*35.46*	*33.66*	*32.42*	*31.46*	*30.67*	*28.46*	*25.51*	*35.23*	*30.46*	*28.07*	*26.60*	*25.58*	*24.83*	*22.98*	*20.78*
Method	**Barbara** 512 × 512								**Boat** 512 × 512							
BayesShrink	35.78	31.25	28.86	27.32	26.22	25.34	23.14	21.36	35.99	31.98	29.94	28.55	27.52	26.71	24.74	22.44
BiShrink 7 × 7	36.76	32.52	30.14	28.51	27.29	26.33	23.91	21.47	36.18	32.46	30.47	29.08	28.03	27.20	25.05	22.52
ProbShrink 3 × 3	36.75	32.48	30.04	28.40	27.20	26.27	23.86	21.58	36.20	32.53	30.50	29.11	28.05	27.22	25.12	22.69
BLS-GSM 3 × 3	**37.05**	**32.89**	**30.54**	**28.93**	**27.72**	**26.76**	**24.25**	21.53	36.46	32.89	**30.89**	**29.49**	28.43	27.58	25.34	22.64
Our bivariate (21)	36.71	32.18	29.66	27.98	26.76	25.83	23.70	**21.76**	**36.70**	**32.90**	30.85	29.47	**28.44**	**27.63**	**25.50**	**22.97**
Best redundant	*37.69*	*33.90*	*31.71*	*30.16*	*28.96*	*27.99*	*25.32*	*22.47*	*36.94*	*33.53*	*31.64*	*30.32*	*29.30*	*28.48*	*26.28*	*23.65*
Method	**Crowd** 512 × 512								**Goldhill** 512 × 512							
BayesShrink	34.60	29.31	26.53	24.73	23.45	22.47	20.07	17.46	35.93	31.94	29.96	28.69	27.79	27.13	25.41	23.32
BiShrink 7 × 7	34.71	29.48	26.70	24.88	23.57	22.57	20.13	17.40	36.17	32.27	30.32	29.07	28.15	27.44	25.57	23.26
ProbShrink 3 × 3	34.42	29.29	26.59	24.83	23.56	22.58	20.15	17.43	36.07	32.30	30.35	29.07	28.13	27.43	25.62	23.47
BLS-GSM 3 × 3	34.79	29.63	26.91	25.12	23.84	22.85	20.39	17.51	36.37	32.61	30.68	29.41	28.47	27.73	25.73	23.30
Our bivariate (21)	**34.86**	**29.77**	**27.11**	**25.38**	**24.13**	**23.17**	**20.75**	**17.97**	**36.53**	**32.69**	**30.76**	**29.52**	**28.60**	**27.89**	**26.06**	**23.82**
Best redundant	*34.96*	*30.05*	*27.49*	*25.83*	*24.63*	*23.69*	*21.24*	*18.33*	*36.88*	*33.24*	*31.37*	*30.13*	*29.22*	*28.50*	*26.60*	*24.30*

Note: output PSNRs have been averaged over ten noise realizations. The best redundant results are obtained using the *BLS-GSM* 3 × 3 with an 8-orientations full steerable pyramid; results slightly differ from the ones published in [9], because no boundary extension has been applied here.

coefficients. It only gives an indication on its expected magnitude. Here, we, thus, propose to use the parent y_p as a discriminator between high SNR wavelet coefficients and low SNR wavelet coefficients, leading to the following general pointwise denoising function:

$$\theta(y, y_p) = f(y_p) \sum_{k=1}^{K} a_k \varphi_k(y) + (1 - f(y_p)) \sum_{k=1}^{K} b_k \varphi_k(y). \tag{18}$$

The linear parameters a_k and b_k are then solved for by minimizing the MSE estimate ϵ defined in Theorem 1, for the linear parameters a_k and b_k. The optimal coefficients are obtained in the same way as in Section II-C and involve a solution similar to (11).

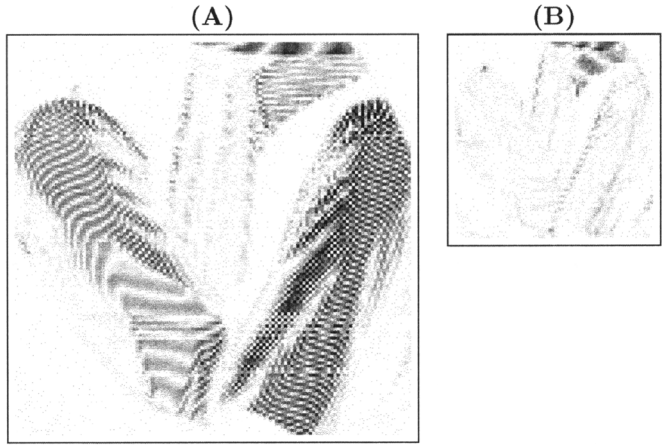


Fig. 10. (a) Zoom at Barbara's trousers at the finest scale of an orthonormal wavelet transform: the stripes are clearly visible. (b) Zoom at Barbara's trousers at the next coarser scale: the stripes are not visible anymore.

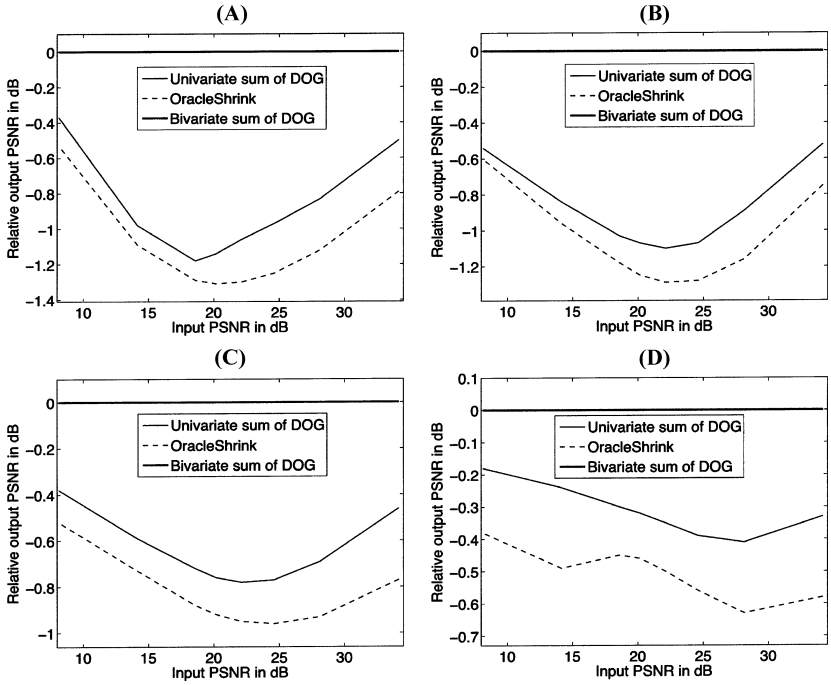


Fig. 11. Comparison of our interscale dependent thresholding function (21) with the best possible soft-threshold *OracleShrink* and with our simple univariate denoising function (14). (a) *Peppers* 256 × 256. (b) *House* 256 × 256. (c) *Lena* 512 × 512. (d) *Barbara* 512 × 512.

A first thought choice for the function f in (18) is simply the Heaviside function

$$H(y_p) = \begin{cases} 1, & \text{if } |y_p| \geq T \\ 0, & \text{if } |y_p| < T \end{cases} \tag{19}$$

where T can be interpreted as a decision factor. However, since the classification will not be perfect (i.e., some small parent coefficients may correspond to high-magnitude child coefficients, and *vice versa*), it is more appropriate to use a smoother decision function. Instead, we, thus, propose to use

$$f(y_p) = e^{-\frac{y_p^2}{2T^2}}. \tag{20}$$

As in the univariate case (Section III), we suggest to use a sum of DOG with $K = 2$ terms for each class of wavelet coefficients and[9] $T = \sqrt{6}\sigma$, leading to the following bivariate denoising function:

$$\begin{aligned} \theta(y, y_p; \mathbf{a}, \mathbf{b}) &= e^{-\frac{y_p^2}{12\sigma^2}} \theta_0(y; \mathbf{a}) + \left(1 - e^{-\frac{y_p^2}{12\sigma^2}}\right) \theta_0(y; \mathbf{b}) \\ &= e^{-\frac{y_p^2}{12\sigma^2}} \left(a_1 + a_2 e^{-\frac{y^2}{12\sigma^2}}\right) y \\ &\quad + \left(1 - e^{-\frac{y_p^2}{12\sigma^2}}\right) \left(b_1 + b_2 e^{-\frac{y^2}{12\sigma^2}}\right) y. \end{aligned} \tag{21}$$

Table II quantifies the improvement introduced by this new way of integrating the interscale information, as compared to the usual expansion of the parent subband.

[9]Side investigations have shown that the T needed in (20) and the one optimized in Section III can be chosen identical for optimal performances and equal to $\sqrt{6}\sigma$.

V. Experimental Results

In this section, we compare our interscale dependent thresholding function (21) with some of the best state-of-the-art techniques: Sendur's *et al.* bivariate MAP estimator with local variance estimation, Portilla's *BLS-GSM* and Pižurica's *ProbShrink*.

In all comparisons, we use a critically sampled orthonormal wavelet basis with eight vanishing moments (*sym8*) over four decomposition stages.

A. PSNR Comparisons

We have tested the various denoising methods for a representative set of standard 8-bit grayscale images such as *Al*, *Barbara*, *Boat*, *Crowd*, *Goldhill* (size 512 × 512) and *Peppers*, *House*, *Bridge* (size 256 × 256), corrupted by simulated additive Gaussian white noise at eight different power levels $\sigma \in [5, 10, 15, 20, 25, 30, 50, 100]$, which corresponds to PSNR decibel values [34.15, 28.13, 24.61, 22.11, 20.17, 18.59, 14.15, 8.13]. The denoising process has been performed over ten different noise realizations for each standard deviation and the resulting PSNRs averaged over these ten runs. The parameters of each method have been set according to the values given by their respective authors in the corresponding referred papers. Variations in output PSNRs are, thus, only due to the denoising techniques themselves. This reliable comparison was only possible thanks to the kindness of the various authors who have provided their respective Matlab codes on their personal websites.

Table III summarizes the results obtained. To the noteworthy exception of *Barbara*, our results are already competitive with the best techniques available that consider nonredundant orthonormal transforms. We stress again that our processing consists of a simple pointwise threshold, driven by interscale infor-

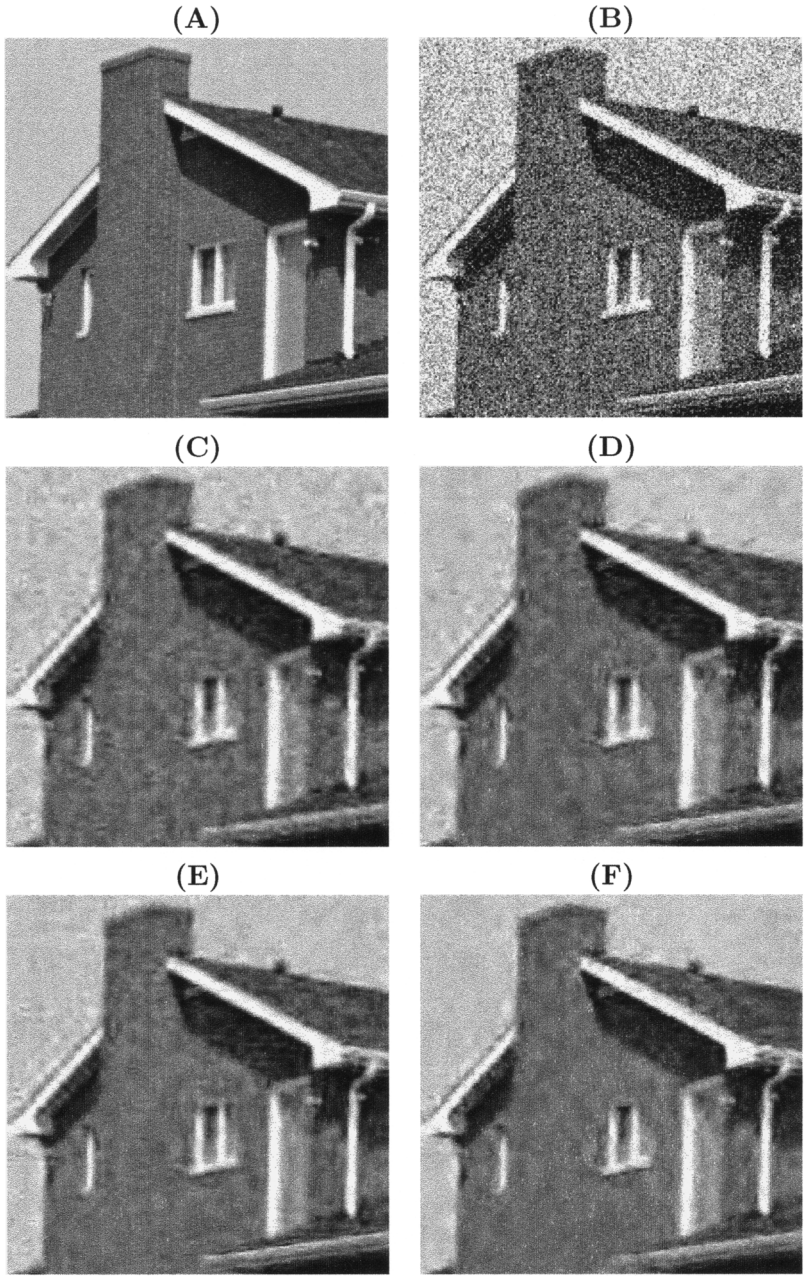


Fig. 12. (a) Part of the noise-free 256 × 256 *House* image. (b) Noisy version of it: PSNR = 18.59 dB. (c) Denoised result using the *BayesShrink*: PSNR = 27.57 dB. (d) Denoised result using the *BiShrink* 7 × 7: PSNR = 28.19 dB. (e) Denoised result using the *BLS-GSM* 3 × 3: PSNR = 28.73 dB. (f) Denoised result using our interscale dependent thresholding function (21): PSNR = 28.96 dB.

mation; i.e., without taking intrascale dependencies into consideration, contrary to the best performing methods (*ProbShrink*, *BiShrink* and *BLS-GSM*).

When looking closer at the results, we observe the following.

- Our method outperforms the classical *BayesShrink* by more than +1 dB on average.
- Our method gives better results than Sendur's *BiShrink* 7× 7 which integrates both the inter- and the intrascale dependencies (average gain of +0.6 dB).
- Our method gives better results than Pižurica's *ProbShrink* 3 × 3 which integrates the intrascale dependencies (average gain of +0.4 dB).
- We obtain similar or sometimes even better results than Portilla's *BLS-GSM* 3 × 3 for most of the images.
- For the *Barbara* image, our method is among the worst performers together with the pointwise *BayesShrink*. Our explanation for this is that some local information (especially the texture in Barbara's trousers) is completely lost at coarser scales (see Fig. 10). Interscale correlations may be too weak for this image, which indicates that an efficient denoising process may require intrascale information as well.
- The gap between our nonredundant SURE-based approach and the best up-to-date redundant results lies in the range of 0.5–1 dB for most images.

It is instructive to compare the results (see Fig. 11) obtained with our interscale dependent thresholding function (21), with the ones obtained with our simple univariate denoising function

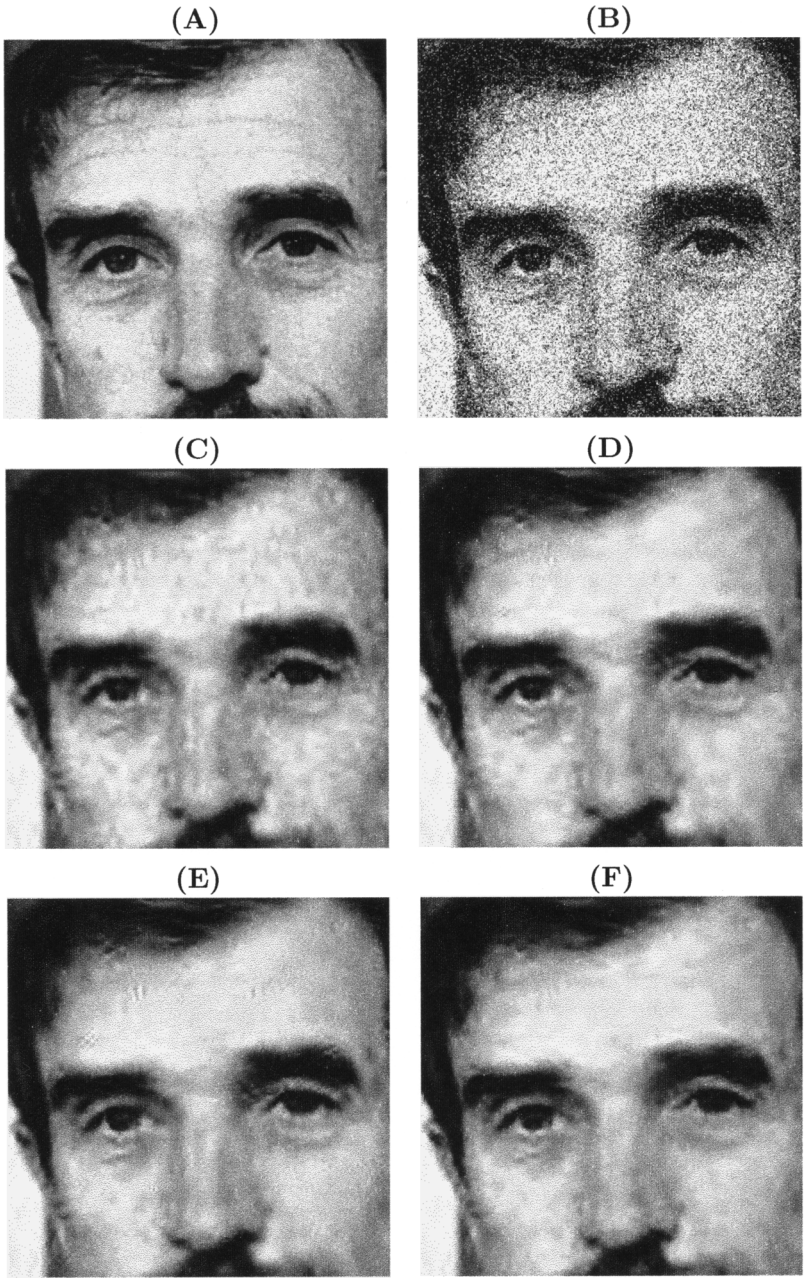


Fig. 13. (a) Part of the noise-free 512×512 *Al* image. (b) Noisy version of it: PSNR $= 14.15$ dB. (c) Denoised result using the *BayesShrink*: PSNR $= 26.71$ dB. (d) Denoised result using the *BiShrink* 7×7: PSNR $= 27.12$ dB. (e) Denoised result using the *BLS-GSM* 3×3: PSNR $= 27.34$ dB. (f) Denoised result using our interscale dependent thresholding function (21): PSNR $= 27.66$ dB.

(14). The improvement (often more than $+1$ dB) is quite significant for most standard images (see Fig. 11). Yet, for images that have substantial high-frequency contents, the integration of interscale dependencies does not lead to such an impressive gain. On the same graphs, we have also included the results obtained with the *OracleShrink*, showing a systematic underperformance with regards to even our simple univariate denoising function.

B. Visual Quality

Although there is no consensual objective way to judge the visual quality of a denoised image, two important criteria are widely used: the visibility of processing artifacts and the conservation of image edges. Processing artifacts usually result from a modification of the spatial correlation between wavelet coefficients (often caused by the zeroing of small neighboring coefficients) and are likely to be reduced by taking into account intrascale dependencies. Instead, image edge distortions usually arise from modifications of the interscale coefficient correlations. The amplitude of these modifications is likely to be reduced by a careful consideration of interscale dependencies in the denoising function.

Since our algorithm only includes interscale considerations, we expect it to be specifically robust to noise with regards to

TABLE IV
RELATIVE COMPUTATION TIME OF VARIOUS DENOISING TECHNIQUES

Method	Unit of time [U]	
	256×256 images	512×512 images
BayesShrink	1.0	3.9
BiShrink 7×7	1.4	5.4
ProbShrink 3×3	2.8	6.6
BLS-GSM 3×3	7.8	30.0
Univariate sum of DOG (14)	1.2	4.5
Bivariate sum of DOG (21)	2.0	7.9
Redundant *BLS-GSM* 3×3	*81.5*	*311.8*

Note: The computation times have been averagde over twenty runs.

edge preservation. Additionally, we would like to stress that our method exhibits the fewest number of artifacts, which we attribute to the fact that we are never forcing any wavelet coefficients to zero. These observations are illustrated in Figs. 12 and 13.

C. Computation Time

It is also interesting to evaluate the various denoising methods from a practical point of view: the computation time. Indeed, the results achieved by overcomplete representation are admittedly superior than the ones obtained by critically sampled wavelet transforms, but their weakness is the time they require (nearly 27 s on a Power Mac G5 workstation with 1.8-GHz PowerPC 970 CPU for 256×256 images to obtain the redundant results reported in Table III). With our simple univariate method (14), the whole denoising process (including four iterations of an orthonormal wavelet transform) lasts approximately 0.4 s for 256×256 images (1.6 s for 512×512 images), using a similar workstation. With our interscale dependent thresholding function (21), the whole denoising task takes between 0.6–0.7 s for 256×256 images and about 2.7 s for 512×512 images. To compare with, Portilla's BLS-GSM with a 3×3 window size lasts approximately 10 s for 512×512 images, using the same orthonormal transform. Besides giving competitive results, our method is, thus, also much faster.

Table IV summarizes the relative computation time of the various methods considered in this paper. Note that the main part of the *ProbShrink* is contained in a precompiled file, making its execution time a bit faster than the other algorithms which are fully implemented in Matlab.

VI. CONCLUSION

We have presented a new approach to orthonormal wavelet image denoising that does not need any prior statistical modelization of the wavelet coefficients. This approach is made possible thanks to the existence of an efficient estimate of the MSE between noisy and clean image—the SURE—that is based on the noisy data alone. Its minimization over a set of denoising processes automatically provides a near-optimal solution in the sense of the *a posteriori* MSE. For efficiency reasons, we have chosen this set to be a linear span of basic nonlinear mappings.

Using this approach, we have designed an image denoising algorithm that takes into account interscale dependencies, but discards intrascale correlations. In order to compensate for features misalignment, we have developed a rigorous procedure based on the relative group delay between the scaling and wavelet filters—*group delay compensation*. The information brought by this new interscale predictor is used to classify smoothly between high- and low-SNR wavelet coefficients.

The comparison of the denoising results obtained with our algorithm, and with the best state-of-the-art nonredundant techniques (that integrate both inter- and intrascale dependencies), demonstrate the efficiency of our SURE-based approach which gave the best output PSNRs for most of the images. The visual quality of our denoised images is moreover characterized by fewer artifacts than the other methods.

We are currently working on an efficient integration of the intrascale correlations within the SURE-based approach. Our goal is to show that the consideration of inter- and intrascale dependencies brings denoising gains that rival the quality of the best redundant techniques such as *BLS-GSM*.

REFERENCES

[1] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.
[2] ——, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[3] L. Breiman, "Better subset regression using the non-negative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, Nov. 1995.
[4] N. G. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Roy. Soc. A.*, Sep. 1999.
[5] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statist. Sin.*, vol. 7, pp. 855–874, 1997.
[6] ——, "Wavelet shrinkage denoising using the non-negative garrote," *J. Comput. Graph. Statist.*, vol. 7, no. 4, pp. 469–488, 1998.
[7] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, 1981.
[8] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1135–1151, Sep. 2000.
[9] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[10] A. Pižurica and W. Philips, "Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 645–665, Mar. 2006.
[11] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, Nov. 2002.
[12] ——, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.
[13] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *J. Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, May 2001.
[14] J.-L. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, Jun. 2002.
[15] M. S. Crouse, R. D. Nowak, and R. G. Baraniuki, "Wavelet-based signal processing using Hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[16] X.-P. Zhang and M. D. Desai, "Adaptive denoising based on SURE risk," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 265–267, Oct. 1998.
[17] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814–1830, Nov. 2005.

[18] P. L. Combettes and J.-C. Pesquet, "Wavelet-constrained image restoration," *Int. J. Wavelets, Multires. Inf. Process.*, vol. 2, no. 4, pp. 371–389, Dec. 2004.
[19] J.-C. Pesquet and D. Leporini, "A new wavelet estimator for image denoising," in *Proc. 6th Int. Conf. Image Processing and its Applications*, Jul. 14–17, 1997, vol. 1, pp. 249–253.
[20] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, 1961, vol. 1, pp. 361–379.
[21] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.
[22] I. Daubechies, "Ten lectures on wavelets," presented at the CBMS-NSF Regional Conf. Ser. Applied Mathematics, Mar. 1992.
[23] M. K. Mihçak, Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, Dec. 1999.
[24] F. Abramovitch, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy Statist. Soc. B*, vol. 60, no. 4, pp. 725–749, 1998.
[25] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 3, pp. 165–168, Mar. 1980.
[26] E. P. Simoncelli, *Bayesian Interference in Wavelet Based Models*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, Mar. 1999, vol. 141, ch. 18, pp. 291–308.
[27] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: Wiley-Interscience, Apr. 1999.

Florian Luisier was born in Switzerland in 1981. In 2005, he received the M.S. degree in microengineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Swizerland. He is currently pursuing the Ph.D. degree within the Biomedical Imaging Group (BIG), EPFL.

His research interests mainly include multiresolution analysis and the restoration of biomedical images.

Thierry Blu (M'96–SM'06) was born in Orléans, France, in 1964. He received the "Diplôme d'ingénieur" from the École Polytechnique, France, in 1986 and from Télécom Paris (ENST), France, in 1988, and the Ph.D. degree in electrical engineering from ENST in 1996 for a study on iterated rational filterbanks, applied to wideband audio coding.

He is with the Biomedical Imaging Group, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, on leave from the France Télécom National Center for Telecommunications Studies (CNET), Issy-les-Moulineaux, France. At EPFL, he teaches the theory of Signals and Systems for Microengineering and Life Science students. His research interests include (multi)wavelets, multiresolution analysis, multirate filterbanks, interpolation, approximation and sampling theory, image denoising, psychoacoustics, optics, wave propagation, etc.

Dr. Blu is the recipient of the 2003 best paper award (SP Theory and Methods) from the IEEE Signal Processing Society. He is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and, since 2006, for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Michael Unser (M'89–SM'94–F'99) received the M.S. (summa cum laude) and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1981 and 1984, respectively.

From 1985 to 1997, he was a Scientist with the National Institutes of Health, Bethesda, MD. He is now a Professor and Director of the Biomedical Imaging Group, EPFL. His main research topics are biomedical image processing, splines, and wavelets. He is the author of over 120 published journal papers in these areas.

Dr. Unser has been actively involved with the IEEE TRANSACTIONS ON MEDICAL IMAGING, holding the positions of Associate Editor (1999–2002 and 2006–present), member of steering committee, and Associate Editor-in-Chief (2003–2005). He has acted as an Associate Editor or member of the editorial board for eight more international journals, including the *IEEE Signal Processing Magazine*, the IEEE TRANSACTIONS ON IMAGE PROCESSING (1992–1995), and the IEEE SIGNAL PROCESSING LETTERS (1994–1998). He organized the first IEEE International Symposium on Biomedical Imaging (ISBI 2002). He currently chairs the technical committee of the IEEE-SP Society on Bio Imaging and Signal Processing (BISP), and well as the ISBI steering committee. He is a recipient of three Best Paper Awards from the IEEE Signal Processing Society.

The SURE-LET Approach to Image Denoising

Thierry Blu, *Senior Member, IEEE*, and Florian Luisier

***Abstract*—We propose a new approach to image denoising, based on the *image-domain minimization* of an estimate of the mean squared error—*Stein's unbiased risk estimate* (SURE). Unlike most existing denoising algorithms, using the SURE makes it needless to hypothesize a statistical model for the noiseless image. A key point of our approach is that, although the (nonlinear) processing is performed in a transformed domain—typically, an undecimated discrete wavelet transform, but we also address nonorthonormal transforms—this minimization is performed in the image domain. Indeed, we demonstrate that, when the transform is a "tight" frame (an undecimated wavelet transform using orthonormal filters), separate subband minimization yields substantially worse results. In order for our approach to be viable, we add another principle, that the denoising process can be expressed as a linear combination of elementary denoising processes—*linear expansion of thresholds* (LET). Armed with the SURE and LET principles, we show that a denoising algorithm merely amounts to solving a *linear system of equations* which is obviously fast and efficient. Quite remarkably, the very competitive results obtained by performing a simple threshold (image-domain SURE optimized) on the undecimated Haar wavelet coefficients show that the SURE-LET principle has a huge potential.**

I. Introduction

DURING acquisition and transmission, images are often corrupted by additive noise. The main aim of an image denoising algorithm is then to reduce the noise level, while preserving the image features.

Transform domain image denoising—the most popular approaches to process noisy images consist in first applying some linear—often multiscale—transformation, then performing a usually nonlinear—and sometimes multivariate—operation on the transformed coefficients, and finally reverting to the image domain by applying an inverse linear transformation. Among the many denoising algorithms to date, we would like to cite the following ones.

- *Portilla et al.* [1]:[1] The authors' main idea is to model the neighborhoods of coefficients at adjacent positions and scales as a Gaussian scale mixture (GSM); the wavelet estimator is then a Bayes least squares (BLS). The resulting denoising method, consequently called *BLS-GSM*, is the most efficient up-to-date approach in terms of peak signal-to-noise ratio (PSNR).
- *Pižurica et al.* [2]:[2] Assuming a generalized Laplacian prior for the noise-free data, the authors' approach called *ProbShrink* is driven by the estimation of the probability that a given coefficient contains significant information—notion of "signal of interest".
- *Sendur et al.* [3], [4]:[3] The authors' method, called *BiShrink*, is based on new non-Gaussian bivariate distributions to model interscale dependencies. A nonlinear bivariate shrinkage function using the maximum *a posteriori* (MAP) estimator is then derived. In a second paper, these authors have extended their approach by taking into account the intrascale variability of wavelet coefficients.

These techniques have been devised for both redundant and nonredundant transforms.

While the choice of the transformation is easily justified by well-accepted general considerations—e.g., closeness to the Karhunen–Loève transformation, "sparsity" of the transformed coefficients, "steerability" of the transformation—the nonlinear operation that follows is more frequently based on *ad hoc* statistical hypotheses on the transformed coefficients that are specific to each author. The final performance of the algorithms—typically, PSNR results—is, thus, inconclusively related to the accuracy of this modelization.

SURE-LET denoising—In this paper, we want to promote quite a different point of view, which avoids any *a priori* hypotheses on the noiseless image—in particular, *no* random process modelization—but for the usual white Gaussian noise assumption. This approach is made possible by the existence of an excellent unbiased estimate of the mean squared error (MSE) between the noiseless image and its denoised version—Stein's unbiased risk estimate (*SURE*). If we evaluate denoising performances by comparing PSNRs, then this MSE is precisely the quantity that we want to minimize. Similar to the MSE, the SURE takes the form of a quadratic expression in terms of the denoised image (see Theorem1).

Our approach, thus, consists in reformulating the denoising problem as the search for the denoising process that will minimize the SURE—in the image domain. In practice, the process is completely characterized by a set of parameters. Now, to take full advantage of the *quadratic* nature of the SURE, we choose to consider only denoising processes that can be expressed as a *linear combination* of "elementary" denoising processes—linear expansion of thresholds (*LET*). This "SURE-LET" stategy is computationally very efficient because minimizing the SURE for the unknown weights gives rise to a mere linear system of equations, which in turn allows to

Manuscript received March 20, 2007; revised July 13, 2007. This work was supported in part by the Center for Biomedical Imaging (CIBM) of the Geneva-Lausanne Universities and the EPFL, in part by the foundations Leenaards and Louis-Jeantet, and in part by the Swiss National Science Foundation under Grant 200020-109415. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pier Luigi Dragotti.

The authors are with the Biomedical Imaging Group (BIG), Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: thierry.blu@epfl.ch; florian.luisier@epfl.ch).

Digital Object Identifier 10.1109/TIP.2007.906002

[1]Available at http://www.io.csic.es/PagsPers/JPortilla/denoise/software/index.htm.

[2]Available at http://telin.ugent.be/~sanja/.

[3]Available at http://taco.poly.edu/WaveletSoftware/denoise2.html.

consider processes described by quite a few parameters. There is, however, a tradeoff between the sharpness of the description of the process which increases with the number of parameters, and the predictability of the MSE estimate, which is inversely related to the number of parameters. We have already applied our approach within a nonredundant, orthonormal wavelet framework, and showed that a simple thresholding function that takes interscale dependences into account is very efficient, both in terms of computation time and image denoising quality[4] [5].

SURE-related literature—Despite its simple MSE justification (a mere integration by parts), the SURE does not belong to the toolbox of the standard signal processing practitioner—although it is, of course, much better established among statisticians. The best-known use of the SURE in image denoising is Donoho's SureShrink algorithm [6] in which a soft-threshold is applied to the *orthonormal* wavelet coefficients, and where the threshold parameter is optimized separately in each subband through the minimization of the SURE. Otherwise, the approach that is most closely related to SURE-LET—but for a multichannel image denoising application—is the contribution by Pesquet and his collaborators [7]–[9] which perform *separate in-band* minimization of the SURE applied to a denoising process that contains both nonlinear and linear parameters.

Yet, the specificity of SURE-LET for redundant or nonorthonormal transforms lies in the fact that this minimization is performed *in the image domain*. While it is true that, due to some Parseval-like MSE conservation, image domain MSE/SURE minimization is equivalent to separate in-band MSE/SURE minimization whenever the analysis transformation is—nonredundant—orthonormal [5], this is grossly wrong as soon as the transformation is, either *redundant* (even when it is a "tight frame") or *nonorthonormal*. This is actually the observation made by those who apply soft-thresholding to an undecimated wavelet transform: the SureShrink threshold determination yields substantially worse results than an empirical choice (see Fig. 3). Unfortunately, this may lead practitioners to wrongly conclude that the SURE approach is unsuitable for redundant transforms, whereas a correct diagnosis should be that it is the independent subband approach that is flawed.

Organization of the paper—In Section II, we expose the multivariate SURE theory for vector functions, and sketch the principles of our linear parametrization strategy; we also address practical issues like how the SURE is modified depending on the choice for boundary conditions, and provide explicit SURE formulæ for pointwise thresholding. In Section III, because we want to exemplify the power of the SURE-LET approach, we restrict the processing to simple pointwise thresholds in the transformed domain and show that, by using an undecimated Haar wavelet transform, a SURE image-domain minimization yields very competitive results with the best up-to-date algorithms [1], [2], [4] (Section IV-C). In comparison, without any optimization attempts in our implementation, the SURE-LET method is quite CPU-time friendly. Yet, a huge margin of improvement can be envisioned if intrascale and interscale dependencies are taken into account. Both the competitiveness and robustness of our method validate our new approach as an attractive solution for image denoising.

[4]See our demo http://bigwww.epfl.ch/demo/suredenoising/.

II. Theoretical Background

We consider the standard simplified denoising problem: given noisy data $y_n = x_n + b_n$, for $n = 1 \dots N$ where b_n is a white Gaussian noise of variance σ^2, find a reasonably good estimate $\hat{\mathbf{x}}$ of $\mathbf{x} = \{x_n\}_{n=1,2,\dots N}$. Our goal is, thus, to find a function of the noisy data alone $\mathbf{F}(\mathbf{y}) = (f_n(\mathbf{y}))_{n=1,2,\dots N} = \hat{\mathbf{x}}$ which will minimize the MSE defined by

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N}|\hat{x}_n - x_n|^2$$
$$\Updownarrow$$
$$\text{MSE} = \frac{1}{N}\|\hat{\mathbf{x}} - \mathbf{x}\|^2. \tag{1}$$

A. Unbiased Estimate of the MSE

Since we do not have access to the original signal $\mathbf{x}$, we cannot compute $\|\hat{\mathbf{x}} - \mathbf{x}\|^2/N$—the *Oracle* MSE. However, without any assumptions on the noise-free data, we will see that it is possible to replace this quantity by an unbiased estimate which is a function of $\mathbf{y}$ only. This has an important consequence: contrary to what is frequently done in the literature, the noise-free signal is not modeled as a random process in our framework—we do not even require $\mathbf{x}$ to belong to a specific class of signals. Thus, the observed randomness of the noisy data originates only from the Gaussian white noise $\mathbf{b}$.

The following lemma which states a version of Stein's lemma [10], shows how it is possible to replace an expression that contains the unknown data $\mathbf{x}$ by another one with the same expectation, but containing the known data $\mathbf{y}$ only.

Lemma 1: Let $\mathbf{F}(\mathbf{y})$ be an N-dimensional vector function such that $\mathcal{E}\{|\partial f_n(\mathbf{y})/\partial y_n|\} < \infty$ for $n = 1, \dots, N$. Then, under the additive white Gaussian noise assumption, the expressions $\mathbf{F}(\mathbf{y})^{\mathrm{T}}\mathbf{x}$ and $\mathbf{F}(\mathbf{y})^{\mathrm{T}}\mathbf{y} - \sigma^2\text{div}\{\mathbf{F}(\mathbf{y})\}$ have the same expectation

$$\mathcal{E}\left\{\sum_{n=1}^{N} f_n(\mathbf{y})x_n\right\} = \mathcal{E}\left\{\sum_{n=1}^{N} f_n(\mathbf{y})y_n\right\} - \sigma^2\mathcal{E}\Bigg\{\underbrace{\sum_{n=1}^{N}\frac{\partial f_n(\mathbf{y})}{\partial y_n}}_{\text{div}\{\mathbf{F}(\mathbf{y})\}}\Bigg\} \tag{2}$$

where $\mathcal{E}\{\cdot\}$ stands for the mathematical expectation operator.

Proof: We use the fact that a Gaussian white probability density $q(b_n)$ satisfies $b_n q(b_n) = -\sigma^2 q'(b_n)$. Thus, denoting by $\mathcal{E}_{b_n}\{\cdot\}$ the partial expectation over the nth component of the noise, we have the following sequence of equalities:[5]

$$\begin{aligned}
\mathcal{E}_{b_n}\{f_n(\mathbf{y})x_n\} &= \mathcal{E}_{b_n}\{f_n(\mathbf{y})y_n\} - \mathcal{E}_{b_n}\{f_n(\mathbf{y})b_n\} \\
&= \mathcal{E}_{b_n}\{f_n(\mathbf{y})y_n\} - \int f_n(\mathbf{y})b_n q(b_n)db_n \\
&= \mathcal{E}_{b_n}\{f_n(\mathbf{y})y_n\} + \sigma^2\int f_n(\mathbf{y})q'(b_n)db_n \\
&= \mathcal{E}_{b_n}\{f_n(\mathbf{y})y_n\} - \sigma^2\int \frac{\partial f_n(\mathbf{y})}{\partial y_n}q(b_n)db_n \\
&\quad \text{(by parts)} \\
&= \mathcal{E}_{b_n}\{f_n(\mathbf{y})y_n\} - \sigma^2\mathcal{E}_{b_n}\left\{\frac{\partial f_n(\mathbf{y})}{\partial y_n}\right\}.
\end{aligned}$$

[5]To be fully rigorous, we need to assume that $f_n(\mathbf{y})q(y_n - x_n)$ tends to zero with $|y_n|$, which is very broadly ensured whenever $f_n(\mathbf{y})$ is bounded by some *fastly increasing* function, like $\exp(\|\mathbf{y}\|^2/2\sigma'^2)$ where $\sigma' > \sigma$.

TABLE I
COMPARISON OF SOME OF THE MOST EFFICIENT DENOISING METHODS

σ	5	10	15	20	25	30	50	100	5	10	15	20	25	30	50	100
Input PSNR	34.15	28.13	24.61	22.11	20.17	18.59	14.15	8.13	34.15	28.13	24.61	22.11	20.17	18.59	14.15	8.13
Method	**Peppers** 256×256								**House** 256×256							
BiShrink [4]	37.18	33.38	31.28	29.80	28.67	27.76	25.28	22.11	38.35	34.71	32.89	31.63	30.64	29.83	27.54	24.51
ProbShrink [2]	37.45	33.75	31.71	30.25	29.15	28.24	25.72	22.48	38.51	35.15	33.43	32.19	31.21	30.38	27.98	24.76
BLS-GSM [1]	37.32	33.77	31.74	30.31	29.21	28.33	25.90	**22.67**	38.67	35.34	33.60	32.35	31.35	30.52	28.21	25.09
UWT SURE-LET	**37.63**	**34.00**	**31.97**	**30.53**	**29.40**	**28.48**	**25.94**	22.60	**38.71**	**35.52**	**33.81**	**32.60**	**31.66**	**30.89**	**28.58**	**25.25**
UWT Oracle	*37.64*	*34.01*	*31.98*	*30.55*	*29.43*	*28.51*	*26.00*	*22.72*	*38.71*	*35.53*	*33.83*	*32.63*	*31.70*	*30.94*	*28.68*	*25.44*
Method	**Al** 512×512								**Bridge** 256×256							
BiShrink [4]	38.72	35.34	33.51	32.24	31.26	30.45	28.15	24.99	35.18	30.47	28.11	26.62	25.56	24.77	22.88	20.73
ProbShrink [2]	38.67	35.42	33.68	32.45	31.51	30.66	28.46	25.40	35.08	30.36	27.97	26.52	25.52	24.79	23.03	20.93
BLS-GSM [1]	**38.98**	**35.57**	**33.81**	**32.60**	**31.67**	**30.91**	**28.73**	**25.75**	**35.26**	30.49	28.11	26.65	25.66	24.92	23.11	20.98
UWT SURE-LET	38.88	35.43	33.60	32.36	31.42	30.66	28.57	25.69	35.23	**30.54**	**28.24**	**26.82**	**25.83**	**25.10**	**23.27**	**21.09**
UWT Oracle	*38.88*	*35.43*	*33.61*	*32.37*	*31.43*	*30.67*	*28.59*	*25.76*	*35.23*	*30.55*	*28.25*	*26.83*	*25.84*	*25.11*	*23.30*	*21.16*
Method	**Barbara** 512×512								**Boat** 512×512							
BiShrink [4]	37.35	33.51	31.37	29.87	28.72	27.79	25.30	22.46	36.72	33.17	31.30	29.98	28.96	28.14	25.97	23.31
ProbShrink [2]	37.39	33.49	31.24	29.60	28.33	27.30	24.54	22.02	36.69	33.29	31.34	29.97	28.91	28.06	25.83	23.17
BLS-GSM [1]	**37.79**	**34.02**	**31.84**	**30.29**	**29.10**	**28.12**	**25.44**	**22.59**	36.98	**33.58**	**31.70**	**30.37**	**29.35**	**28.54**	**26.35**	**23.70**
UWT SURE-LET	36.98	32.65	30.16	28.45	27.18	26.23	24.13	22.26	**37.13**	33.53	31.57	30.22	29.20	28.39	26.20	23.61
UWT Oracle	*36.98*	*32.65*	*30.16*	*28.45*	*27.19*	*26.24*	*24.14*	*22.29*	*37.13*	*33.54*	*31.58*	*30.23*	*29.21*	*28.40*	*26.22*	*23.65*
Method	**Crowd** 512×512								**Goldhill** 512×512							
BiShrink [4]	34.86	29.85	27.28	25.61	24.40	23.47	21.05	18.18	36.78	33.11	31.23	29.99	29.08	28.37	26.52	24.19
ProbShrink [2]	34.79	30.00	27.47	25.80	24.57	23.61	21.14	18.24	36.76	33.20	31.33	30.12	29.22	28.53	26.71	24.51
BLS-GSM [1]	34.97	30.07	27.52	25.87	24.67	23.73	21.29	18.37	**36.98**	**33.36**	**31.50**	**30.28**	**29.39**	**28.69**	**26.85**	24.61
UWT SURE-LET	**35.10**	**30.20**	**27.64**	**25.96**	**24.74**	**23.78**	**21.32**	**18.43**	36.85	33.20	31.37	30.17	29.30	28.61	26.83	**24.69**
UWT Oracle	*35.10*	*30.20*	*27.64*	*25.96*	*24.74*	*23.78*	*21.33*	*18.45*	*36.85*	*33.21*	*31.37*	*30.18*	*29.30*	*28.62*	*26.85*	*24.75*

Note: Output PSNRs have been averaged over eight noise realizations.

Now, taking the expectation over the remaining components of the noise, we get

$$\mathcal{E}\left\{f_n(\mathbf{y})x_n\right\} = \mathcal{E}\left\{f_n(\mathbf{y})y_n\right\} - \sigma^2\mathcal{E}\left\{\frac{\partial f_n(\mathbf{y})}{\partial y_n}\right\}.$$

Since the expectation is a linear operator, (2) follows directly.■

By applying Lemma 1 to the expression of the MSE, we then get Stein's unbiased risk—or MSE—estimate (SURE).

Theorem 1: Under the same hypotheses as Lemma 1, the random variable

$$\epsilon = \frac{1}{N}\left\|\mathbf{F}(\mathbf{y}) - \mathbf{y}\right\|^2 + \frac{2\sigma^2}{N}\mathrm{div}\left\{\mathbf{F}(\mathbf{y})\right\} - \sigma^2 \tag{3}$$

is an unbiased estimator of the MSE, i.e.,

$$\mathcal{E}\{\epsilon\} = \frac{1}{N}\mathcal{E}\left\{\left\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\right\|^2\right\}.$$

Proof: By expanding the expectation of the MSE, we have

$$\begin{aligned}\mathcal{E}\left\{\left\|\mathbf{F}(\mathbf{y}) - \mathbf{x}\right\|^2\right\} &= \mathcal{E}\left\{\left\|\mathbf{F}(\mathbf{y})\right\|^2\right\} - 2\mathcal{E}\left\{\mathbf{F}(\mathbf{y})^{\mathrm{T}}\mathbf{x}\right\} \\ &\quad + \mathcal{E}\left\{\left\|\mathbf{x}\right\|^2\right\} \\ &= \mathcal{E}\left\{\left\|\mathbf{F}(\mathbf{y})\right\|^2\right\} - 2\mathcal{E}\left\{\mathbf{F}(\mathbf{y})^{\mathrm{T}}\mathbf{y}\right\} \\ &\quad + 2\sigma^2\mathcal{E}\left\{\mathrm{div}\left\{\mathbf{F}(\mathbf{y})\right\}\right\} + \mathcal{E}\left\{\left\|\mathbf{x}\right\|^2\right\}\end{aligned}$$

where we have applied Lemma 1. Since the noise $\mathbf{b}$ has zero mean, we can replace $\mathcal{E}\{\|\mathbf{x}\|^2\}$ by $\mathcal{E}\{\|\mathbf{y}\|^2\} - N\sigma^2$. A rearrangement of the $\mathbf{y}$ terms then provides the result of Theorem 1. ■

We want to emphasize here the fact that in image denoising applications the number of samples is usually large—typically 256^2—and, thus, the estimate ϵ has a small variance—typically $\propto 1/N$. This estimate is, thus, close to its expectation, which is indeed the true MSE of the denoising process.

B. SURE-LET Approach

Our general denoising strategy consists in expressing the denoising process, $\mathbf{F}(\mathbf{y})$, as a linear combination (LET: linear expansion of thresholds) of given elementary processes, $\mathbf{F}_k(\mathbf{y})$

$$\mathbf{F}(\mathbf{y}) = \sum_{k=1}^{K} a_k\mathbf{F}_k(\mathbf{y}). \tag{4}$$

Here, the unknown weights a_k are specified by minimizing the SURE given by (3). It is also possible, in order to evaluate the performance of the algorithm, to compare the result with what the minimization of the MSE would provide—i.e., the *Oracle* optimization (see Table I). A limitation of the LET approach is that the elementary denoising functions $\mathbf{F}_k$ have to fulfill the hypothesis of Lemma 1 (differentiability); moreover, the number of parameters K must not be "too large" compared to the number of pixels (typically, less than 100 for usual 256×256 images), in order for the variance of the SURE to remain small.

The linearity of the expansion (4) is a crucial advantage for solving the minimization problem, because the SURE is quadratic in $\mathbf{F}(\mathbf{y})$. The coefficients a_k are, thus, the solution of a linear system of equations

$$\begin{aligned}\sum_{l=1}^{K}\underbrace{\mathbf{F}_k(\mathbf{y})^{\mathrm{T}}\mathbf{F}_l(\mathbf{y})}_{[\mathbf{M}]_{k,l}}a_l &= \underbrace{\mathbf{F}_k(\mathbf{y})^{\mathrm{T}}\mathbf{y} - \sigma^2\mathrm{div}\left\{\mathbf{F}_k(\mathbf{y})\right\}}_{[\mathbf{c}]_k} \\ &\quad \text{for } k = 1, 2, \ldots K \\ &\Updownarrow \\ \mathbf{Ma} &= \mathbf{c}.\end{aligned} \tag{5}$$

Note that, since the minimum of ϵ always exists, we are ensured that there will always be a solution to this system. When

$\text{rank}(\mathbf{M}) < K$, the function $\mathbf{F}$ is *over-parameterized* and consequently, several sets of parameters a_k yield equivalent results; in that case, we may simply consider the solution provided by the pseudoinverse of $\mathbf{M}$. Of course, it is also possible to reduce the parametrization order K so as to make the matrix $\mathbf{M}$ full rank—at no quality loss.

What this approach suggests is that the practitioner may choose at will (restricted only by the differentiability constraint of Theorem 1) a set of K different denoising algorithms—ideally with complementary denoising behaviors—and optimize a weighting of these algorithms to get the best of them at once.

Among the potentially interesting algorithms are those that work in a transformed domain such as:

- the nonredundant wavelet transforms, either orthogonal or bi-orthogonal [11];
- the classical undecimated wavelet transform [12];
- the *curvelet* [13] transform;
- the *contourlet* [14] transform;
- the steerable pyramids [15], [16];

as well as the discrete cosine transform (*DCT*) or its overcomplete variant: the block discrete cosine transform (*B-DCT*). In the remainder of this paper, we will consider only pointwise thresholding in such transform domains.

C. Pointwise SURE-LET Transform Denoising

Transform domain denoising consists in first defining a couple of linear transformations $\mathcal{D}$—decomposition—and $\mathcal{R}$—reconstruction—such that $\mathcal{R}\mathcal{D} = \text{Identity}$: typically, $\mathcal{D}$ is a bank of decimated or undecimated filters. Once the size of the input and output data are frozen, these linear operators are characterized by matrices, respectively $\mathbf{D} = (d_{i,j})_{(i,j)\in[1;L]\times[1;N]}$ and $\mathbf{R} = (r_{i,j})_{(i,j)\in[1;N]\times[1;L]}$ that satisfy the *perfect reconstruction* property $\mathbf{RD} = \mathbf{Id}$. Then, the whole denoising process boils down to the following steps.

1) Apply $\mathbf{D}$ to the noisy signal $\mathbf{y} = \mathbf{x} + \mathbf{b}$ to get the transformed noisy coefficients $\mathbf{w} = \mathbf{Dy} = (w_i)_{i\in[1;L]}$.
2) Apply a *pointwise* thresholding function $\Theta(\mathbf{w}) = (\theta_i(w_i))_{i\in[1;L]}$.
3) Revert to the original domain by applying $\mathbf{R}$ to the thresholded coefficients $\Theta(\mathbf{w})$, yielding the denoised estimate $\hat{\mathbf{x}} = \mathbf{R}\Theta(\mathbf{w})$.

This algorithm can be summarized as a function of the noisy input coefficients

$$\hat{\mathbf{x}} = \mathbf{F}(\mathbf{y}) = \mathbf{R}\Theta(\mathbf{Dy}). \tag{6}$$

The SURE-LET approach suggests to express $\mathbf{F}$ as a *linear expansion* of denoising algorithms $\mathbf{F}_k$, according to

$$\mathbf{F}(\mathbf{y}) = \sum_{k=1}^{K} a_k \underbrace{\mathbf{R}\Theta_k(\mathbf{Dy})}_{\mathbf{F}_k(\mathbf{y})} \tag{7}$$

where $\Theta_k(\cdot)$ are elementary pointwise thresholding functions.

As we have noticed in the previous subsection [see (5)], retrieving the parameters a_k boils down to the resolution of a linear system of equations. Note that this linear parametrization does not imply a linear denoising; indeed, the thresholding functions Θ_k can be chosen nonlinear.

In the SURE-LET framework, Theorem 1 can be reformulated in the following way.

Corollary 1: Let $\mathbf{F}$ be defined according to (6) where Θ denotes *pointwise* thresholding. Then the MSE between the original and the denoised signal can be *unbiasedly* estimated by the following random variable:

$$\epsilon = \frac{1}{N}\left\|\mathbf{F}(\mathbf{y}) - \mathbf{y}\right\|^2 + \frac{2\sigma^2}{N}\boldsymbol{\alpha}^{\mathrm{T}}\Theta'(\mathbf{Dy}) - \sigma^2 \tag{8}$$

where

- $\boldsymbol{\alpha} = \text{diag}\{\mathbf{DR}\} = \{[\mathbf{DR}]_{1,1}, [\mathbf{DR}]_{2,2}, \ldots, [\mathbf{DR}]_{L,L}\}$ is a vector made of the diagonal elements of the matrix $\mathbf{DR}$;
- $\Theta'(\mathbf{Dy}) = \Theta'(\mathbf{w}) = (\theta_i'(w_i))_{i\in[1;L]}$.

In particular, when $\mathbf{D} = [\mathbf{D}_1^{\mathrm{T}}, \mathbf{D}_2^{\mathrm{T}}, \ldots, \mathbf{D}_J^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{R} = [\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_J]$ where $\mathbf{D}_i$, $\mathbf{R}_i$ are $N_i \times N$ and $N \times N_i$ matrices, then $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\alpha}_2^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_J^{\mathrm{T}}]^{\mathrm{T}}$ where $\boldsymbol{\alpha}_i = \text{diag}\{\mathbf{D}_i\mathbf{R}_i\}$.

Proof: By applying Theorem 1, we only have to prove that in the SURE-LET framework

$$\text{div}\left\{\mathbf{F}(\mathbf{y})\right\} = \boldsymbol{\alpha}^{\mathrm{T}}\Theta'(\mathbf{Dy}). \tag{9}$$

By using the reconstruction formula $\mathbf{F}(\mathbf{y}) = \mathbf{R}\Theta(\mathbf{w})$, i.e., $f_n(\mathbf{y}) = \sum_{l=1}^{L} r_{n,l}\theta_l(w_l)$, and the decomposition formula $\mathbf{w} = \mathbf{Dy}$, i.e., $w_l = \sum_{k=1}^{N} d_{l,k}y_k$, we can successively write the following equalities:

$$\begin{aligned}
\text{div}\left\{\mathbf{F}(\mathbf{y})\right\} &= \sum_{n=1}^{N}\frac{\partial f_n(\mathbf{y})}{\partial y_n} \\
&= \sum_{n=1}^{N}\sum_{l=1}^{L} r_{n,l}\theta_l'(w_l)\frac{\partial w_l}{\partial y_n} \\
&= \sum_{n=1}^{N}\sum_{l=1}^{L} r_{n,l}\theta_l'(w_l)d_{l,n} \\
&= \sum_{l=1}^{L}\theta_l'(w_l)\underbrace{\sum_{n=1}^{N} d_{l,n}r_{n,l}}_{[\mathbf{DR}]_{l,l}}
\end{aligned} \tag{10}$$

and, finally, conclude that $\text{div}\{\mathbf{F}(\mathbf{y})\} = \text{diag}\{\mathbf{DR}\}^{\mathrm{T}}\Theta'(\mathbf{Dy})$. ∎

As it appears in this corollary, the computation of the divergence term—i.e., of $\text{diag}\{\mathbf{DR}\}$—is a crucial point.

1) Evaluation of the Divergence Term—$\boldsymbol{\alpha}$: In the general case where $\mathbf{D}$ and $\mathbf{R}$ are known only by their action on vectors, and not explicitly by their matrix coefficients—typically, when only $\mathcal{D}$ and $\mathcal{R}$ are specified—the analytical expression for $\boldsymbol{\alpha}$ is quite painful to compute: in order to build $\boldsymbol{\alpha}$, for each $l = 1, 2, \ldots L$ it is necessary to compute the reconstruction $\mathbf{f}_l = \mathbf{R}\mathbf{e}_l$ (where $[\mathbf{e}_l]_n = \delta_{n-l}$ is the canonical basis of $\mathbb{R}^L$), then the decomposition $\mathbf{D}\mathbf{f}_l$ and keep the lth component. Given that L is of the order of 256^2—and even much more in the case of redundant transforms—this process may be extremely costly, even considering that it has to be done only once. Fortunately, we can always compute a very good approximation of it using the following numerical algorithm.

For $i = 1 \cdots I$

1) Generate a normalized Gaussian white noise $\mathbf{b}_i \in \mathbb{R}^L$.

2) Apply the reconstruction matrix to $\mathbf{b}_i$ to get the vector $\mathbf{r}_i = \mathbf{R}\mathbf{b}_i$ of size $N \times 1$.
3) Apply the decomposition matrix to $\mathbf{r}_i$ to get the vector $\mathbf{b}'_i = \mathbf{D}\mathbf{R}\mathbf{b}_i$ of size $L \times 1$.
4) Compute the element-by-element product of $\mathbf{b}'_i$ with $\mathbf{b}_i$ to get a vector of L coefficients $\mathbf{v}_i = \mathrm{diag}\{\mathbf{b}'_i\mathbf{b}_i^{\mathrm{T}}\}$, which can be viewed as a realization of the random vector $\mathbf{v} = \mathrm{diag}\{\mathbf{DRbb}^{\mathrm{T}}\}$.

end

An approximate value $\hat{\boldsymbol{\alpha}}$ for $\mathrm{diag}\{\mathbf{DR}\}$ is finally obtained by averaging the realizations $\mathbf{v}_i$ over I runs (typically, $I = 1000$ provides great accuracy)

$$\hat{\boldsymbol{\alpha}} = \frac{1}{I}\sum_{i=1}^{I}\mathbf{v}_i. \tag{11}$$

The above algorithm is justified by the following lemma.

Lemma 2: Let $\mathbf{b}$ be a normalized Gaussian white noise with L components. Then, we have the following equality:

$$\mathcal{E}\left\{\mathrm{diag}\{\mathbf{DRbb}^{\mathrm{T}}\}\right\} = \mathrm{diag}\{\mathbf{DR}\}. \tag{12}$$

Proof:

$$\begin{aligned}\mathcal{E}\left\{\mathrm{diag}\{\mathbf{DRbb}^{\mathrm{T}}\}\right\} &= \mathrm{diag}\Big\{\mathbf{DR}\underbrace{\mathcal{E}\{\mathbf{bb}^{\mathrm{T}}\}}_{\mathbf{Id}}\Big\}\\ &= \mathrm{diag}\{\mathbf{DR}\}.\end{aligned}$$

■

The numerical computation of $\mathrm{diag}\{\mathbf{DR}\}$ can be performed offline for various image sizes, since it does not depend specifically on the image—but for its size—nor on the noise level.

2) Influence of the Boundary Extensions: One of the main drawbacks of any transform-domain denoising algorithm is the potential generation of boundary artifacts by the transform itself. Decreasing these effects is routinely done by performing boundary extensions, the most popular choice being symmetric extension and periodic extension. Thus, the effect of these extensions boils down to replacing the transformation $\mathbf{D}$ by another transformation, $\mathbf{D}'$.

Indeed, usual boundary extensions are linear preprocessing applied to the available data $\mathbf{y}$ and can, therefore, be expressed in a matrix form. In particular, for a given boundary extension of length E, i.e., characterized by an $E \times N$ matrix $\mathbf{H}$, the denoising process becomes

$$\begin{aligned}\mathbf{F}(\mathbf{y}) &= [\mathbf{Id}_N\ \mathbf{0}_{N\times E}]\mathbf{R}_{N+E}\Theta\left(\mathbf{D}_{N+E}\begin{bmatrix}\mathbf{y}\\ \mathbf{Hy}\end{bmatrix}\right)\\ &= \mathbf{R}'\Theta(\mathbf{D}'\mathbf{y})\end{aligned}$$

where $\mathbf{D}_{N+E}$ (resp., $\mathbf{R}_{N+E}$) is the matrix corresponding to the linear transformation $\mathcal{D}$ (resp., $\mathcal{R}$) when the input signal is of size $N + E$. Any boundary handling can, therefore, be seen as a modification of the decomposition matrix $\mathbf{D}$ that must be taken into account when computing the divergence term, namely $\mathrm{diag}\{\mathbf{D}'\mathbf{R}'\}$. This is where Lemma 2 is particularly useful: although the implementation of the transformations $\mathcal{D}$ and $\mathcal{R}$ with the adequate boundary extensions may be straightforward, the explicit computation of the coefficients of the matrices $\mathbf{R}'$ and $\mathbf{D}'$ is tedious—and Lemma 2 avoids this computation.

3) Applications to Standard Linear Transforms: In some particular cases of linear transforms, it is possible to easily compute $\mathrm{diag}\{\mathbf{DR}\}$ analytically, as shown in the following.

a) Nonredundant transforms: Here, we assume that the number of samples is preserved in the transform domain, and more precisely:

- $\mathbf{D}$ is a full rank matrix of size $N \times N$;
- $\mathbf{R}$ is also a full rank matrix of size $N \times N$.

Then, the following lemma shows how to compute the divergence term of Corollary 1.

Lemma 3: When $\mathbf{D}$ is nonredundant, the divergence term $\boldsymbol{\alpha}$ in (8) is given by

$$\boldsymbol{\alpha} = \underbrace{[1, 1, \ldots, 1]}_{L\text{ times}}^{\mathrm{T}}. \tag{13}$$

Proof: Because $\mathbf{DR} = \mathbf{RD} = \mathbf{Id}$, we have $\boldsymbol{\alpha} = \mathrm{diag}\{\mathbf{DR}\} = \mathrm{diag}\{\mathbf{Id}\}$. ■

Note that, when additionally the transformation is *orthonormal*, the reconstruction matrix is simply the transpose of the decomposition matrix, i.e., $\mathbf{R} = \mathbf{D}^{\mathrm{T}}$. Consequently, in corollary 1, the SURE becomes

$$\begin{aligned}\epsilon &= \frac{1}{N}\left\|\mathbf{F}(\mathbf{y}) - \mathbf{y}\right\|^2 + \frac{2\sigma^2}{N}\boldsymbol{\alpha}^{\mathrm{T}}\Theta'(\mathbf{Dy}) - \sigma^2\\ &= \frac{1}{N}\left\|\Theta(\mathbf{Dy}) - \mathbf{Dy}\right\|^2 + \frac{2\sigma^2}{N}\boldsymbol{\alpha}^{\mathrm{T}}\Theta'(\mathbf{Dy}) - \sigma^2\\ &\quad (\text{orthogonality of } \mathbf{R})\\ &= \frac{1}{N}\sum_{i=1}^{N}\Big((\theta_i(w_i) - w_i)^2 + 2\sigma^2\theta'_i(w_i)\Big) - \sigma^2\end{aligned} \tag{14}$$

where w_i is the ith component of $\mathbf{Dy}$; i.e., it is a sum of the specific MSE estimates for each transformed coefficient w_i. The optimization procedure can, thus, be performed *separately* in the transform domain [5]. This is specific to orthonormal transforms: nonredundant biorthogonal transforms do not enjoy this property; i.e., the optimization in the transform domain is *not* equivalent to the optimization in the image domain. Yet, Lemma 3 still applies and is actually particularly useful for applying our SURE minimization strategy.

b) Undecimated filterbank transforms: Here, we will consider linear redundant transforms characterized by J analysis filters $\tilde{G}_i(z) = \sum_n \tilde{g}_i[n]z^{-n}$ and J synthesis filters $G_i(z) = \sum_n g_i[n]z^{-n}$ as shown in Fig. 1.

A periodic boundary extension implementation of this structure gives rise to decomposition and reconstruction matrices $\mathbf{D}$ and $\mathbf{R}$ made of J circulant submatrices—i.e., diagonalized with an N-point DFT matrix—$\mathbf{D}_i$ and $\mathbf{R}_i$ of size $N \times N$ each, with coefficients

$$\begin{aligned}[\mathbf{D}_i]_{k,l} &= \sum_n \tilde{g}_i[l - k + nN]\\ [\mathbf{R}_i]_{k,l} &= \sum_n g_i[k - l + nN].\end{aligned}$$

We then have the following lemma to be used in Corollary 1:

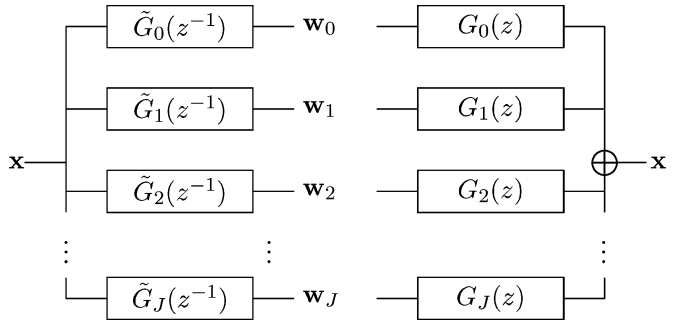


Fig. 1. Undecimated J-band analysis-synthesis filterbank.

Lemma 4: When $\mathbf{D}$ and $\mathbf{R}$ are periodically extended implementations of the analysis-synthesis filterbank of Fig. 1, the divergence term $\boldsymbol{\alpha}$ in (8) is given by $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\alpha}_2^{\mathrm{T}}, \dots \boldsymbol{\alpha}_J^{\mathrm{T}}]^{\mathrm{T}}$ where

$$\boldsymbol{\alpha}_i = \left(\sum_n \gamma_i[nN] \right) \underbrace{[1, 1, \dots, 1]^{\mathrm{T}}}_{N\text{times}}. \tag{15}$$

and where $\gamma_i[n]$ is the nth coefficient of the filter $\tilde{G}_i(z^{-1})G_i(z)$. The extension to filterbanks in higher dimensions is straightforward—the summation in (15) running over a multidimensional index n.

Proof: According to Corollary 1, we have to compute $\boldsymbol{\alpha}_i = \mathrm{diag}\{\mathbf{D}_i\mathbf{R}_i\}$. Since $\mathbf{D}_i$ and $\mathbf{R}_i$ are circulant matrices the product $\mathbf{D}_i\mathbf{R}_i$ is also circulant and is built using the N-periodized coefficients of the filter $\tilde{G}_i(z^{-1})G_i(z)$, i.e.,

$$[\mathbf{D}_i\mathbf{R}_i]_{k,l} = \sum_n \gamma_i[k - l + nN]$$

the diagonal of which yields (15). ∎

It is often assumed that $\tilde{G}_i$ and G_i satisfy the biorthogonality condition

$$\sum_{k=0}^{M-1} \underbrace{\tilde{G}_i(z^{-1}e^{-j2\pi k/M_i})\tilde{G}_i(ze^{j2\pi k/M_i})}_{\Gamma_i(ze^{j2\pi k/M_i})} = \lambda_i \tag{16}$$

where M_i is a divisor of N, because undecimated filterbanks are usually obtained from critically sampled filterbanks—for which (16) holds with $\lambda_i = M_i$. In this case, since (16) actually specifies the coefficients $\gamma_i[nM_i]$, we find that $\boldsymbol{\alpha}_i = \lambda_i/M_i[1, 1, \dots, 1]^{\mathrm{T}}$.

An example of such a transform is the standard undecimated wavelet transform (UWT) which uses $J + 1$ ($3J + 1$ in two dimensions) orthonormal filters (see Fig. 2). In that case, the equivalent filters are given by

$$\begin{aligned}\tilde{G}_i(z) &= 2^i G_i(z) \\ &= H(z)H(z^2)\dots H\left(z^{2^{i-2}}\right) G\left(z^{2^{i-1}}\right) \\ &\quad \text{for } i = 1, 2, \dots, J \\ \tilde{G}_{J+1}(z) &= 2^J G_{J+1}(z) = H(z)H(z^2)\dots H\left(z^{2^{J-1}}\right).\end{aligned}$$

They satisfy (16) for $\lambda_i = 1$. This shows that $\boldsymbol{\alpha}_i = 2^{-i}[1, 1, \dots, 1]^{\mathrm{T}}$ for all $i = 1, 2, \dots J$ and $\boldsymbol{\alpha}_{J+1} = 2^{-J}[1, 1, \dots, 1]^{\mathrm{T}}$. In a 2-D separable framework, these values are extended straightforwardly, taking into account that the 2-D filters still satisfy (16) for $\lambda_i = 1$: the general result is, thus, that $\boldsymbol{\alpha}_i$ is given by the (2-D) downsampling factor $1/M_i$.

III. Example of a SURE-LET Denoising Algorithm

In Section II-C, we have proposed a general form of denoising function (7), which involves several degrees of freedom: the linear transformation, the number K of linear parameters, and the thresholding functions Θ_k. This section studies a possible choice. The denoising performance of the resulting algorithm will be evaluated in the next section.

First, we will restrict ourselves to the undecimated wavelet transform,[6] although other linear transforms may in some cases be more advisable—e.g., the undecimated DCT, the curvelet transform, etc…

A. Choosing an Efficient Thresholding Function

A pointwise thresholding function is likely to be efficient if it satisfies the following minimal properties:

- differentiability: required to apply Theorem 1—this rules out pure hard-thresholds;
- anti-symmetry: we assume that the coefficients are not expected to exhibit a sign preference;
- linear behavior for large coefficients: because when a coefficient is large, it can be kept unmodified—noise corruption is negligible.

A good choice has been experimentally found to be of the form

$$\begin{aligned}\theta_i(w) &= a_{i,1}t_1(w) + a_{i,2}t_2(w) \\ \text{where} \quad t_1(w) &= w \text{ and } t_2(w) = w\left(1 - e^{-\left(\frac{w}{3\sigma}\right)^8}\right)\end{aligned} \tag{17}$$

in each band i. The nonlinear term, $t_2(w)$, can be seen as a regular approximation of a *Hard-threshold*.

Similarly to what was observed empirically in other settings [5], [17], adding more thresholding functions only bring marginal (∼0.1–0.2 dB) improvement to the overall denoising quality.

B. Solving for the Linear Parameters

Finding the parameters $a_{i,k}$ that minimize the MSE estimate ϵ amounts to solving the linear system of (5) in which it is necessary to replace $\mathbf{F}(\mathbf{y})$ by

$$\mathbf{F}(\mathbf{y}) = \sum_{i=1}^{J}\sum_{k=1}^{2} a_{i,k}\mathbf{F}_{i,k}(\mathbf{y}) + \text{lowpass}$$

where $\mathbf{F}_{i,k}$ is the image obtained by zeroing all the bands $i' \neq i$ and processing the subband i with the thresholding function $t_k(w)$. Note that, as usual in denoising algorithms, the $(J+1)$th band, lowpass, is not processed.

As shown in Section II-C3b, the divergence term in (5) has an exact expression, namely $\mathrm{div}\{\mathbf{F}_{i,k}(\mathbf{y})\} = 4^{-i}\sum_{w\in\mathrm{band}i} t_k'(w)$. Alternatively, in particular, in the case of nonperiodic boundary image extensions, it is possible to use the approximate algorithm presented in Section II-C2.

C. Summary of the Algorithm

1) Perform a boundary extension on the noisy image.

[6] In our tests, the best performer was the Haar wavelet.

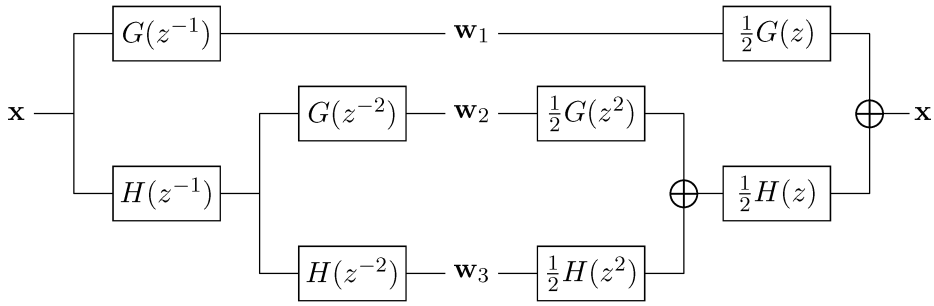


Fig. 2. Classical undecimated wavelet filterbank for 1-D signal.

2) Perform an UWT on the extended noisy image.
3) **For** $i = 1 \ldots J$ (number of bandpass subbands), **For** $k =$ 1, 2:
   a) Apply the *pointwise* thresholding functions t_k defined in (17) to the current subband $\mathbf{w}_i$.
   b) Reconstruct the processed subband by setting all the other subbands to zero to obtain $\mathbf{F}_{i,k}(\mathbf{y})$.
   c) Compute the first derivative of t_k for each coefficient of the current subband $\mathbf{w}_i$ and build the corresponding coordinate of $\mathbf{c}$ as exemplified by (5).

   end

   end
4) Compute the matrix $\mathbf{M}$ and deduce the optimal—in the minimum SURE sense—linear parameters $a_{i,k}$'s using the matrix formulation of (5).
5) The noise-free image $\hat{\mathbf{x}}$ is finally estimated by the sum of each $\mathbf{F}_{i,k}$ weighted by its corresponding SURE-optimized $a_{i,k}$.

IV. Results

A. Wavelet-Domain Versus Image-Domain Optimization

Before comparing our SURE-LET approach with the best state-of-the-art algorithms, we demonstrate here that, in order to optimize the denoising process, it is essential to perform the minimization in the *image-domain*. Instead, an *independent wavelet subband* processing is suboptimal, often by a significant margin, even in a "tight" frame representation. This is because we usually do not have energy preservation between the denoised "tight" frame coefficients $\hat{\mathbf{w}}$ and the reconstructed image $\hat{\mathbf{x}} = \mathbf{R}\hat{\mathbf{w}} : \|\mathbf{R}\hat{\mathbf{w}}\| \neq \|\hat{\mathbf{w}}\|$. This is not in contradiction with the well-known energy conservation between the "tight" frame coefficients $\mathbf{w} = \mathbf{D}\mathbf{y}$ and the noisy image $\mathbf{y} : \|\mathbf{D}\mathbf{y}\| = \|\mathbf{y}\|$.

In Fig. 3, we compare a classical wavelet domain SURE-based optimization of our thresholding function (17) with the image domain optimization based on Lemma 4 in the framework of the undecimated *Haar* wavelet transform. We notice that the rigorous image domain optimization provides large improvements—up to $+1$ dB—over the independent in-band optimization. A closer examination of the "optimal" thresholds in both cases indicates that this difference may be related to the difference between the slopes of these functions around zero: the image-domain solution is actually much flatter, making it able to suppress small coefficients almost exactly.

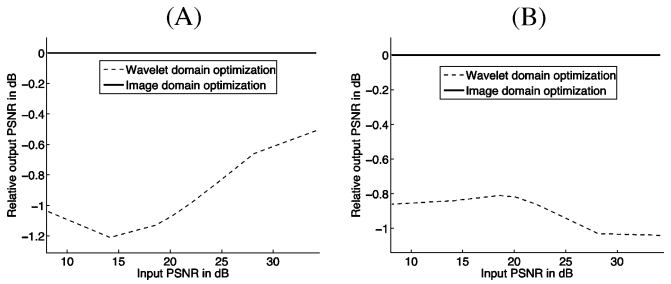


Fig. 3. Comparison of the proposed *SURE-LET* denoising procedure with a SURE-based denoising algorithm optimized in the wavelet domain when using the undecimated wavelet (*Haar*) transform: (a) *House*; (b) *Al*.

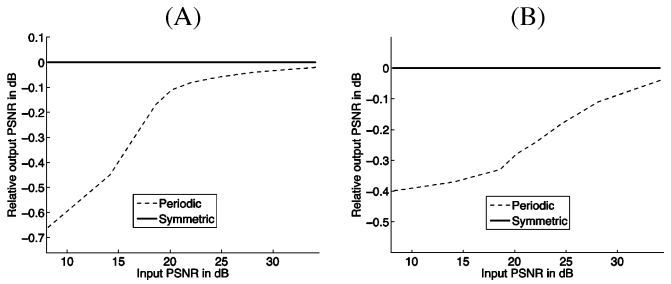


Fig. 4. Influence of the boundary extensions when using the undecimated wavelet (*Haar*) transform: (a) *Peppers*; (b) *House*.

B. Periodic Versus Symmetric Boundary Extensions

It is also worth quantifying the effects of particular boundary extensions. In Fig. 4, we compare symmetric boundary extensions (rigorous SURE computation, as described in Section II-C-2) with the periodic ones. As it can be observed, the symmetric boundary extension can lead to up to $+0.5$ dB of PSNR improvements over the periodic one.

C. Comparison With State-of-the-Art Denoising Schemes

We have compared our Haar wavelet SURE-LET denoising algorithm with some of the best state-of-the-art techniques for which the code is freely distributed by the authors: *BiShrink* [4] (dual tree complex wavelet decomposition), *ProbShrink* [2] (undecimated Daubechies *symlets*) and *BLS-GSM* [1] (full steerable—eight orientations per scale—pyramidal decomposition). Depending on the size of the images, 256×256 or 512×512, we have performed 4 or 5 decomposition levels.

For a reliable comparison, we have run all the algorithms[7] on a comprehensive set of standard grayscale[8] images of size

[7]We have used the same parameters as those suggested by the authors in their respective papers and softwares.

[8]8-bit images with pixels values between 0 and 255.

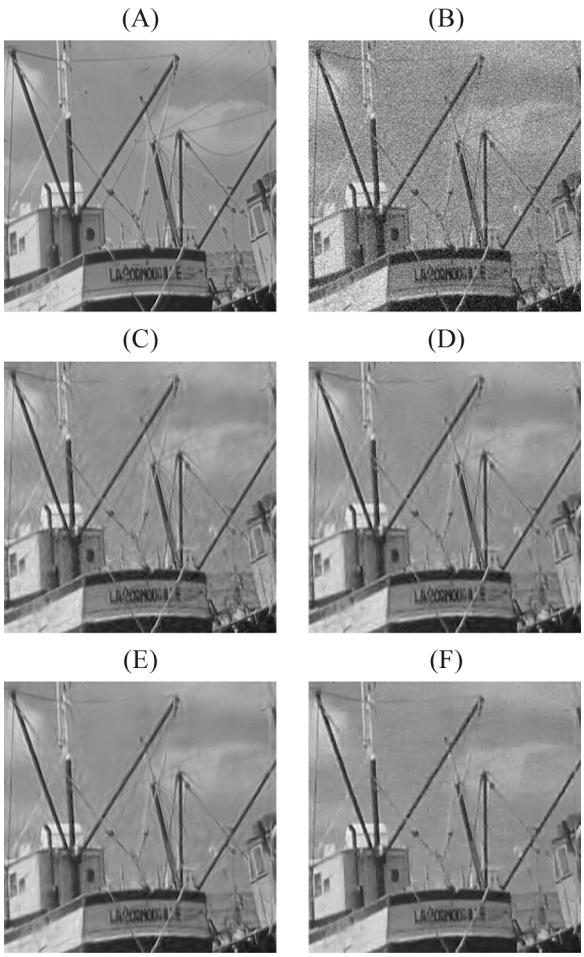


Fig. 5. (a) Part of the noise-free *Boat* image. (b) A noisy version of it: $\mathrm{PSNR} = 22.11$ dB. (c) *BiShrink* denoising result: $\mathrm{PSNR} = 29.99$ dB. (d) *ProbShrink* denoising result: $\mathrm{PSNR} = 29.97$ dB. (e) *BLS-GSM* denoising result: $\mathrm{PSNR} = 30.36$ dB. (f) *UWT SURE-LET* denoising result: $\mathrm{PSNR} = 30.24$ dB.

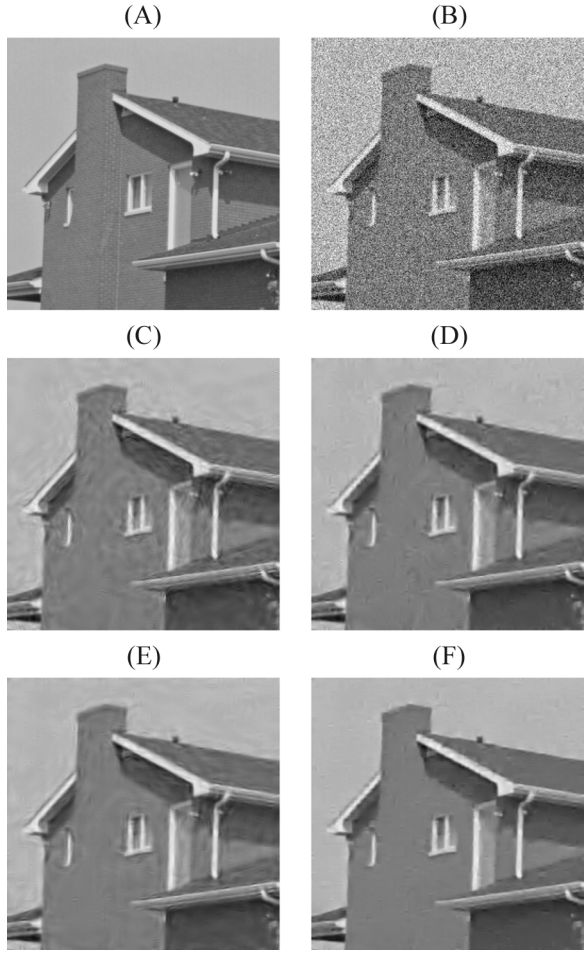


Fig. 6. (a) Noise-free *House* image. (b) A noisy version of it: $\mathrm{PSNR} = 18.59$ dB. (c) *BiShrink* denoising result: $\mathrm{PSNR} = 29.77$ dB. (d)*ProbShrink* denoising result: $\mathrm{PSNR} = 30.33$ dB. (e) *BLS-GSM* denoising result: $\mathrm{PSNR} = 30.50$ dB. (f) *UWT SURE-LET* denoising result: $\mathrm{PSNR} = 30.90$ dB.

256×256 (*Peppers*, *House*, *Bridge*) and of size 512×512 (*Al*, *Barbara*, *Boat*, *Crowd*, *Goldhill*), each one corrupted with additive Gaussian white noise at eight different power levels $\sigma \in$[5,10,15,20,25,30,50,100], which corresponds to PSNR decibel values [34.15,28.13,24.61,22.11,20.17,18.59,14.15,8.13]. We have then averaged the output PSNRs over eight noise realizations (the different algorithms are applied to the same noise realizations).

Table I reports the PSNR results we have obtained with the various denoising methods, the best results being shown in boldface. As we can notice, our algorithm (*UWT SURE-LET*) matches the best state-of-the-art results for most of the images, except for Barbara where it may be argued that, either a finer subband decomposition, or a more sophisticated, multivariate, thresholding function should be used in order to capture the texture information that characterizes this image. Note also how the SURE minimization is close to the MSE one (*Oracle* in Table I), which is an evidence of the robustness of the SURE-LET approach.

We want to stress that the denoising algorithm we propose in this section is limited to a pointwise thresholding, contrary to the above mentioned algorithms which involve some kind of multivariate thresholding. Because it simply boils down to solving a linear system of equations, our algoritm is quite fast compared to *BLS-GSM* which has the best denoising results. More precisely, the execution of our current un-optimized Matlab implementation of the whole denoising task lasts on average 3.5 s for 256×256 images and about 26 s for 512×512 on a Power Mac G5 with CPU speed of 1.8 GHz and 1 GB of memory, whereas Portilla *et al. BLS-GSM* lasts, respectively, 25 and 100 s on the same workstation. Note that the main part of our computational time is dedicated to the independent reconstruction of all the subbands.

Other preliminary tests indicate that if, for images like Barbara, we choose transforms that have more subbands (such as the undecimated DCT), our simple pointwise thresholding strategy may provide slightly better results than *BLS-GSM* (typically, $+0.2$ dB); moreover, it is possible to select a transform or the other based only on the SURE values. We may also envision that thresholding schemes that involve inter and intrascale dependences substantially improve the denoising performance, as this is the case with orthonormal wavelet transforms [5].

We can finally notice in Figs. 5 and 6 that our *SURE-LET* denoising procedure gives quite a decent visual quality compared to the best state-of-the-art spatially adaptive method.

V. Conclusion

We have presented a new approach to image denoising that is especially useful when redundant or nonorthonormal transforms are involved. In this paper, we have emphasized the theoretical part of our approach and its implementation aspects, in order to make the SURE-LET principle easily applicable for others. Accordingly, we did not try to take advantage of all the degrees of freedom (multivariate thresholding, increased number of parameters, more sophisticated transforms) to make our example of algorithm optimal. And yet, the obtained results are quite competitive with the best state-of-the-art denoising algorithms—which require involved statistical image models. This indicates that there is a substantial margin of improvement of SURE-LET type algorithms.

Acknowledgment

The authors would like to thank Prof. M. Unser for useful discussions, and the anonymous reviewers for their suggestions.

References

[1] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[2] A. Pižurica and W. Philips, "Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 654–665, Mar. 2006.

[3] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, Nov. 2002.

[4] L. Sendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 438–441, Dec. 2002.

[5] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Inter-scale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 593–606, Mar. 2007.

[6] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[7] J.-C. Pesquet and D. Leporini, "A new wavelet estimator for image denoising," in *Proc. 6th Int. Conf. Image Processing and Its Applications*, Jul. 14–17, 1997, vol. 1, pp. 249–253.

[8] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814–1830, Nov. 2005.

[9] C. Chaux, L. Duval, A. Benazza-Benyahia, and J.-C. Pequet, "A nonlinear Stein based estimator for multichannel image denoising," *IEEE Trans. Signal Process.*, to be published.

[10] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, 1981.

[11] I. Daubechies, "Ten lectures on wavelets," presented at the CBMS-NSF Regional Conf. Ser. Applied Mathematics, Mar. 1992.

[12] G. Nason and B. W. Silverman, *The Stationary Wavelet Transform and Some Statistical Applications*. New York: Springer-Verlag, 1995, vol. 103.

[13] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, Jun. 2002.

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

[15] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[16] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[17] M. Raphan and E. P. Simoncelli, "Learning to be Bayesian without supervision," presented at the NIPS Conf., Dec. 2006.

[18] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[19] M. K. Mihçak, Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, Dec. 1999.

[20] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[21] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1522–1531, Sep. 2000.

[22] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[23] X.-P. Zhang and M. D. Desai, "Adaptive denoising based on SURE risk," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 265–267, Oct. 1998.

[24] N. G. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Roy. Soc. A*, Sep. 1999.

[25] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *J. Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, May 2001.

[26] F. Abramovitch, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy. Statist. Soc.*, ser. B, vol. 60, no. 4, pp. 725–749, 1998.

[27] E. P. Simoncelli, *Bayesian Denoising of Visual Images in the Wavelet Domain*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, Mar. 1999, vol. 141.

[28] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: Wiley-Interscience, Apr. 1999.

[29] P. L. Combettes and J.-C. Pesquet, "Wavelet-constrained image restoration," *Int. J. Wavelets, Multires. Inf. Process.*, vol. 2, no. 4, pp. 371–389, Dec. 2004.

Thierry Blu (M'96–SM'06) was born in Orléans, France, in 1964. He received the "Diplôme d'ingénieur" from École Polytechnique, France, in 1986 and from Télécom Paris (ENST), France, in 1988, and the Ph.D. degree in electrical engineering from ENST in 1996 for a study on iterated rational filterbanks, applied to wideband audio coding.

He is with the Biomedical Imaging Group at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, on leave from the France Telecom R&D center in Issy-les-Moulineaux. At EPFL, he teaches the theory of signals and systems for microengineering and Life Science students. His research interests include (multi)wavelets, multiresolution analysis, multirate filterbanks, interpolation, approximation and sampling theory, image denoising, psychoacoustics, optics, and wave propagation.

Dr. Blu was the recipient of two best paper awards from the IEEE Signal Processing Society (2003 and 2006). From 2002 and 2006, he was an Associate Editor for the IEEE Transactions on Image Processing and, since 2006, he has been an Associate Editor for the IEEE Transactions on Signal Processing.

Florian Luisier was born in Switzerland in 1981. In 2005, he received his M.S. degree in microengineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Swizerland. He is currently pursuing the Ph.D. degree with the Biomedical Imaging Group (BIG), EPFL.

His research interests mainly include multiresolution analysis and the restoration of biomedical images.