Geometrical Priors for Noisefree Wavelet Coefficients in Image Denoising

Maarten Jansen
Adhemar Bultheel

to appear in *Bayesian Inference in Wavelet Based Models*, Lecture Notes in Statistics, Mueller, P. and Vidakovic, B. (Eds.), New York: Springer-Verlag

ABSTRACT Wavelet threshold algorithms replace wavelet coefficients with small magnitude by zero and keep or shrink the other coefficients. This is basically a local procedure, since wavelet coefficients characterize the local regularity of a function. Although a wavelet transform has decorrelating properties, structures in images, like edges, are never decorrelated completely, and these structures appear in the wavelet coefficients. We therefore introduce a geometrical prior model for configurations of large wavelet coefficients and combine this with the local characterization of a classical threshold procedure into a Bayesian framework. The threshold procedure selects the large coefficients in the actual image. This observed configuration enters the prior model, which, by itself, only describes configurations, not coefficient values. In this way, we can compute for each coefficient the probability of being "sufficiently clean".

1 Introduction

Wavelet thresholding (Donoho and Johnstone, 1994; Donoho and Johnstone, 1995) is a popular method for the reduction of noise in images. It assumes that the original, non-corrupted image can be represented in a sparse way by a small number of large wavelet coefficients. In the case of an orthogonal transform, i.i.d. noise is spread out equally over all coefficients. Selecting the coefficients with the largest magnitude therefore removes most of the noise, while preserving the essential image information. In the case of correlated noise (Johnstone and Silverman, 1997) or a non-orthogonal transform, the multiresolution structure of a wavelet transform justifies the application of scale-dependent thresholds.

Whereas typical threshold algorithms are based on this heuristical approach that the largest coefficients capture the essential image features, *Bayesian* methods, including (Simoncelli and Adelson, 1996; Vidakovic,

1994; Ruggeri and Vidakovic, 1995) start from a full model for wavelet coefficients of the following type:

$$\boldsymbol{W} = \boldsymbol{V} + \boldsymbol{N}$$

This is an additive model for wavelet coefficients where $\boldsymbol{N}$ is the noise vector, $\boldsymbol{V}$ is the uncorrupted wavelet coefficient vector, and $\boldsymbol{W}$ is the input (empirical) wavelet coefficient vector. Both the noise and the noise-free data are viewed as realizations of a probability distribution. A Bayesian method considers $\boldsymbol{W}$ as an observation from which it computes (in principle) the posterior distribution of $\boldsymbol{V}$. This is used to estimate the uncorrupted data coefficients. For example, minimization of the L_2-loss leads to the posterior mean:

$$\hat{\boldsymbol{V}} = \mathrm{E}(\boldsymbol{V}|\boldsymbol{W})$$

For some models of $\boldsymbol{V}$ and $\boldsymbol{N}$, Bayesian estimators can be interpreted as threshold procedures, although often more complicated than the simple hard- or soft-thresholding. Because of the complete modelling, this approach seems less heuristic. However, most models carry hyperparameters, that have to be chosen, often on a heuristic basis.

In this Chapter, we start from a wavelet threshold procedure and discuss why — especially for images — thresholding is a too local operation: important wavelet coefficients are concentrated around image structures, like edges. Section 3 therefore introduces a geometrical prior model for structures of important coefficients. This is typically a Gibbs Random Field. Using a geometrical prior that supports important coefficient *configurations*, we "correct" the result of threshold procedure. Our model introduces the notion of important and noisy coefficients. Both these classes have an according probability density. Other approaches with a similar *mixture model* for wavelet coefficients include (Malfait and Roose, 1997; Malfait, 1995; Chipman et al., 1997; Crouse et al., 1998; Clyde et al., 1998; Abramovich et al., 1997).

2 Wavelet Thresholding

2.1 Wavelet terminology

A discrete wavelet transform is a decomposition of an input vector into coefficients at different resolution levels. Each resolution represents image features of a specific scale. In two dimensions, we use a tensor-product like extension of a one-dimensional transform, the so called square wavelet-transform: at each level the coefficients belong to one of three *subbands* or *components*, corresponding to three orientations in the image: the coefficients carry vertical, horizontal, or diagonal information. Instead of the

classical fast wavelet transform, we use the so called non-decimated, redundant, or stationary wavelet transform (Nason and Silverman, 1995; Pesquet et al., 1996; Malfait and Roose, 1997). While the fast transform has linear complexity, this alternative is of $\mathcal{O}(N \log N)$ complexity, both in computations and in memory requirements. On the other hand, the stationary transform is translation invariant and generates the same number of coefficients in each subband at each scale. This facilitates interscale coefficient comparisons. Moreover, the reconstruction procedure (Nason and Silverman, 1995) from this redundant data representation is of course not unique. For manipulated data, a linear combination of all possible reconstruction schemes causes an additional smoothing of the result (Nason and Silverman, 1995).

2.2 Threshold schemes

A wavelet threshold algorithm typically consists of three steps: first, the observational data are transformed into empirical wavelet coefficients. The next step is a manipulation of the coefficients and finally, an inverse transform of the modified coefficients yields the result.

Of course, the second step is the crucial one. Apart from the choice of the wavelet basis, all strategies and options are concentrated in this step. As mentioned before, heuristical arguments or at least assumptions about the noise-free image regularity, motivate threshold rules and threshold selection procedures. The manipulation is based on a *classification* of the coefficients: the most common threshold procedures use a binary classification: a coefficient is either dominated by noise *or* sufficiently important. The magnitude of the coefficient is the *criterion* of classification: we consider the absolute value of each coefficient as a measure of significance. This could of course be generalized: the measure of significance $\boldsymbol{M}$ is a function of the observation:

$$\boldsymbol{M} = m(\boldsymbol{W}),$$

Wavelet coefficients with a significance below a threshold δ, are classified as noisy. With each wavelet coefficient W_s, the algorithm associates a 'label' or 'mask' variable X_s such that:

$$X_s = \begin{cases} 0, & \text{if } W_s \text{ is noisy according} \\ & \text{to the criterion, i.e. if} M_s < \delta \\ 1, & \text{then } W_s \text{ is sufficiently clean, i.e. if} \\ & M_s \geq \delta \end{cases}$$

In these and following equations, s represents the 'multidimensional' index of a wavelet coefficient on a given resolution level j and for a given component (vertical, horizontal, or diagonal): $s = (k, l)$. To avoid overloaded notations we omit the resolution level j and the component m in our equations, and use the simple index s. So, if no confusion is possible, we write

W_s instead of $W^m_{j;s}$ or $W^m_{j;k,l}$. This classification is followed by the modification step: If $W_{\delta s}$ denotes the modified coefficient, with subscript δ refering to the threshold value, we write:

$$W_{\delta s} = h(W_s, M_s, X_s),$$

for some *action* $h(W_s, M_s, X_s)$. The classic *hard-threshold* procedure corresponds to

$$h(W_s, M_s, X_s) = W_s X_s.$$

It keeps the 'uncorrupted' coefficients and replaces the noisy ones by zero.

2.3 Thresholding is local

Wavelet basis functions are localized in space and scale (frequency). As a consequence, manipulating a coefficient has a local effect, both in space and frequency. This is an important advantage of wavelet based methods.

On the other hand, usual *classification* rules are local too, and do not take into account all the correlations that exist among neighboring coefficients. Although a wavelet transform has decorrelating properties, this decorrelation is not complete (a wavelet transform is sometimes seen as an *approximation* of a Karuhnen-Loève-transform). We distinguish two types of correlations:

1. Important image features correspond to large coefficients at different scales: these coefficients are of course correlated. This correlation can be used as the significance criterion in the qualification of wavelet coefficients in a threshold procedure (Xu et al., 1994): coefficients with a small interlevel correlation are replaced by zero. Alternatively, this multiscale structure can be incorporated in a prior model (Crouse et al., 1998).

2. The second type of correlation is within one scale: especially in images, important coefficients tend to be clustered on the location of edges.

In this chapter, we neglect the first type of correlations, and we concentrate on the second type of clustering.

2.4 Threshold mask images

Figure 1 shows an image with artificial, additive, homoscedastic correlated noise. This noise was the result of a convolution of white noise with a FIR-highpass-filter (A FIR or finite impuls response filter has a finite number of taps). The signal-to-noise ratio is 4.97 dB, where we define signal-to-noise ratio as:

$$SNR = 10 \log_{10} \left(\frac{\frac{1}{N} \sum_{i=1}^{N} (f_i - \overline{f})^2}{\sigma^2} \right),$$



FIGURE 1. An image (Left) with artificial, additive correlated noise (Right), SNR = 4.97dB. The noise is the result of a convolution of white noise with a FIR-highpass-filter.

with $\overline{f} = \frac{1}{N}\sum_{i=1}^{N} f_i$. f_i is the gray intensity of pixel i. We apply a redundant wavelet transform. As wavelet filter, we use the variation on the CDF-(spline)-filters "with less dissimilar lengths" (Cohen et al., 1992; Antonini et al., 1992). We choose a basis with four primal and four dual vanishing moments. These wavelets are rather popular in image processing: they are smooth and have compact support.

Since we know the uncorrupted image, we can compute at each level the threshold that minimizes the mean square error (MSE) of the wavelet coefficients. Since we use a bi-orthogonal (i.e. not an orthogonal) transform, this is not exactly the same as a minimization of the MSE of the pixel values, although a *stable* basis (with Riesz constants close to each other (Daubechies, 1992)) guarantees a quasi-equivalence. Moreover, the eventual objective is visual quality, and there seems no reason why minimization in terms of original pixels is better than minimization in terms of wavelet coefficients. Of course, the user views images in the pixel domain, but we do not interpret an image pixel by pixel. Although a discussion about the human visual system is far beyond the scope of this text, a multiresolution analysis is said to be closer to the way we view an image. Note that MSE minimization is equivalent to SNR maximization. The black pixels in the binary image of Figure 2 indicate the empirical coefficients of the horizontal subband at the one but finest scale with magnitude larger than the minimum MSE threshold (The one but finest scale in the wavelet transform is two scales below the original image resolution). These coefficients are kept, while the white pixels correspond to coefficients that are replaced by zero. Although we recognize image structures, like edges, there also many isolated black pixels.

In practical situations, the minimum MSE threshold cannot be computed exactly. So, the next picture shows the same *mask* or *label* image for an

estimate of the minimum MSE threshold, based on generalized cross validation (GCV) procedure (Weyrich and Warhola, 1995; Jansen et al., 1997). This method is as fast as the wavelet transform, requires no estimation for the noise deviation σ and it is asymptotically optimal, i.e. if the number of data is getting large, the estimated threshold approaches the minimum MSE threshold. These properties are preserved — on a level-dependent basis — if the noise is correlated in the wavelet representation (Jansen and Bultheel, 1998). Some other methods, like (Berkner and Wells, Jr., 1998) are based on a "probabilistic upper bound" for the noise in the wavelet domain. This upper bound and the corresponding threshold change if the noise on the wavelet coefficients is correlated. GCV-procedures, however, aim at a minimum MSE threshold and they can be used level by level without further modifications in the case of correlated noise and non-orthogonal transforms.

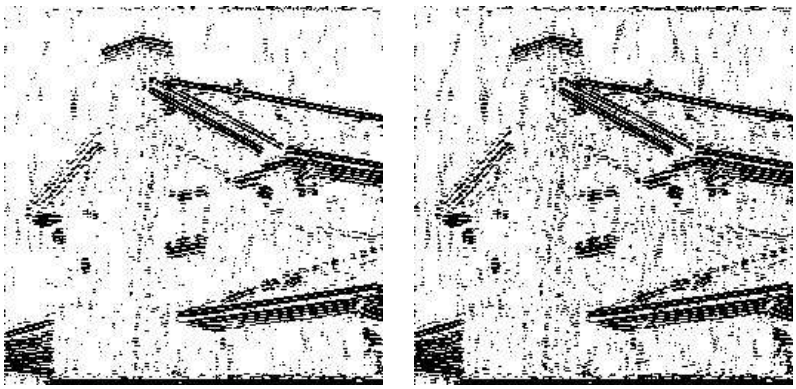
If we apply the same threshold to the noise-free coefficients, we get the left picture in Figure 3. We see that many of the isolated pixels have disappeared: they were due to noise.

However, the minimum MSE threshold is based on a global compromise between noise and data: this is not the best thing we can do: instead of applying one threshold for all coefficients at a given level, we would like to decide for each coefficient separately what is best: keeping or killing. If we know the noise deviation σ and the noise-free magnitude V_s, then the expected MSE is minimized by replacing W_s with zero if $|V_s| < \sigma$ and keeping it otherwise (Donoho and Johnstone, 1994). This is an ideal *diagonal projection*: instead of applying the minimum MSE threshold (or its estimate) to the noise-free coefficients, we choose a threshold on these uncorrupted coefficients equal to the noise deviation. The resulting label image is shown in Figure 3 (Right picture). Since neither V_s nor σ are known in practice, this is of course an ideal case.

To conclude this discussion, we compare the result from hard-thresholding with level- and subband-dependent GCV-threshold with the result from diagonal projection. Only the three finest scales were processed. Signal-to-noise ratio is respectively 18.64 dB and 21.05 dB.

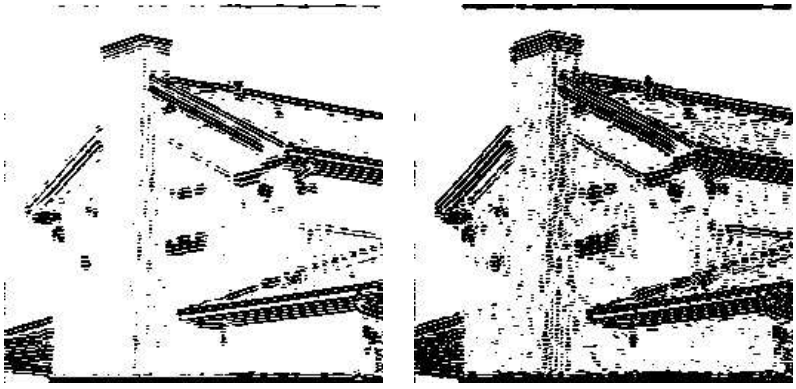
2.5 Binary image enhancement methods

A comparison of the different label images clearly illustrates that thresholding each coefficient separately does not take into account the image structures. An obvious way to recover the optimal mask of Figure 3 (b), is trying classic enhancement methods. Figure 5 (a) shows the label image after applying a median filter to the approximate minimum MSE labels in Figure 2. Another possibility is the application of so called erosion-dilation methods: these methods proceed in two steps: in the first step, black pixels with less than, for instance, two black neighbors are removed. This erosion can be repeated several times, before going to the dilation. This second



(a) (b)

FIGURE 2. Mask or label images, corresponding to the horizontal component of the one but finest scale. Black pixels represent coefficients with magnitude above the threshold. Left: using the minimum MSE threshold. Right: using a GCV estimate of this threshold.



(a) (b)

FIGURE 3. Same mask images as in Figure 2, here based on noise-free coefficients. Left: black pixels indicate noise-free coefficients with magnitude above the previous threshold. Right: black pixels indicate noise-free coefficients with magnitude larger than noise deviation. This corresponds to *diagonal projection*: if an "oracle" (Donoho and Johnstone, 1994) tells us whether or not a coefficient is dominated by noise, this is the best thing we can do.



(a) (b)

FIGURE 4. Left: the result from thresholding with level- and subband-dependent GCV-threshold. Only the three finest scales were processed. SNR = 18.64 dB. Right: result from the optimal (clairvoyant) diagonal projection. SNR = 21.05 dB.

step tries to restore the eroded objects by turning white background pixels into black object pixels, if there is already an object in the neighborhood. (This neighborhood is typically a 3×3 box containing the actual pixel in its center.) Figure 5 (b) contains the result of this operation. It is hard to



(a) (b)

FIGURE 5. Result of elementary binary image enhancement methods on the approximate minimum MSE label image in Figure 2. Left: a median filter; Right: an erosion-dilation procedure.

preserve the fine edge structures, while removing the noisy pixels. These operations act on the label images only and forget about the background behind them: these pixels come from a wavelet coefficient classification. We would prefer a method that can deal with the geometry *and* the local

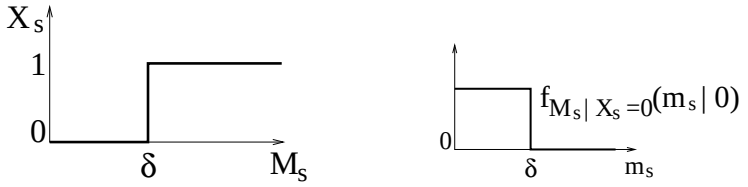


FIGURE 6. Left: The deterministic classification function for coefficient magnitude thresholding: if a coefficient magnitude M is below the threshold value δ, it is classified as noisy ($X = 0$), otherwise it is called sufficiently clean ($X = 1$). Right: this deterministic approach is a special case of the Bayesian model, where the conditional density is zero for coefficient magnitudes above the threshold if $X = 0$ and beneath the threshold if $X = 1$.

coefficient values at the *same* time. Bayes' rule tells us how we can do so.

3 A Bayesian procedure

3.1 A modified classification procedure

The classification in a threshold algorithm is a deterministic function of the empirical coefficients: thresholding on magnitudes corresponds to a simple step function, as illustrated in Figure 6. Recall that, in this text, the measure of significance is the coefficient magnitude: $\boldsymbol{M} = |\boldsymbol{W}|$. However, it would be interesting to examine measures based on interscale-correlations: e.g. $M^m_{j;s} = \prod_i W^m_{j+i;s}$, where j is the current resolution level and m the orientation ($m = \text{HOR}, \text{VER}, \text{DIAG}$). Because we want to take the spatial configurations into account, we give up this tight relation between a coefficient value and its classification.

We introduce a *prior model* for coefficient classification configurations $\boldsymbol{X}$. This prior should express the belief that clusters of noise-free coefficients have a higher probability than configurations with many isolated labels. In particular, edge-shaped clusters should be promoted.

Next, the *conditional model* (likelihood function) states that if the classification label for a coefficient equals one, this coefficient is *probably* above the threshold. A zero label means that the corresponding coefficient is probably small. The classical, deterministic approach can be seen as an extreme case of this probability model, where, for example, a label $X = 0$ tells that the coefficient is *certainly* in the range $[-\delta, \delta]$ (see right plot Figure 6).

If we have a prior distribution $P(\boldsymbol{X} = \boldsymbol{x})$ and a conditional model $f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$, then Bayes' rule allows to compute the *posterior* probability:

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}) f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})}{f_{\mathbf{M}}(\boldsymbol{m})}$$

In a given experiment, where $\boldsymbol{m}$ is fixed, the denominator is a constant.

As we explain later, it is sufficient to know the *relative* probabilities of configurations, and therefore we write:

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}) = C \cdot P(\boldsymbol{X} = \boldsymbol{x}) f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$$

3.2 The a priori model

As explained above, we are looking for a multivariate model for a binary image $\boldsymbol{X}$. Expressing a probability function for all 2^{NM} possible values in a N by M label image may be cumbersome. We therefore construct the model starting from local descriptions of clustering. First we write the prior probability function as:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z} \exp[-H(\boldsymbol{x})],$$

with the partition function Z:

$$Z = \sum_{\boldsymbol{x}} \exp[-H(\boldsymbol{x})].$$

$H(\boldsymbol{x})$ is the *energy function* of a configuration $\boldsymbol{x}$: the lower the energy, the higher the prior probability. To express that this energy comes from *local* interactions only, we first define for each pixel index s in the lattice S its set of neighbors, i.e. the set of indices $\partial s \subset S$ that interact with s. We assume that $s \notin \partial s$ and $s \in \partial t \Leftrightarrow t \in \partial s$. A set of indices that are all neighbors to each other is called a *clique*. The set of all cliques is

$$\mathcal{C} = \{C \in 2^S \mid \forall s, t \in C : t \in \partial s\}$$

The Gibbs Random Field (GRF) model (Besag, 1974; Geman and Geman, 1984; Winkler, 1995; Bello, 1994; Charbonnier et al., 1992; Luettgen et al., 1993) states that energy is the sum of clique potential functions:

$$H(\boldsymbol{x}) = \alpha \sum_{C \in \mathcal{C}} U_C(\boldsymbol{x}_C).$$

As the equation indicates, the clique potential U_C only depends on the values of $\boldsymbol{x}$ in the sites that belong to C.

A well-known example is the Ising model: the neighbors of a pixel with index s is a 3×3 submatrix, excluding s in its center. The model only considers pairs of neighbors. Other cliques in the system have no energy. The total energy is:

$$H(\boldsymbol{x}) = \sum_{\{s,t\} \in \mathcal{C}} x_s x_t$$

For our experiments, we use a slightly different model, in a 5×5-neighborhood system. We only consider 3×3 cliques, i.e. the largest possible

type in a 5 by 5 neighborhood system. The potentials of all other types of cliques are set to zero. For a 3 by 3 clique C we set:

$$U_C(\boldsymbol{x}_C) = \frac{\sum_{s\in C} x_s \sum_{t\in C\cap\partial s}(1-2x_s x_t)}{\sum_{s\in C} x_s},$$

i.e. for each $x_s = 1$, we subtract the number of neighbors within the clique with value one from the number of neighbors with zero value. The sum of these results is divided by the number of labels with value $x_s = 1$, to obtain a mean value. The idea behind this potential function is to compute a kind of "average degree of isolation" within the clique for the pixels with value one. The background pixels are considered to be neutral. Unlike the classical Ising model, this function is not symmetric for interchanges of ones and zeros.

3.3 The conditional model

We also need a conditional density $f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x})$. Whereas the prior describes the clustering of significant wavelet coefficients, this conditional model deals with the *local* significance measure. Therefore we write

$$f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x}) = \prod_{s\in S} f_{M_s|X_s}(m_s|x_s).$$

This density expresses that if the label $X_s = 1$, i.e. if the corresponding wavelet coefficient is sufficiently noise-free, a large value of M_s is probable. Figure 7 shows how this density uses a threshold parameter δ. Another parameter is the transition width Δ between the high and the low density level. This parameter expresses the uncertainty of the classification. For instance, if $X_s = 1$, the coefficient is important, and its significance is normally above the threshold. Due to noise, it may happen to be smaller than the threshold. The more noise, the more probable this event. We therefore choose:

$$\Delta \sim \sigma$$

Since in practical cases, σ is unknown, we replace it with an estimate or another deviation-like quantity. For example, if we assume that most of the coefficients below the threshold are pure noise, we can compute the total amount of energy in these coefficients as a measure for the noise variance.

We can write:

$$f_{M_s|X_s}(m_s|x_s) = \frac{1}{Y_s}\exp[-\beta V_s(m_s|x_s)],$$

where Y_s is a normalization factor and $V_s(m_s|x_s)$ is a conditional potential function, such as illustrated in figure 8. The conditional density now

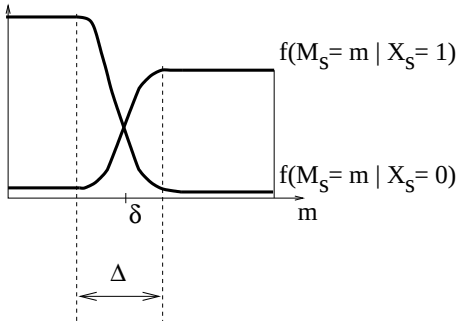


FIGURE 7. Conditional density $f(M_s|X_s)$. If $X_s = 1$, then large values of M_s are probable and vice versa. This function uses a threshold δ as a parameter.

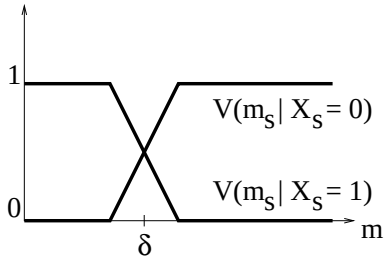


FIGURE 8. Conditional potential function.

becomes:

$$f_{\mathbf{M}|\mathbf{X}}(\boldsymbol{m}|\boldsymbol{x}) = \frac{1}{Y}\exp[-V(\boldsymbol{m}|\boldsymbol{x})],$$

where

$$Y = \int_{\boldsymbol{m}} \exp[-V(\boldsymbol{m}|\boldsymbol{x})]\, d\boldsymbol{m} = \prod_s Y_s,$$

and

$$V(\boldsymbol{m}|\boldsymbol{x}) = \beta \sum_{s\in S} V_s(m_s|x_s).$$

3.4 Posterior probabilities

From Bayes' rule, we can compute the posterior probabilities

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}).$$

With these probabilities, a Bayesian decision rule leads to an estimation of the optimal label. The Maximum A Posteriori (MAP) procedure chooses the mask $\boldsymbol{x}$ with the highest posterior probability. The Maximal Marginal Posterior (MMP) rule is a more local approach: it computes in each site s

the marginal probabilities:

$$P(X_s = 1|\boldsymbol{M} = \boldsymbol{m}) = \sum_{\boldsymbol{x}} x_s P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m}),$$

and if this probability is more than 0.5, the pixel gets value 1. Both decision rules have a binary outcome: each coefficient is classified as noisy ($X = 0$) or relatively uncorrupted ($X = 1$). We would like to exploit the entire posterior probability: the posterior mean value

$$E(X_s|\boldsymbol{M} = \boldsymbol{m}) = P(X_s = 1|\boldsymbol{M} = \boldsymbol{m})$$

preserves all information. It is a minimum least squares estimator. This classification leads to a posterior 'expected action':

$$\begin{aligned}\mathrm{E}(W_{\delta s}|\boldsymbol{M}) = {} & h(W_s, m_s(\boldsymbol{W}), 1)P(X_s = 1|\boldsymbol{M}) \\ & + h(W_s, m_s(\boldsymbol{W}), 0)P(X_s = 0|\boldsymbol{M}).\end{aligned}$$

If $h(W_s, M_s, X_s) = X_s W_s$, this is:

$$\mathrm{E}(W_{\delta s}|\boldsymbol{M}) = W_s \mathrm{E}(X_s|\boldsymbol{M} = \boldsymbol{m}) = W_s P(X_s = 1|\boldsymbol{M}).$$

Unlike most thresholding methods, this is not a binary procedure: using the posterior probability leads to a more continuous approach.

3.5 Stochastic sampling

The computation of $P(X_s = 1|\boldsymbol{W})$ involves the probability of all possible configurations $\boldsymbol{x}$. Because of the enormous number of configurations, this is an intractable task. The sum we have to compute is of the following form:

$$\mu_s = \sum_{\boldsymbol{x}} f_s(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M} = \boldsymbol{m})$$

where in this case $f_s(\boldsymbol{x}) = x_s$ and $\mu_s = P(X_s = 1|\boldsymbol{M} = \boldsymbol{m})$.

To estimate this type of sum (or integral for random variables on a continuous line), one typically uses stochastic samplers. These methods generate subsequent samples $\boldsymbol{X}^{(i)}$, not selected uniformly, but in proportion to their probability. This allows to approximate the matrix of required marginal probabilities by the mean value of the generated masks:

$$\hat{\mu}_s = \sum_i X_s^{(i)}.$$

Mostly, the samples are generated, not independently of each other, but in a chain, hence the name Markov Chain Monte Carlo (MCMC) estimation.

The next sample is generated, starting from the previous one. One advantage of this procedure is that knowledge of the *relative* probabilities of the candidates is sufficient. The probability ratio of two subsequent samples:

$$r^{(i)} = \frac{P(\boldsymbol{X}^{(i+1)}|\boldsymbol{M})}{P(\boldsymbol{X}^{(i)}|\boldsymbol{M})}$$

is the only quantity needed by the algorithm, as we explain below, and if

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{M}) \sim \frac{1}{Z}\exp[-H(\boldsymbol{x})]\frac{1}{Y}\exp[-V(\boldsymbol{m}|\boldsymbol{x})],$$

there is no need for the enormous computation of the partition function ZY.

We use the classic Metropolis MCMC sampler. The chain of states is started from an initial state $\boldsymbol{X}^{(0)}$. The successive samples $\boldsymbol{X}^{(i)}$ are then produced as follows: a candidate intermediate state is generated by a local random perturbation of the actual state. Then the probability ratio r of the actual state and its perturbation is computed. Since the GRF-model is based on local potential functions, only positions s whose mask labels are switched by the perturbation or which have a switched label in their neighborhood ∂s are involved in the computation. If the candidate has a higher probability than the actual state, i.e. if the probability ratio is larger than one, then the new state is accepted, otherwise it is accepted with probability equal to r. To generate a completely new sample, we repeat this local switching for all locations in the grid.

3.6 Algorithm overview

This is a schematic overview of the subsequent steps of the algorithm:

1. Compute the stationary wavelet transform $\boldsymbol{W}$ of the input.

2. At each level and for each component, select the appropriate threshold. This threshold generates an initial label image $\boldsymbol{X}^{(0)}$.

3. Run a stochastic sampler to estimate for each coefficient at the given resolution level the probability $P(X_s|\boldsymbol{W})$. Use $\boldsymbol{X}^{(0)}$ from the previous step as the starting sample. A Markov Chain Monte Carlo algorithm produces the sequence of samples.

4. $\hat{W}_{\delta s} \leftarrow W_s P(X_s = 1|\boldsymbol{W})$.

5. Inverse wavelet transform yields the result.

4 Results and discussion

The expression of the posterior energy:

$$H(\boldsymbol{x}|\boldsymbol{m}) = H(\boldsymbol{x}) + V(\boldsymbol{m}|\boldsymbol{x}) + C^*,$$

where the constant C^* comes from the denominator in Bayes' rule and where:

$$H(\boldsymbol{x}) = \alpha \sum_{C \in \mathcal{C}} U_C(\boldsymbol{x}_C)$$

and

$$V(\boldsymbol{m}|\boldsymbol{x}) = \beta \sum_{s \in S} V_s(m_s|x_s),$$

leaves us with two parameters α and β. They determine the rigidity of prior and conditional model. The higher these numbers, the larger the energy difference between the two states. If the values are low, moving from one state to another is easy: a low β value indicates that the threshold in the conditional model is not a very strict boundary, while a small α means that clustering is not so important. The choice $\alpha = 0$ disregards spatial structures. It only considers the uncertainty in the conditional model, and leads to a kind of 'smoothed hard-thresholding'.

From our numerical experiments, we learned that the best results were obtained for $\alpha \approx 2$ and $\beta \approx 90$. Ten MCMC iterations were sufficient to obtain the labels in Figure 9, provided we use the labels of the threshold procedure as the initial state. More iterations (up to 100) did not improve the output quality. We used the minimum GCV-threshold. This image comes close to the labels obtained by applying the minimum MSE threshold to the noise-free coefficients in Figure 3(a). The right picture is the corresponding result of the algorithm, applied to all subbands in the three finest resolution levels. Signal-to-noise ratio is 18.47 dB.

Figure 10(a) illustrates the influence of the prior if β is set to 1 (instead of $\beta = 90$). The prior indeed causes edge-like structures, but without the aid of the conditional, these structures do not show a strong correspondence to the important image features.

For a good parameter choice, the resulting label image "comes close to" the mask obtained by applying the threshold to the noise-free coefficients. However, since this example shows a *decrease* in signal-to-noise ratio, one could wonder if starting a from threshold value, different from the minimum MSE threshold, could yield better results. Choosing the threshold $\delta = \sigma$ should lead to an approximation of the ideal diagonal projection method of Figure 3(b). The result appears in Figure 10(b). Of course, we do not know the exact value of σ in each subband, so we assume that the energy of the coefficients below the threshold is a good measure for the noise variance. SNR is now 18.78dB.

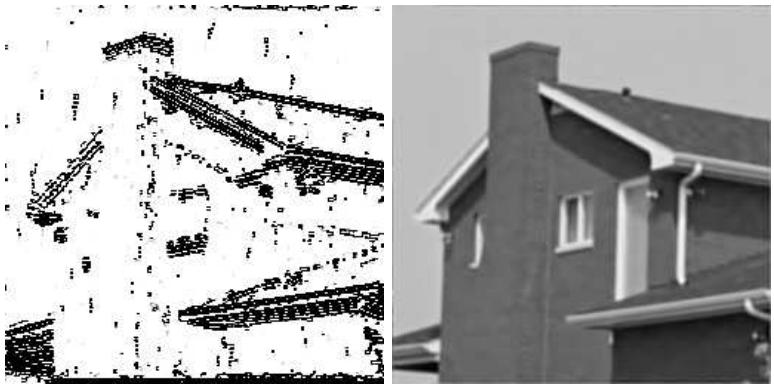
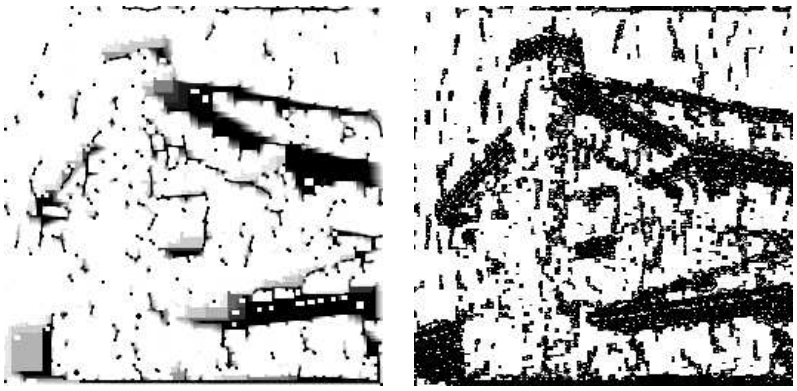


FIGURE 9. Left: label image after ten MCMC iterations. Consequently, this image has 11 grey values. A pixel value is an estimate of the marginal posterior probability $P(X_s = 1|\boldsymbol{M})$. We used the minimum GCV-threshold. Right: the algorithm output. Three resolution levels were processed. Signal-to-noise ratio is 18.47 dB.



(a) (b)

FIGURE 10. (a) The same label image as in the previous case, but now with conditional parameter $\beta = 1$ instead of $\beta = 90$. The influence of the prior causes edge-like structures, but without the aid of the conditional, these structures correspond too little with the really important features. (b) The labels of the horizontal component at the one but finest scale if we take $\delta = \sigma$.

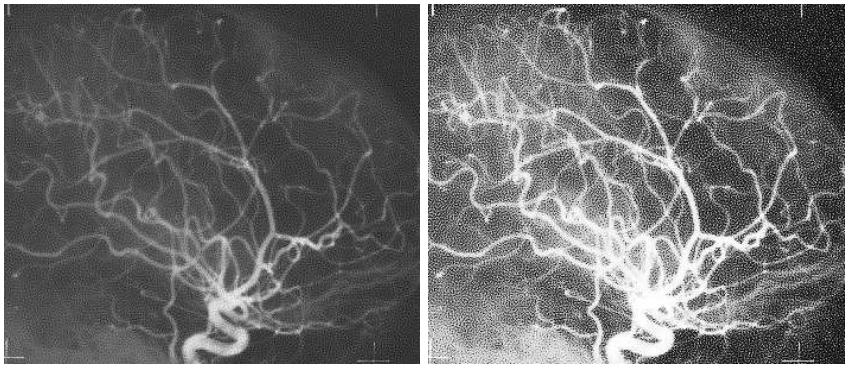


FIGURE 11. Digital subtraction angiogram of cerebral blood vessels with artificially added noise. SNR = 4.58dB. The picture on the right is the result of a contrast enhancement technique, called histogram equalization.

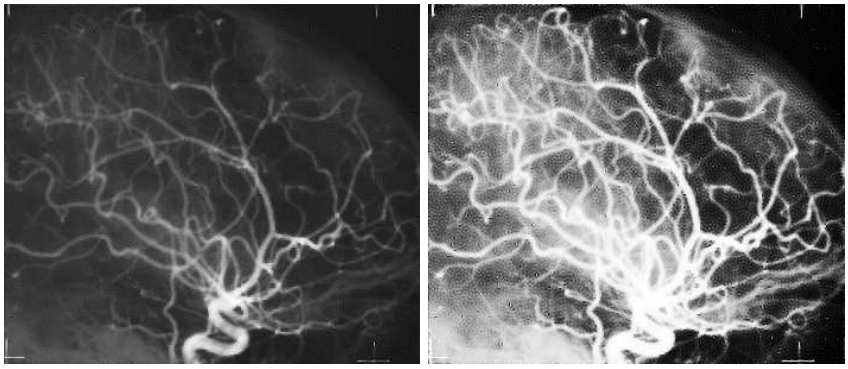


FIGURE 12. Result of classic threshold algorithm on DSA-image. SNR = 17.02dB.

Several possible improvements still need to be investigated. First, the choice of the hyperparameters α and β could maybe done on empirical basis. Also, these hyperparameters could be made subband-dependent: label images at different subbands show different characteristics. In general, there is an everlasting search for appropriate priors: priors based on local potential functions are attractive because of their easy computation, but at the same time, a prior should be selective on a more global region, to distinguish edges and other image features. A threshold selection based on interscale correlations could deal with interlevel dependencies.

We now apply the procedure to a DSA-image (Digital Subtraction Angiography) of cerebral blood vessels, with artificially added noise. The SNR value of the input image in Figure 11 is 4.58dB. Because of the fine image structure, we operate only on the highest resolution level. A classical threshold algorithm leads to Figure 12 with SNR = 17.02dB. A Bayesian

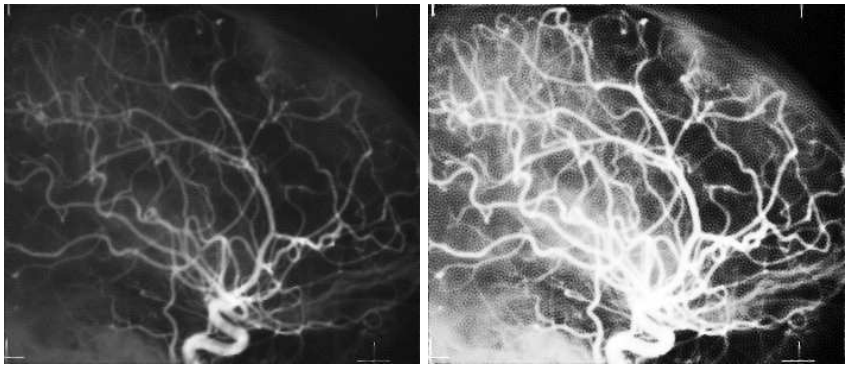


FIGURE 13. Result of Bayesian procedure. The initial labels were the threshold labels from the previous figure. SNR = 17.32dB.

correction on this image yields the result in Figure 13. SNR = 17.32dB. Some of the vessels are clearly better reproduced.

5 Conclusions

This chapter has introduced a geometrical prior model for configurations of important wavelet coefficients. This model can be used in combination with threshold algorithms. The threshold defines an initial binary "label image", where labels $X_s = 1$ indicate coefficients with large magnitude. Unlike more heuristic approaches, like erosion-dilation methods, Bayes' rule leads to a mathematically based procedure that cleans up the label images, taking into account prior beliefs and observed coefficients at the same time. The output are the posterior probabilities of the coefficients being sufficiently noise-free. These probabilities can be used to replace the binary threshold procedure with a more continuous approach. Although results are promising, further research for priors and for optimal hyperparameters remains necessary.

Acknowledgement

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors. The first author is financed by a grant from the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT).

6 References

Abramovich, F., Sapatinas, F., and Silverman, B. W. (1997). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 58.

Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image coding using the wavelet transform. *IEEE Transactions on Image Processing*, 1(2):205–220.

Bello, M. G. (1994). A combined Markov Random Field and wave-packet transform-based approach for image segmentation. *IEEE Transactions on Image Processing*, 3(6):834–846.

Berkner, K. and Wells, Jr., R. O. (1998). Correlation-dependent model for denoising via nonorthogonal wavelet transforms. Technical report, C. M. L., Dept. Mathematics, Rice University.

Besag, J. E. (1974). Spatial interaction and the spatial analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236.

Charbonnier, P., Blanc-Féraud, L., and Barlaud, M. (1992). Noisy image restoration using multiresolution Markov Random Fields. *Journal of Visual Communication and Image Representation*, 3(4):338–346.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* To Appear.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85:391–402.

Cohen, A., Daubechies, I., and Feauveau, J. (1992). Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560.

Crouse, M. S., Nowak, R., and Baraniuk, R. G. (1998). Wavelet-based signal processing using Hidden Markov Models. *IEEE Transactions on Signal Processing*, Special Issue on Wavelets and Filterbanks. To Appear.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* CBMS-NSF Regional Conf. Series in Appl. Math., Vol. 61. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Jansen, M. and Bultheel, A. (1998). Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Transactions on Image Processing.* to appear.

Jansen, M., Malfait, M., and Bultheel, A. (1997). Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, To Appear.

Luettgen, M. R., Karl, W. C., Willsky, A. S., and Tenney, R. R. (1993). Multiscale representations of Markov Random Fields. *IEEE Transactions on Signal Processing*, 41(12):3377–3395.

Malfait, M. (1995). Using wavelets to suppress noise in biomedical images. In Aldroubi, A. and Unser, M., editors, *Wavelets in Medicine and Biology,* Chapter 8, pages 191–208. CRC Press.

Malfait, M. and Roose, D. (1997). Wavelet based image denoising using a Markov Random Field a priori model. *IEEE Transactions on Image Processing*, 6(4):549–565.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In Antoniadis, A. and Oppenheim, G., editors, *Wavelets and Statistics*, Lecture Notes in Statistics, pages 281–299.

Pesquet, J.-C., Krim, H., and Carfantan, H. (1996). Time invariant orthonormal wavelet representations. *IEEE Transactions on Signal Processing*, 44(8):1964–1970.

Ruggeri, F. and Vidakovic, B. (1995). A Bayesian decision theoretic approach to wavelet thresholding. Preprint 95-35, Duke University, Durham, NC.

Simoncelli, E. P. and Adelson, E. (1996). Noise removal via Bayesian wavelet coring. In *proceedings 3rd International Conference on Image Processing.*

Vidakovic, B. (1994). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. Preprint 94-24, Duke University, Durham, NC.

Weyrich, N. and Warhola, G. T. (1995). De-noising using wavelets and cross validation. In Singh, S., editor, *Approximation Theory, Wavelets and Applications*, volume 454 of *NATO ASI Series C*, pages 523–532.

Winkler, G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods.* Applications of Mathematics. Springer.

Xu, Y., Weaver, J. B., Healy, D. M., and Lu, J. (1994). Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE Transactions on Image Processing*, 3(6):747–758.