A Joint Inter- and Intrascale Statistical Model for Bayesian Wavelet Based Image Denoising

Aleksandra Pižurica, Wilfried Philips, Ignace Lemahieu, and Marc Acheroy

***Abstract*—This paper presents a new wavelet-based image denoising method, which extends a recently emerged "geometrical" Bayesian framework. The new method combines three criteria for distinguishing supposedly useful coefficients from noise: coefficient magnitudes, their evolution across scales and spatial clustering of large coefficients near image edges. These three criteria are combined in a Bayesian framework. The spatial clustering properties are expressed in a prior model. The statistical properties concerning coefficient magnitudes and their evolution across scales are expressed in a joint conditional model. The three main novelties with respect to related approaches are: (1) the interscale-ratios of wavelet coefficients are statistically characterized, and different local criteria for distinguishing useful coefficients from noise are evaluated; (2) a joint conditional model is introduced, and (3) a novel anisotropic Markov Random Field prior model is proposed. The results demonstrate an improved denoising performance over related earlier techniques.**

***Index Terms*—Image denoising, interscale ratios, Markov random field, statistical modeling, wavelets.**

I. Introduction

In image denoising, a compromise has to be found between noise reduction and preserving significant image details. To achieve a good performance in this respect, a denoising algorithm has to adapt to image discontinuities. The wavelet representation naturally facilitates the construction of such spatially adaptive algorithms. It compresses the essential information in a signal into relatively few, large coefficients; these large coefficients correspond to the main image details at different resolution scales. Due to this property, additive noise can be effectively suppressed even by simple thresholding [7] of the wavelet coefficients. Bayes estimation assuming independent wavelet coefficients [25], [2], [22] results also in relatively simple shrinkage of their magnitudes.

Advanced wavelet based denoising methods make use of image context [1], interscale dependencies [20], [30], [10], [5], [24] or intrascale (spatial) correlations [13], [26], [11], [12] between image wavelet coefficients. Combining inter- and intrascale dependencies in a *decimated*, orthonormal wavelet basis is known to provide only minor improvements in estimation performance [15]. However, in a *nondecimated* wavelet basis, such a combination has clear advantages in terms of both quantitative image quality measures and visual quality of the results, as was demonstrated, e.g., in [17]. The method of [17] combines the inter and intrascale dependencies in a powerful and elegant way: a bilevel Markov Random Field (MRF) model is used to encode prior knowledge about spatial clustering of wavelet coefficients, i.e., to encode "geometrical properties" of image details. The interscale dependencies between wavelet coefficients are encountered via interscale ratios, which appear as significance measures. The statistical properties of these significance measures are expressed in a conditional probability density model, and combined with the prior model in a Bayesian framework. Hence the name "geometrical Bayesian approach". The conditional model from [17] is heuristic and parameterized, which complicates its practical implementation. The authors of [11], [12] further motivate the whole approach theoretically and develop practical algorithms from it. However, they do not employ interscale statistics, but instead consider the magnitude of a wavelet coefficient as its significance measure, and derive corresponding, realistic conditional models.

In this paper, we further extend the geometrical Bayesian approach. The three main novelties are (1) a statistical characterization of different significance measures for wavelet coefficients, and a comparative evaluation of their performance; (2) a joint conditional model, which combines local inter- and intrascale statistical properties and (3) a novel, anisotropic MRF prior model. We start from different interscale-ratio formulations, [17], [10], and present them in a unifying and a more general way. Instead of using earlier heuristic models for such ratios, we determine empirically their realistic conditional probability densities given pure noise, and given noisy edges. Using the empirical densities and employing a statistical decision theory, we compare the performance of different significance measures. Such a performance evaluation clearly motivates a joint significance measure, which relies on both coefficient magnitudes and on their evolution across scales. The resulting, joint conditional model offers a superior denoising performance with respect to earlier ones that use interscale ratios only [17], or coefficient magnitudes only [11], [12]. A new anisotropic prior model preserves well finest image details. As compared to the isotropic one from [17], it introduces minor increase in complexity, but preserves image details significantly better.

The paper is organized as follows. In Section II, back-

A. Pizurica* and W. Philips are with the Department for Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium. E-mail: Aleksandra.Pizurica@telin.rug.ac.be, philips@telin.rug.ac.be, Tel: +32 9 264 34 12, Fax: +32 9 264 42 95

I. Lemahieu is with the Department for Electronics and Information Systems (ELIS/MEDISIP), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium. E-mail: ignace.lemahieu@rug.ac.be, Fax: 32-9-264.35.94, Tel: 32-9-264.42.32

M. Acheroy is with the Royal Military Academy, Av. de la Renaissance 30, B-1000 Brussels, Belgium. E-mail: acheroy@elec.rma.ac.be, Fax: +32 2 737 64 72, Tel.: + 32 2 737 64 70

The research is funded by the Belgian Ministry of Defense, in the scope of the Humanitarian Demining project, HUDEM.

ground concepts are briefly reviewed. A simulation method for characterizing statistical properties of different significance measures is presented in Section III, and a performance evaluation in Section IV. The proposed joint significance measure and the practical implementation of the conditional model are described in Section V. In Section VI, a new prior model is proposed. The results are presented and discussed in Section VII. Finally, concluding remarks are given in Section VIII.

II. Background

A. Wavelet decomposition and notation

For a comprehensive treatment of the wavelet transform we refer to [3], [6], [18], [21], [28]. In a wavelet decomposition of an image a wavelet coefficient $w_{j,l}^D$ represents its bandpass content at resolution scale 2^j, spatial position l and orientation D. Typically, three orientation subbands are used, leading to three *detail images* at each scale, characterized by horizontal, vertical and diagonal directions. A decomposition from [19] results in *two* detail images at each scale; these details can be interpreted as partial derivatives of smoothed image in horizontal and vertical directions. This decomposition is often used for edge detection, local regularity characterization and denoising [20], [10]. We use it for the results throughout the paper. In Section VII, we also present results for a classical 3-subband decomposition, which was implemented with the same quadratic spline wavelet from [19], like in [17], [30]. In both cases, we use a non-decimated representation [4], [11], [17]. It facilitates use of interscale dependencies, and offers a better denoising performance than orthonormal bases. The improvement is achieved at the expense of computing time: more coefficients need to be stored and processed. In this respect, the 2-subband decomposition is more attractive.

Whenever there can be no confusion, we omit the indices that denote the scale and the orientation. Random variables are denoted by capital letters and their realizations by the corresponding small letters. Boldface letters are used for vectors. A given detail image is thus represented as vector $\mathbf{w} = \{w_1, ..., w_n\}$. The set of indices $L = \{1, ..., n\}$ is a set of pixels on a regular rectangular lattice. In the spirit of [11], [12], [17], we assign a measure of significance m_l, and a binary label x_l to each wavelet coefficient w_l. The label value $x_l = -1$ denotes that w_l represents mainly noise, and the value $x_l = 1$ denotes that w_l is a "significant" coefficient. The significance measure can be, e.g., the magnitude of a coefficient, the local Lipschitz exponent [21] or the interscale correlation [30] at the corresponding location, etc. A set $\mathbf{m} = \{m_1, ..., m_n\}$ is called *significance map*, and a set $\mathbf{x} = \{x_1, ..., x_n\}$ is called *mask*. We assume that the wavelet coefficients are corrupted with additive white Gaussian noise. An observed wavelet coefficient w_l is thus given by $w_l = y_l + n_l$, where y_l is the noise-free wavelet coefficient and n_l are independent identically distributed (i.i.d.) random variables. The magnitude of a wavelet coefficient is denoted by $\omega_l = |w_l|$.

B. Spatial dependencies and MRF

A detailed analysis of MRF models can be found elsewhere, e.g., in [14], [29]. Here we briefly review the basic concepts of these models that are of interest for our denoising method. Let $L \setminus l$ denote the set of all pixels in L except l itself. The Markov property of a random field $\mathbf{X}$ is $P(X_l = x_l | \mathbf{X}_{L\setminus l} = \mathbf{x}_{L\setminus l}) = P(X_l = x_l | \mathbf{X}_{\partial(l)} = \mathbf{x}_{\partial(l)})$, where $\partial(l)$ is the *neighborhood* of the pixel l. Most often used are the first-order neighborhood (four nearest pixels) and the second-order neighborhood (eight nearest pixels). A set of pixels, which are all neighbors of one another is called a *clique*. For example, for the first-order neighborhood cliques consist of one or two pixels, and for the second order neighborhood cliques consist of up to four pixels. The joint probability $P(\mathbf{X} = \mathbf{x})$ of a MRF is a special case of the Gibbs distribution, $\exp(-H(\mathbf{x})/T)/Z$, with partition constant Z and temperature T, where the energy $H(\mathbf{x})$ can be decomposed into contributions of *clique potentials* $V_C(\mathbf{x})$ over all possible cliques. The clique potential $V_C(\mathbf{x})$ is a function of *only* those labels x_l, for which $l \in C$. In practice, one chooses the appropriate clique potential functions to give preference to certain local spatial dependencies, e.g., to assign higher prior probability to edge continuity. In [17], the isotropic auto-logistic MRF model, with the second order neighborhood N_l, was applied to binary masks. In this case, and for the label set $\{-1, 1\}$ that we use here, the joint probability of the MRF becomes[1]

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\Big(-\sum_{l \in L} V_{N_l}(\mathbf{x})\Big), \quad V_{N_l}(\mathbf{x}) = -\gamma \sum_{k \in N_l} x_l x_k, \tag{1}$$

where γ is a positive constant. In the remainder of this text, we call the MRF model from Eq (1), the isotropic MRF model. In Section V, we shall introduce a new, *anisotropic* model, which adapts to the presence of micro-edges in a given neighborhood. Another prior model, which uses a different reasoning, and adapts to an average degree of isolation of labels 1 was proposed in [12].

C. A probabilistic shrinkage approach

We shall extend the probabilistic shrinkage approach of [17]. Each wavelet coefficient is multiplied with the marginal probability that it is noise-free, given the computed significance map:

$$\hat{y}_l = P(X_l = 1 | \mathbf{M} = \mathbf{m}) w_l. \tag{2}$$

A heuristic motivation for this rule is: the shrinkage factor $P(X_l = 1 | \mathbf{M} = \mathbf{m})$ is always between zero and one, and it suppresses more those coefficients that are less likely to represent a useful signal, given the significance map $\mathbf{m}$ for the whole detail image. A theoretic motivation in the form of an "expected posterior action" was given in [11], [12]. The exact computation of the marginal probability $P(X_l = 1 | \mathbf{M} = \mathbf{m})$ requires the summation of the

[1] In [17], the set of labels was $\{0, 1\}$, and this expression was represented in a less compact way, but it is only a matter of notation.

posterior probabilities $P(\mathbf{X} = \mathbf{x}|\mathbf{M} = \mathbf{m})$ of all possible configurations $\mathbf{x}$ for which $x_l = 1$. Since this is an intractable task, one typically alleviates it by using a relatively small, but "representative" subset of all possible configurations. Such a representative subset is obtained via an importance-type sampling: the probability that a given mask is sampled is proportional to its posterior probability. An estimate of $P(X_l = 1|\mathbf{M} = \mathbf{m})$ is then obtained by computing the fractional number of all sampled masks for which $x_l = 1$. An importance sampling for Markov models is typically realized via Monte Carlo Markov Chain (MCMC) [16] samplers, such as the Metropolis [14] and the Gibbs sampler [9]. We apply the Metropolis sampler. From each configuration $\mathbf{x}$, it generates a new, "candidate" configuration $\mathbf{x}^c$ by switching the binary label at a randomly chosen position l. The decision about accepting the change is based on the ratio p of the posterior probabilities of the two configurations. After applying the Bayes rule $P(\mathbf{X} = \mathbf{x}|\mathbf{M} = \mathbf{m}) = P(\mathbf{X} = \mathbf{x}) f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}|\mathbf{x})/f_{\mathbf{M}}(\mathbf{m})$, assuming the conditional independence[2] as in [11], [12], [17] $f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}|\mathbf{x}) = \prod_{l\in L} f_{M_l|X_l}(m_l|x_l)$, and a MRF prior model from Eq (1), the ratio p reduces to

$$p = \frac{f_{M_l|X_l}(m_l|x_l^c)}{f_{M_l|X_l}(m_l|x_l)} \exp\Big(V_{N_l}(\mathbf{x}) - V_{N_l}(\mathbf{x}^c)\Big). \qquad (3)$$

If $p > 1$ the local change is accepted, because it has produced a mask with a higher posterior probability. If $p < 1$, the change is accepted with probability p. After all labels in the mask have been updated in this way, one iteration is completed. It was found in [11], [17] that 10 such iterations suffice to estimate the marginal probabilities, provided that the initial mask is well chosen.

III. Choice of significance measures and statistical characterization

For the performance of the above described denoising approach the choice of the significance measure m_l and the characterization of its conditional densities $f_{M_l|X_l}(m_l|x_l)$ are important. In [17], m_l was defined as an averaged interscale ratio of the magnitudes of wavelet coefficients, but a *heuristic* model was used for its conditional densities. In [11] and [12], the magnitude of the wavelet coefficient was used instead, and its realistic conditional densities were considered. To our knowledge, there has been no attempt to objectively compare the performance of these two and possibly other significance measures m_l, neither to investigate realistic densities of interscale ratios.

Two different formulations of interscale ratios have been used in recent literature with the same intention: to estimate roughly the local regularity of an image and accordingly to make a distinction between useful edges and noise. In [17], the ratio of the magnitudes of wavelet coefficients was averaged over a certain number d of resolution scales: $m_l = (1/d)\sum_{j=1}^{d} |w_{j+1,l}/w_{j,l}|$; it is a rough estimate of 2^α, where α is the local Lipschitz exponent [21]. Another formulation was given in [10]. There, the authors first define the integral $\mathcal{N}f(s, x_0) = \int_{|x-x_0|\le Ks} |\mathcal{W}f(s, x)|dx$, where $\mathcal{W}f(s, x)$ is the wavelet transform at resolution scale s, K is the support width of the mother wavelet and $|x-x_0| \le Ks$ defines the cone of influence of the point x_0 in the scale-space (s, x). The ratio $\mathcal{N}f(2^{j+1}, x_0)/\mathcal{N}f(2^j, x_0)$ was proved to be an estimate of $2^{\alpha+1}$ and was used as a significance measure for the wavelet coefficients at resolution scale 2^j. The statistical properties of these ratios were not investigated and also their averaging through scales was not considered.

A. A discretized approximation of interscale ratios

We formulate the two mentioned interscale ratio measures from [17] and [10] in a unifying and slightly more general way. Let us define for each spatial position l the coefficient $\alpha_{n\to k,l}$ which determines the average rate of increase of the magnitudes of the wavelet coefficients between any two dyadic scales 2^n and 2^k, where $n, k \in Z$ and $k \ge n + 1$

$$\alpha_{n\to k,\,l} \triangleq \log_2\Big(\frac{1}{k-n}\sum_{j=n}^{k-1}\frac{|w_{j+1,l}|}{|w_{j,l}|}\Big). \qquad (4)$$

The logarithm in Eq. (4) is used in order to make $\alpha_{n\to k,l}$ behave as a rough estimate of the local Lipschitz exponent α. This quantity describes the evolution of the individual wavelet coefficients at the spatial position l. Similarly we define a second quantity $\beta_{n\to k,l}$, which describes the "collective" evolution of the wavelet coefficients inside a cone of influence centered at the spatial position l. We denote by $C(j, l)$ the discrete set of wavelet coefficients at the resolution scale 2^j, which belong to the cone of influence of the spatial position l, and we define $\beta_{n\to k,l}$ as

$$\beta_{n\to k,\,l} \triangleq \log_2\Big(\frac{1}{k-n}\sum_{j=n}^{k-1}\frac{|I_{j+1,l}|}{|I_{j,l}|}\Big), \quad I_{j,l} \triangleq \sum_{m\in C(j,l)} |w_{j,m}|, \qquad (5)$$

which is a rough estimate of $\alpha + 1$. We call the quantity $\alpha_{n\to k,l}$ the *average point ratio* (APR) and the quantity $\beta_{n\to k,l}$ the *average cone ratio* (ACR).

B. Statistical characterization via simulation

We apply a simulation method to determine the empirical conditional densities $f_{M_l|X_l}(m_l|x_l)$, of any chosen significance measure m_l and arbitrary type of noise. Practically we restrict ourselves to additive white Gaussian noise. The benefit of knowing the empirical densities is twofold. Firstly, the performance of different significance measures can be objectively compared relying on the statistical estimation theory [27]. Secondly, the empirical densities can be employed in an actual denoising procedure instead of heuristic models.

To determine $f_{M_l|X_l}(m_l|1)$ one needs a statistical model for the actual significant image discontinuities, i.e., the "ground truth" edges. To achieve realistic results, we extract these discontinuities from various natural noise-free

[2] Since an undecimated wavelet transform is assumed, the conditional independence assumption apparently cannot hold. However, in the context of the proposed approach, this simplification greatly reduces the computational effort and yields good results.

Fig. 1. Left: reference images: 1 - "Lena", 2 - "Goldhill", 3 - "Fruits", 4 - "Barbara". Right: reference edge positions for vertical orientation of details at resolution scale 2^2.

images, like those in Fig. 1. For a given reference image, at each resolution scale and orientation the reference edge positions are obtained by thresholding the magnitudes of noise-free wavelet coefficients. For characterizing densities of interscale ratios, the choice of these thresholds is not critical. In particular, we choose a threshold for each resolution scale that is equal to the magnitude of a wavelet coefficients produced by an ideal step edge of amplitude A, where A equals to 1/16 of the dynamic range of the image.

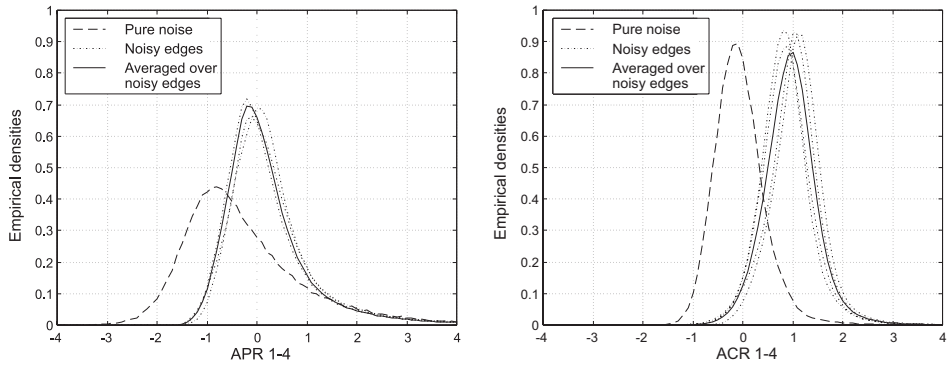
The complete simulation procedure is the following. To estimate $f_{M_l|X_l}(m_l|-1)$, we use the image which consists of pure noise. We compute the discrete wavelet transform of this image and find the histogram of the given significance measure m_l. To estimate $f_{M_l|X_l}(m_l|1)$, we apply a similar procedure, except that the noise is added to the reference image and m_l is computed only in the reference positions of significant coefficients. The whole procedure is iteratively repeated, adding every time random noise with equal variance. Therefore, we refer to this method as simulation.

1. Conditional densities of APR and ACR. The conditional densities $f_{M_l|X_l}(\alpha_{n\to k,l}|-1)$ and $f_{M_l|X_l}(\beta_{n\to k,l}|-1)$, of APR and ACR, respectively, given pure noise do not depend on the noise variance: the increase of noise affects equally the magnitudes of the wavelet coefficients at all scales and the statistical distribution of their averaged ratios does not change. These conditional densities, shown as dashed curves in Fig. 2(a) and Fig. 2(b), are peaked in the vicinity of -1 and 0, respectively. The empirical densities $f_{M_l|X_l}(\alpha_{n\to k,l}|1)$ and $f_{M_l|X_l}(\beta_{n\to k,l}|1)$, for different reference edges from Fig. 1 are also illustrated in Fig. 2(a) and Fig. 2(b), and are peaked in the vicinity of 0 and 1, respectively. We have verified that the choice of a threshold that specifies the reference edges in Fig. 1 is not critical for characterizing these densities.

It should be noticed that the overlap between $f_{M_l|X_l}(\beta_{n\to k,l}|-1)$ and $f_{M_l|X_l}(\beta_{n\to k,l}|1)$ is smaller than the overlap between $f_{M_l|X_l}(\alpha_{n\to k,l}|-1)$ and $f_{M_l|X_l}(\alpha_{n\to k,l}|1)$. It suggests that ACR provides a better separation between noise and useful edges. We examine this further in Section IV. The robustness of interscale ratio statistics with respect to noise level is illustrated in the top of Fig. 3. One can see that the overlap between conditional densities of ACR given noise and given edges does not change much as the noise variance increases.

2. Conditional densities of the magnitudes of the wavelet coefficients. The bottom of Fig. 3 illustrates densities $f_{M_l|X_l}(m_l|x_l)$, for the case where m_l is the magnitude of the wavelet coefficient, which we denote by ω_l. There is no need to simulate $f_{M_l|X_l}(\omega_l|-1)$ since it is simply equal to $(2/\sigma_n\sqrt{(2\pi)})\exp(-\omega_l^2/(2\sigma_n^2))$, $\omega_l > 0$, where σ_n is the standard deviation of noise in the given detail image. For $f_{M_l|X_l}(\omega_l|1)$, a realistic model to be used in the denoising procedure can be computed directly from the noisy image itself (Section V). For comparative performance analysis with respect to interscale ratios, we use empirical densities $f_{M_l|X_l}(\omega_l|1)$.

It is interesting to compare how the overlap between the conditional densities of ω_l in Fig. 3 evolves with the increase of noise level, with respect to that of ACR that is shown in the same figure. A simple visual inspection suggests that for relatively small standard deviations of noise the coefficient magnitude provides a better separation between noise and useful signal, whereas the opposite



(a) (b)

Fig. 2. Conditional densities of (a) APR and (b) ACR, computed from scales $2^1 - 2^4$. Different reference edges from Fig. 1 are used. The standard deviation of added noise is $\sigma = 25$.

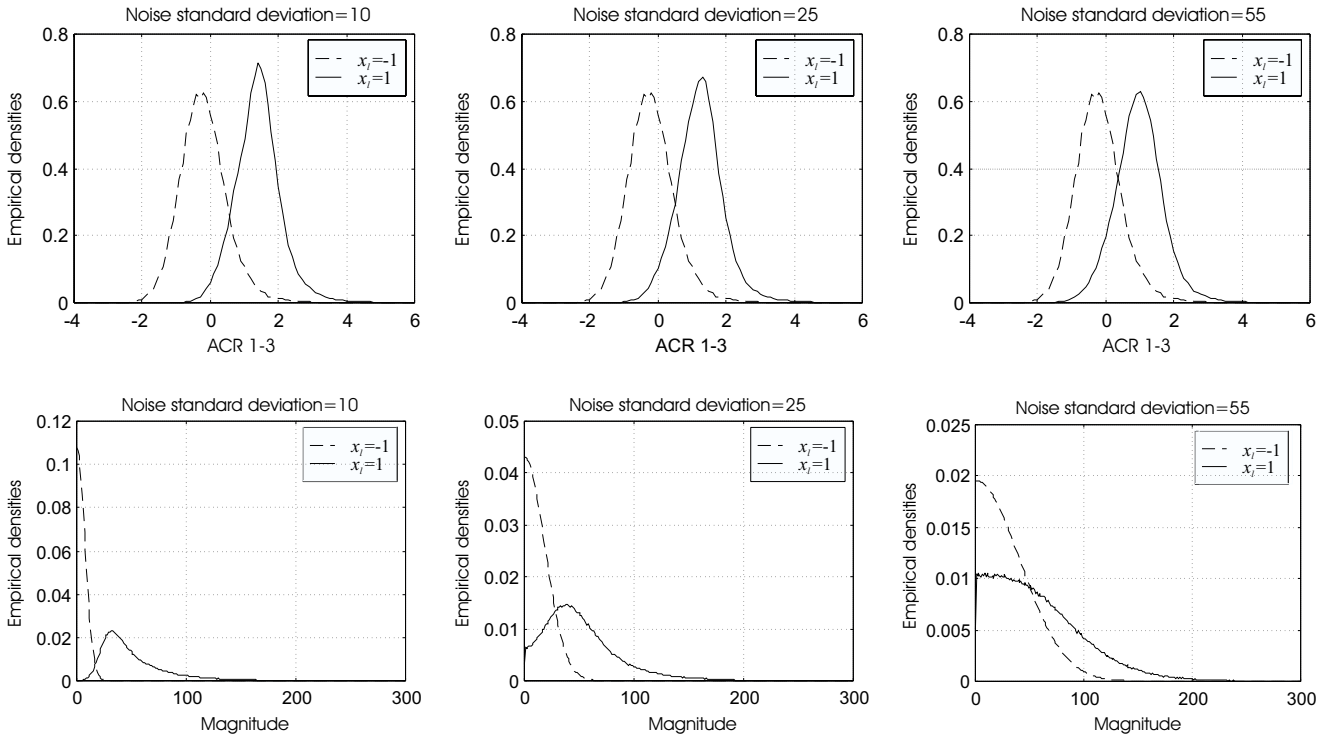


Fig. 3. Comparison between empirical densities of ACR computed from scales $2^1 - 2^3$ (top) and empirical densities of coefficient magnitudes at scale 2^2 (bottom), for reference edges 1 from Fig. 1.

is true for high noise levels. This is further investigated in Section IV. It is not shown here, but one can easily extrapolate, that if a different threshold was chosen for defining the edges of interest in Fig. 1, the densities $f_{M_l|X_l}(\omega_l|1)$ are peaked at correspondingly greater or smaller values ω_l, but important is that their qualitative behaviour does not change.

IV. Performance Evaluation

By performance of a given significance measure we mean its ability to distinguish between noise and useful signal without making use of the prior model. This is equivalent to analyzing the performance under the Bayes labeling [8], [27] with $P(X_l = -1) = P(X_l = 1) = 0.5$. We shall compute the receiver operating characteristics (ROC) [27] for a given standard deviation σ of added noise, and the error probabilities depending on σ. To understand the meaning of ROC, imagine a binary classifier that receives m_l as its input and delivers the label x_l by applying a decision threshold T. It can be shown that the optimal decision threshold T_{opt} (for which the total probability of wrongly classified labels is minimum) is the one for which $f_{M_l|X_l}(T_{opt}|-1) = f_{M_l|X_l}(T_{opt}|1)$. The receiver operating characteristics describes the operation of the classifier for a range of different decision thresholds, below and above T_{opt}. It shows fractional number of the labels $x_l = 1$ that are wrongly classified as $x_l = -1$ (false negatives - FN), versus the fractional number of the labels $x_l = -1$ that

are wrongly classified as $x_l = 1$ (false positives - FP). For $P(X_l = -1) = P(X_l = 1)$, FN and FP are simply determined as

$$FN = \int_{-\infty}^{T} f_{M_l|X_l}(m_l|1)dm_l,$$
$$FP = \int_{T}^{\infty} f_{M_l|X_l}(m_l|-1)dm_l, \quad (6)$$

For the case $T = T_{opt}$, we shall compute the error probability $P_e = 0.5(FN + FP)$.

Recall that we have derived the empirical densities $f_{M_l|X_l}(m_l|1)$ from realistic edges corrupted by noise, and $f_{M_l|X_l}(m_l|-1)$ from pure noise. Therefore, FN, FP and P_e that we consider refer to the ability to detect realistic noisy edges from *pure noise*. We call it expected performance (Section IV-A). The actual performance for different test images, where instead of pure noise one deals with a combination of *noise and image texture* is addressed in Section IV-B.

A. Expected performance

Using the empirical densities derived from the test images in Fig. 1, we compare first the two interscale ratios: APR and ACR. Discussing Fig. 2, we have already remarked that ACR should perform better. The computed ROC in Fig. 4(a) and error probabilities as a function of σ in Fig. 4(b) confirm this observation. Fig. 4(c) illustrates that the expected performance of interscale ratios is improved when they are averaged over a greater number of resolution scales.

Let us now compare the performance of ACR (which was the better one among the two considered interscale-ratios) and the magnitudes of the wavelet coefficients. Fig. 5(a) illustrates that at the finest resolution scale 2^1, ACR offers a better performance than the coefficient magnitude even for small amounts of noise. However, this is not the case at coarser resolution scales. At the scale 2^2, as shown in Fig. 5(b), ACR offers a better performance only for relatively high standard deviations σ of noise.

B. Actual performance

The above presented, expected performance comparison between different significance measures relies on the empirical densities $f_{M_l|X_l}(m_l|x_l)$ and refers to the detection of noisy edges from pure noise. To investigate its validity in practice, we shall apply a decision threshold directly to the computed significance map $\mathbf{m}$, and actually count the number of misclassified labels, with respect to the assumed "ground truth" edges. A reference edge label at the position l will be denoted by x_l^R, and the one that results from thresholding m_l will be denoted by x_l^T. The empirical FP is then equal to $FP_e = \sharp\{l|x_l^R = -1, x_l^T = 1\}/\sharp\{l|x_l^R = -1\}$, and the empirical FN is computed as $FN_e = \sharp\{l|x_l^R = 1, x_l^T = -1\}/\sharp\{l|x_l^R = 1\}$, where $\sharp S$ denotes the cardinality of the set S. It should be noted that in contrast to earlier analyzed performance, FN_e and FP_e, refer to the separation between realistic edges and a combination of noise and image texture, instead of pure noise. Concerning coefficient magnitudes, it does not make a significant difference and the two ways of computing error probabilities give almost the same results. For interscale ratios empirical FN_e and FP_e are greater than those that were derived from the estimated densities, and the difference is image-dependent.[3]

In Fig. 6, we show the empirical error probabilities computed as $(FN_e + FP_e)/2$ for a range of decision thresholds. One can see that main conclusions regarding the comparison between the three significance measures remain valid on these empirical curves. Also, it can be seen that for interscale ratios, a well chosen decision threshold provides nearly optimal performance for different images and for different noise levels, as was expected according to the empirical densities.

V. A joint measure of significance

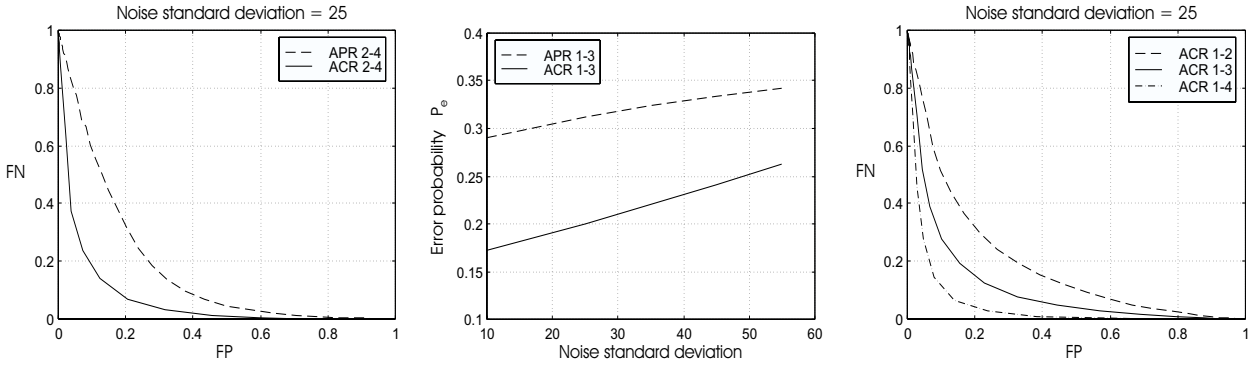
On the basis of the previous analysis, we now propose a joint significance measure m_l, which relies on both the coefficient magnitude ω_l and on the interscale ratio at the corresponding spatial position. It was shown that ACR offers a better performance than APR. Therefore, for wavelet coefficients at resolution scale 2^j we compute $\beta_{1\to j+1,l}$. It was the best choice according to our experiments: averaging ratios over scales coarser than 2^{j+1} suppresses background noise better, but increases the number of falsely selected coefficients in regions adjacent to image edges. To simplify notations, in the remainder we use only the location index l. The proposed significance measure is thus $m_l = (\omega_l, \beta_l)$, where ω_l and β_l are computed from the observed wavelet coefficients, and are realizations of random variables Ω_l and B_l, respectively. We still need to specify conditional densities $f_{M_l|X_l}(m_l|x_l)$ for this joint significance measure m_l. From a practical point of view, it is most convenient to assume that β_l and ω_l are conditionally independent given x_l,

$$f_{M_l|X_l}(m_l|x_l) = f_{\Omega_l|X_l}(\omega_l|x_l) f_{B_l|X_l}(\beta_l|x_l). \quad (7)$$

To determine the validity of this assumption, we find empirically the actual joint conditional densities, $f_{M_l|X_l}(\omega_l, \beta_l|x_l)$ using the simulation method from Section III, adapted for the two-dimensional case. Fig 7 shows the contour plots of the empirical joint densities in comparison with the product of two independent densities from Eq (7). These diagrams suggest that the assumption about the conditional independence is not true, but is acceptable as an approximation.

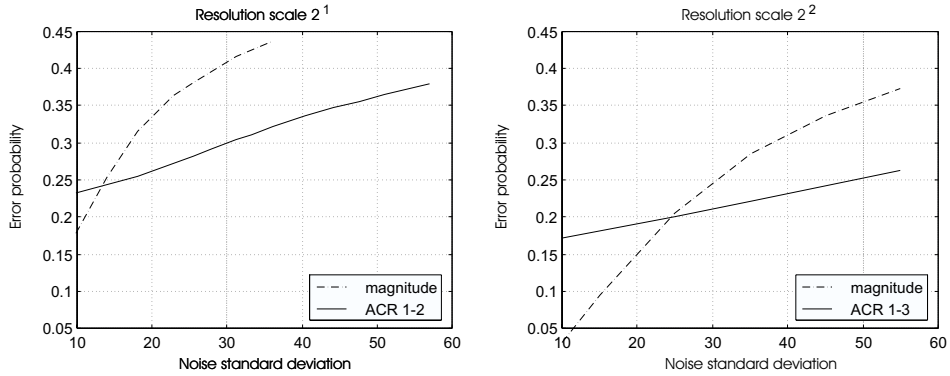
With the proposed conditional probability model (7),

[3] The main reason for this is: the wavelet coefficients that originate from noisy image texture propagate better through scales than coefficients of pure noise; even if they have small magnitudes, they can have large interscale ratios.



(a) (b) (c)

Fig. 4. (a) ROC for APR and for ACR, calculated from the scales $2^2 - 2^4$. (b) Error probabilities for optimal decision thresholds versus standard deviation of noise. (c) ROC for ACR averaged between different scales.



(a) (b)

Fig. 5. A comparative performance of ACR and coefficient magnitudes at the scales (a) 2^1 and (b) 2^2.

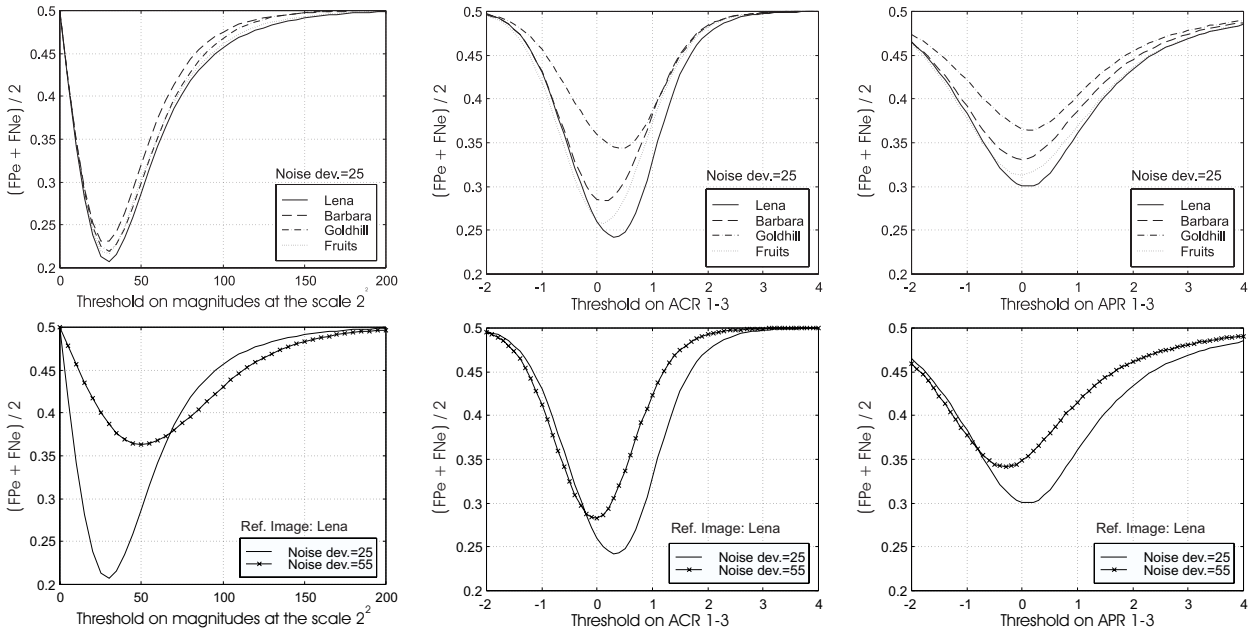


Fig. 6. Empirical error probabilities computed as $(FN_e + FP_e)/2$. Top: for different images and the same noise deviation. Bottom: for the same image and different noise deviations.

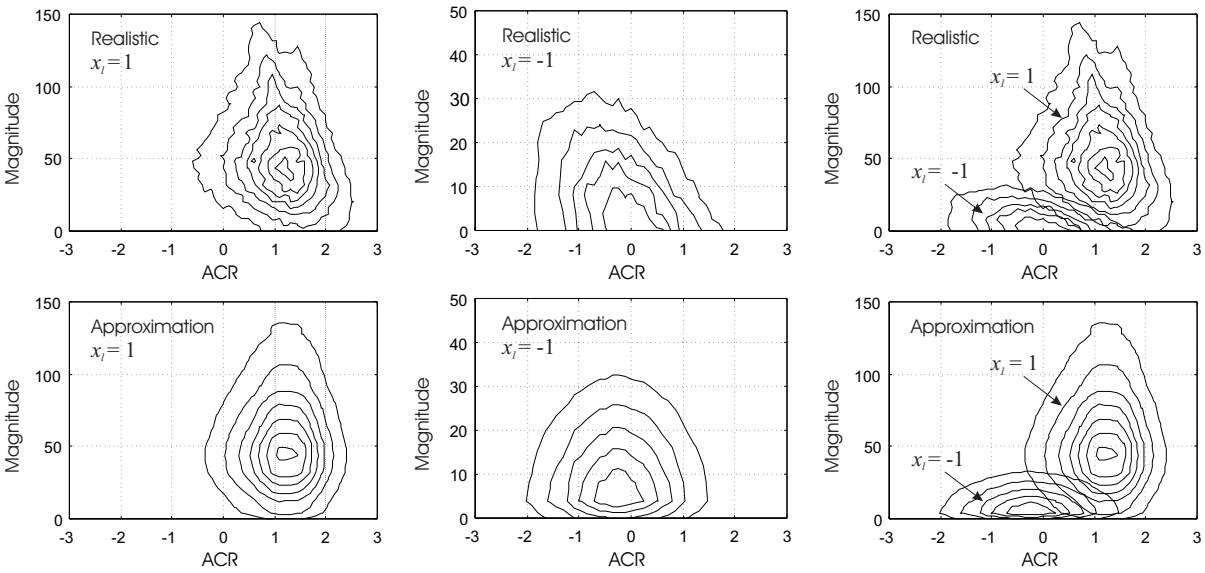


Fig. 7. Joint densities for ACR and coefficient magnitudes at the scale 2^2, for $\sigma = 25.5$. Top row: empirical joint densities. Bottom row: products of empirical one dimensional densities.

the parameter p in Eq (3) becomes

$$p = \Big(\frac{f_{\Omega_l|X_l}(\omega_l|x_l^c)}{f_{\Omega_l|X_l}(\omega_l|x_l)}\Big)^{\lambda_1}\Big(\frac{f_{B_l|X_l}(\beta_l|x_l^c)}{f_{B_l|X_l}(\beta_l|x_l)}\Big)^{\lambda_2} e^{\lambda_3\left(V_{N_l}(\mathbf{x})-V_{N_l}(\mathbf{x}^c)\right)}, \tag{8}$$

where $\lambda_1 = \lambda_2 = 1$ and λ_3 is a positive constant. In practice, it may be useful to allow other values of λ_1 and λ_2 in order to change the relative importance given to ω_l and β_l. For example, in the special case where $\lambda_1 = 0$, this model reduces roughly to the one in [17], except for the fact that there APR was used instead of ACR, and that other, heuristic density model was assumed. For $\lambda_2 = 0$ the joint conditional model basically reduces to the one of [11] and [12]. In Section VII-A we investigate the influence of different values λ_1, λ_2 and λ_3 on denoising results.

In actual denoising procedure, we implement the densities from Eq (8) as follows. Since white Gaussian noise is assumed, $f_{\Omega_l|X_l}(\omega_l|-1) = 2f_N(\omega_l)$, $\omega_l \geq 0$, where $f_N(n)$ is the Gaussian density with zero mean and variance σ_n^2, which is here assumed to be known[4]. Realistic models for $f_{\Omega_l|X_l}(\omega_l|1)$ were already discussed in [12] and we use one such model. In particular, we assume a generalized Laplacian distribution [18] $f_{Y_l}(y_l) \propto \exp(-y_l/s)^p$, for noise-free wavelet coefficients y_l. The parameters s and p are reliably estimated from the input noisy image [25]. The density of the significant noise-free wavelet coefficients $f_{Y_l|X_l}(y_l|1)$ is equal to zero for $|y_l| < T$, and proportional to $\exp(-y_l/s)^p$ for $|y_l| > T$, where T is a threshold that specifies the edges of interest. We adopt the reasoning from [12], according to which $T = \sigma_n$. Since $w_l = y_l + n_l$, where y_l and n_l are statistically independent, the density of w_l is given by convolution [23] $f_{W_l|X_l}(w_l|1) = f_{Y_l|X_l}(y_l|1) * f_N(n)$,

[4] In case σ_n^2 is unknown, a robust estimate of the noise variance is the median absolute deviation of the highest-frequency subband wavelet coefficients, divided by 0.6745 [7].

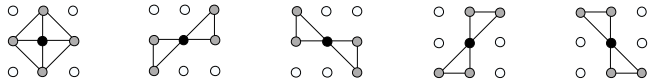


Fig. 8. The proposed sub-neighborhoods.

and it is straightforward that $f_{\Omega_l|X_l}(\omega_l|1) = 2f_{W_l|X_l}(\omega_l|1)$, $\omega_l \geq 0$. For interscale ratios, we have used the empirical densities, because they are robust with respect to noise level. We have verified that using empirical ratios $f_{B_l|X_l}(\beta_l|1)/f_{B_l|X_l}(\beta_l|-1)$ derived from different reference edges and for different noise variance, does not affect denoising performance noticeably. Therefore, one ratio of ACR densities was stored and used for all results presented in this paper.

VI. A New MRF Prior Model

The simplest isotropic MRF model is not well suited to encode prior knowledge about the spatial clustering of the wavelet coefficients. Significant coefficients form relatively narrow clusters with predominantly horizontal, vertical or diagonal directions, depending on the orientation subband. At coarser scales, even the isotropic model with a small neighborhood performs well. However, the problem arises at the finest scales, where details are very thin lines.

The idea behind our model is the following: for each spatial position l, we define a given number of oriented sub-neighborhoods, which contain possible micro-edges centered at the position l. The binary label value $x_l = 1$ (edge) should be assigned a high probability if *any* of the oriented sub-neighborhoods indicates the existence of an edge element in a certain direction. On the contrary, the binary label value $x_l = -1$ (non-edge) should be assigned a high probability only if *no one* of the oriented neighborhoods indicates the existence of a such edge element.

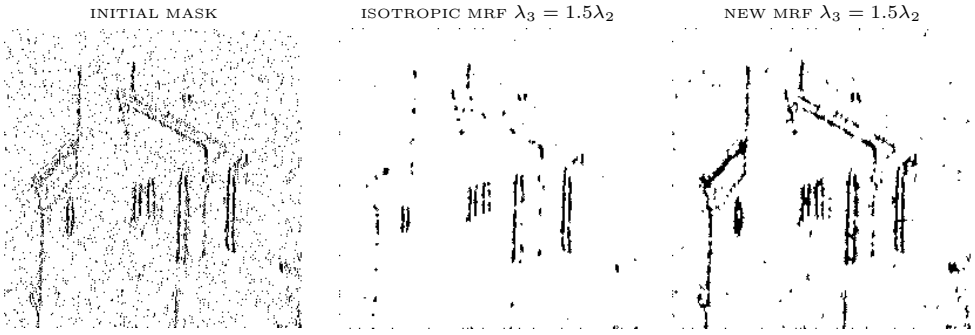


Fig. 9. Masks at the scale 2^1, for SNR=9dB. Left: initial. Middle: using the isotropic MRF model. Right: using the new, anisotropic MRF model. Equal relative influence was given to both prior models.

One, but not the only possible, realization of the above explained idea follows. We choose the sub-neighborhoods $N_{l,i}$, $1 \leq i \leq 5$ shown in Fig. 8; each $N_{l,i}$ contains four neighbors of the central pixel l. We use only pair-wise cliques $\{k,l\}$, with the potential function $V_2(k,l) = -\gamma x_k x_l$, where γ is a positive constant. The potential $V_{N_{l,i}}(\mathbf{x})$ of the sub-neighborhood $N_{l,i}$, is equal to the sum of all pair-wise potentials

$$V_{N_{l,i}}(\mathbf{x}) = \sum_{k \in N_{l,i}} V_2(k,l) = -\gamma \; x_l \sum_{k \in N_{l,i}} x_k. \qquad (9)$$

The potential associated with the *complete* neighborhood N_l of the position l will be denoted by $V_{N_l}(\mathbf{x})$. To determine this potential we follow the idea that was introduced at the beginning of this Section. An edge-label $x_l = 1$ should be assigned a high probability if *any* of $N_{l,i}$ indicates the presence of an edge element. Therefore, $V_{N_l}(\mathbf{x}|x_l = 1) = \min_i\{V_{N_{l,i}}(\mathbf{x})\} = -\gamma \; \max_i\{\sum_{k \in N_{l,i}} x_k\}$. A non-edge label $x_l = -1$ should be assigned a high probability if *none* of $N_{l,i}$ indicates the presence of an edge element. This is accomplished if we choose $V_{N_l}(\mathbf{x}|x_l = -1) = \max_i\{V_{N_{l,i}}(\mathbf{x})\} = \gamma \; \max_i\{\sum_{k \in N_{l,i}} x_k\}$, which is exactly the opposite with respect to $V_{N_l}(\mathbf{x}|x_l = 1)$. We can present both of these potentials with the same expression, if we keep the label x_l as a variable:

$$V_{N_l}(\mathbf{x}) = -\gamma \; x_l \; \max_i\Big\{ \sum_{k \in N_{l,i}} x_k \Big\}. \qquad (10)$$

This completes the specification of the new prior model. The expression for the joint probability is of the form (1), but with the new neighborhood potential $V_{N_l}(\mathbf{x})$, given in (10).

Fig. 9 illustrates the advantage of the new, anisotropic MRF model. The three masks correspond to vertical details at the resolution scale 2^1, for the house image with input SNR=9dB. The initial mask was obtained by thresholding ACR. Starting from this initial mask, we ran 10 iterations of the Metropolis sampler, with the isotropic prior model, and repeat the procedure using the new prior model. In both cases equal relative importance was given to the prior model $\lambda_3 = 1.5\lambda_2$ and $\lambda_1 = 0$. One can see that the isotropic model removes background noise, but it also removes useful edges. The new model, in contrast to this, has the ability to preserve the useful edges well.

VII. Results and Discussion

The practical implementation of the algorithm is briefly described in the Appendix. Its performance will be illustrated on two 256x256 images with artificial noise: the peppers and the house images from Fig. 11. The first one is rich with slow intensity variations and natural texture, while the second one is mainly characterized by sharp edges and flat background. As a quantitative performance measure, we use the signal to noise ratio defined as $SNR = 10\log_{10}(P_{signal}/P_{noise})$, where P_{signal} is the variance of the input image, and P_{noise} the variance of noise. This performance measure was chosen for the sake of clear comparison with the most related methods from literature, in particular the one of [17]. In Table 1, we also indicate a simple calculation of the peak-signal-to-noise-ratio $PSNR = 10\log_{10}(255^2/\textit{mean squared error})$ values from given SNR, for both images.

A. Choice of parameters

The proposed method relies on the magnitudes of wavelet coefficients, their interscale dependencies and on prior knowledge about the spatial clustering. The relative importance given to these three sources is expressed through the parameters λ_1, λ_2 and λ_3, respectively, in Eq (8). To quantify the relative influence of the two significance measures with respect to each other, and with respect to the prior model at the same time, we set $\lambda_3 = \lambda$, $\lambda_1 = k_1\lambda$, and $\lambda_2 = k_2\lambda$. The choice of λ is then equivalent to choosing a different "temperature" [14] of the Metropolis sampler, which is not critical in this method. Therefore, we fix it to a constant value (in particular, $\lambda = 10$) and concentrate on the optimum choice of the relative influences $k_1 = \lambda_1/\lambda_3$ and $k_2 = \lambda_2/\lambda_3$. We computed the SNR for different images, using a step 0.25 for k_1 and k_2 in the range 0 to 4. Fig. 10 demonstrates that the performance of the method is stable for a wide range of selected parameters. It also shows that better results are always achieved by combining both significance measures instead of using

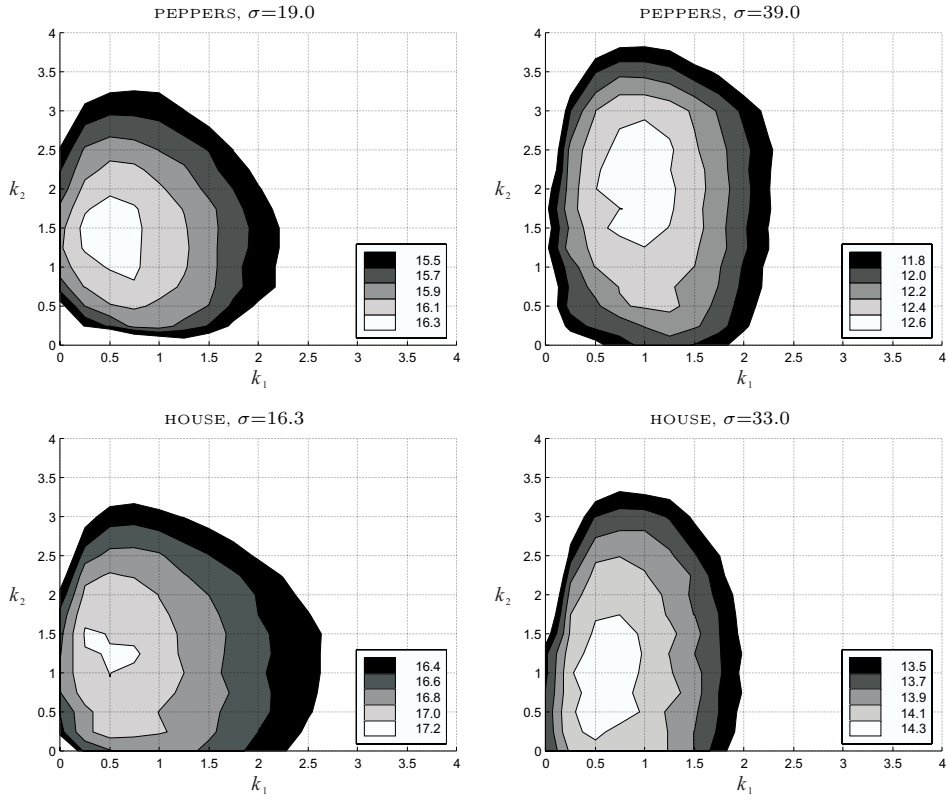


Fig. 10. Resulting SNR corresponding to different values of the parameters $k_1 = \lambda_1/\lambda_3$ and $k_2 = \lambda_2/\lambda_3$, for different test images and different standard deviations σ of added noise.

either of them alone, and that the benefit from a joint conditional model is image-dependent. If we examine the centers of the regions with maximum SNR, then it indeed appears that for higher noise variance the optimal proportion λ_2/λ_1 is bigger, i.e., that the use of interscale statistics is more important. The parameters $\lambda_1/\lambda_3 = 1.5$ and $\lambda_2/\lambda_3 = 0.75$ are near to optimal for all analyzed cases, so we use these parameters for all subsequent results. The visual quality of results in Fig. 11 confirms the advantage of using the joint conditional model over special cases, where $\lambda_1 = 0$ or $\lambda_2 = 0$.

B. Implementation and execution time

We have implemented the proposed method with a 3-subband decomposition as well, using the quadratic spline wavelet from [20] like in [17]. For all results, we used the same parameters $\lambda_1 : \lambda_2 : \lambda_3 = 1.5 : 0.75 : 1$. With respect to the 2-subband case, certain image-dependent improvement is achieved at the expense of computation time. The improvements in terms of SNR, are summarized in Table 1, and the differences in visual quality can be judged from Fig. 12. In all cases, we have used three resolution scales. We have measured the total execution time for 256x256 images, on a computer with Pentium III processor. For the implementations with two and with three orientation subbands, the corresponding execution times are 30s and 45s, respectively.

C. Quantitative results and discussion

Quantitative results for the two test images are summarized in Table 1. The first method, that is used as a basis for comparison, is [17]. It uses 3 orientation subbands and is similar to a special case of our method, with $\lambda_1 = 0$. However, note that [17] uses APR with a heuristic conditional model, the parameters of which need to be estimated for each detail image at each scale and that it applies the isotropic MRF prior model. For the peppers image, our method achieves a significant improvement, which is bigger for higher input SNR. For example, for input SNR=9dB the new method is 1.6dB better, and for input SNR=0dB it is 0.6dB better). For the house image, the results of the new method are almost the same as those in [17]. One can also see from the contour plots in Fig. 10 that for the house image setting $k_1 = \lambda_1/\lambda_3 = 0$ does not incur such a big penalty on the resulting SNR as it does for the peppers image. The computation time is approximately the same for both methods, since the Metropolis sampling procedure, takes the majority of computing time. The differences in prior and conditional models are insignificant in this respect: both prior models practically involve pairwise pixel interactions only and converge equally fast. The performance of the new prior model and the isotropic one from [17] can be visually compared in Fig. 13.

Table 1 also lists the results from [20]. The method of [20] is considered one the best for detecting multiscale edges from a noisy image. However, it makes all textures

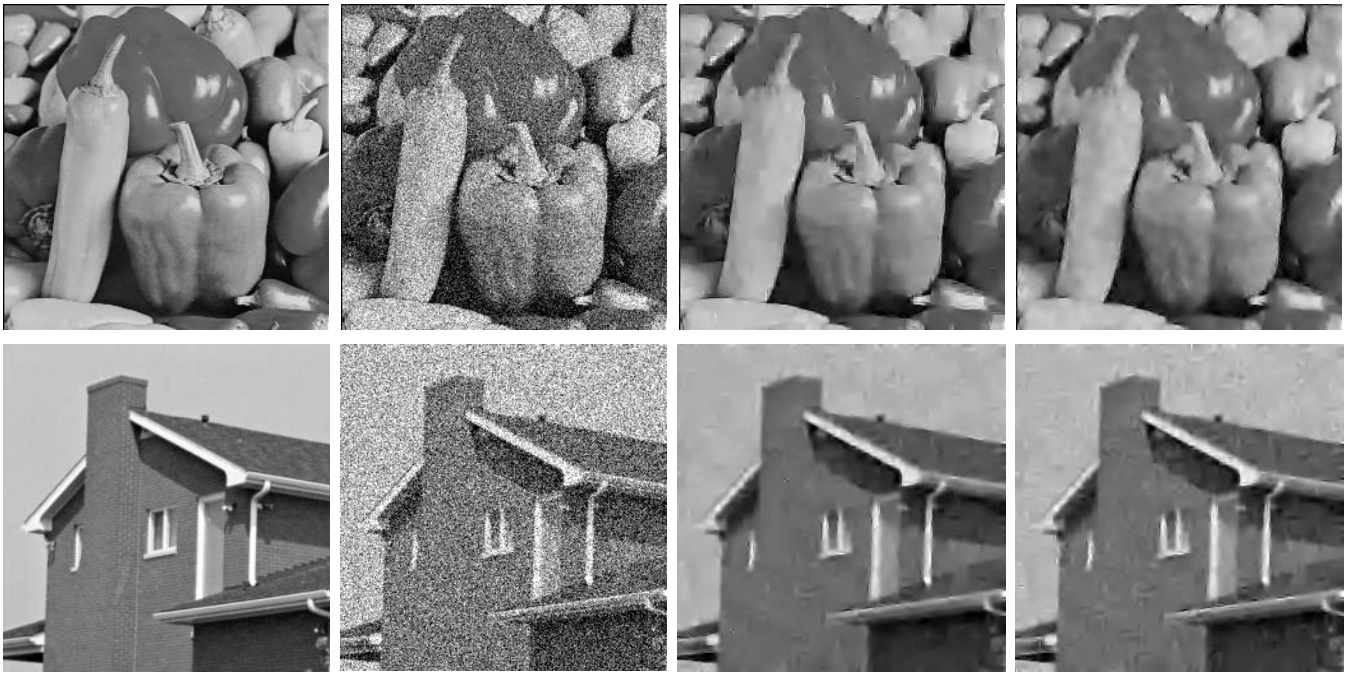


Fig. 11. Influence of the proposed joint conditional model on the visual quality of results. Top: noise-free image, noisy image SNR=3dB, the result of the proposed method using the joint conditional model, and the corresponding result for $\lambda_1 = 0$. Bottom: noise-free image, noisy image SNR=0dB, the result of the proposed method using the joint conditional model, and the corresponding result for $\lambda_2 = 0$. Two orientation subbands are used in all cases.

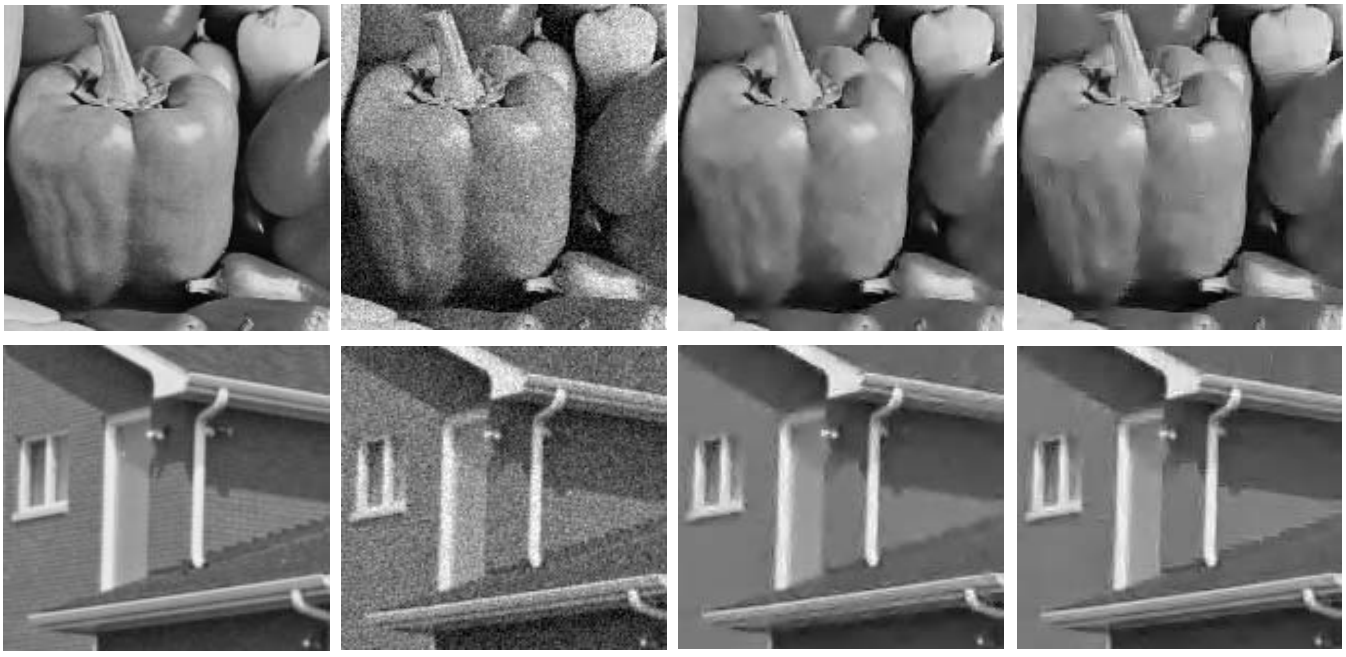


Fig. 12. From left to right: enlarged parts of noise-free images, noisy images $SNR = 9dB$, and the results of the proposed method using 2 and using 3 orientation subbands.

disappear, as was noted by the authors themselves. For this reason our method achieves a bigger improvement with respect to [20] for smaller noise levels, where such textures can still be reconstructed.

We have also included the results of Matlab's spatially adaptive Wiener filter. Its window size was optimized for each input image to produce the maximum SNR. For both test images these sizes were 3x3 for input SNR 9dB, 5x5 for input SNR 6 and 3dB and 7x7 for input SNR 0dB.

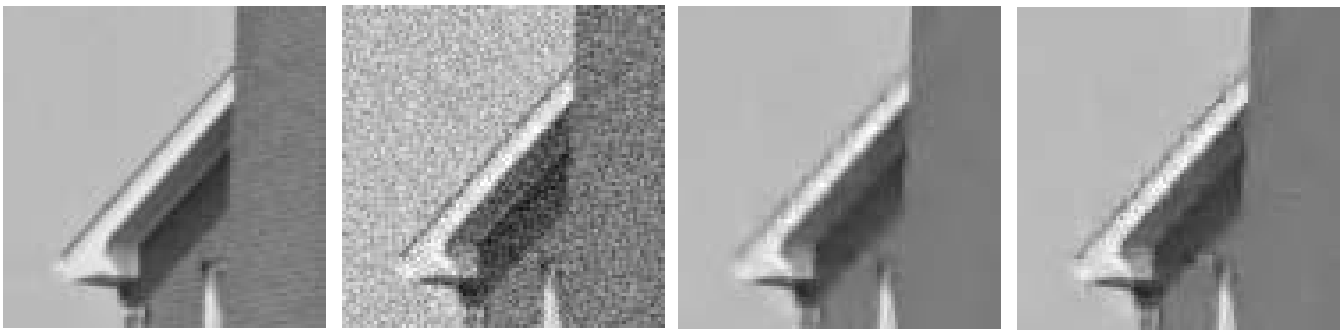


Fig. 13. Comparing the performance of the isotropic and the anisotropic MRF prior model on a part of the house image. From left to right: noise-free image, noisy image $SNR = 9dB$ and the results of the 3-subband version of the proposed method, using isotropic MRF model and using the new MRF model.

TABLE I

COMPARISON OF QUANTITATIVE RESULTS, EXPRESSED IN $SNR[dB]$, FOR DIFFERENT METHODS.*

	PEPPERS				HOUSE			
Input	9	6	3	0	9	6	3	0
New, 2 orientation subbands	16.3	14.6	13.0	11.3	17.2	15.6	14.4	13.2
New, 3 orientation subbands	16.6	15.0	13.4	11.6	17.9	16.4	14.9	13.4
Malfait and Roose [17]	15.0	13.7	12.4	11.0	18.0	16.4	14.9	13.3
Mallat and Hwang [20]	14.6	13.7	13.5	11.0	16.4	15.6	14.2	12.2
Matlab's spatially adapt. Wiener	15.4	13.5	11.7	9.7	15.4	13.7	11.8	10.0

*To compute the corresponding $PSNR[dB]$, 13.6 should be added to all values for the peppers image, and 14.9 should be added to all values for the house image.

VIII. Conclusion

In this paper, several issues were addressed to improve Bayesian image denoising using prior models for spatial clustering. A new MRF prior model was introduced to preserve image details better. A joint significance measure, which combines coefficients magnitudes and their evolution through scales was introduced. For the resulting, joint conditional model a simple practical realization was proposed and motivated via simulations. The advantage of the joint conditional model in terms of noise suppression performance was demonstrated on different images and for different amounts of noise. Some aspects that were analyzed in this paper may be useful for other denoising schemes as well: the realistic conditional densities of interscale ratios obtained via simulations, and objective criteria for evaluating noise suppression performance of different significance measures.

Appendix

The complete procedure in the proposed method is:

- Compute the discrete wavelet transform without down-sampling
- For each orientation and for resolution scales 2^j, $1 \leq j \leq 3$
  - For all coefficients w_l, $1 \leq l \leq n$ compute significance measures $m_l = (\omega_l, \beta_l)$
  - Run the Metropolis algorithm:
    - Initialize binary mask $\mathbf{x}$ and set $N_{local} \leftarrow 0$ and $S_l \leftarrow 0$, for $1 \leq l \leq n$
    - Repeat N_{global} times
      Repeat until all the positions l have been visited
      - Choose l at random; set $x_l^c \leftarrow -x_l$ and $N_{local} \leftarrow N_{local} + 1$
      - Compute p from Eq (8) and generate a random number u from U[0,1]
      - if $p > u$ set $x_l \leftarrow x_l^c$
      - If $x_l = 1$ set $S_l \leftarrow S_l + 1$
  - For $1 \leq l \leq n$: Estimate the wavelet coefficients: $\hat{y}_l = w_l S_l / N_{local}$
- Reconstruct the denoised image by applying the inverse wavelet transform.

Acknowledgement

The authors wish to thank to all three reviewers for a number of constructive suggestions to improve the original text and contents of the manuscript. The authors also thank to Dimitri Van De Ville for fruitful discussions concerning this work.

References

[1] S. G. Chang, B. Yu, and M. Vetterli, "Spatially Adaptive Wavelet Thresholding with Context Modeling for Image Denoising ," *IEEE Trans. on Image Proc.*, vol. 9, pp. 1522–1531, Sep 2000.

[2] H.A. Chipman, E.D. Kolaczyk, and R.E. McCulloch, "Adaptive Bayesian Wavelet Shrinkage," *J. of the Amer. Statist. Assoc.*, vol. 92, pp. 1413–1421, 1997.

[3] A. Cohen and J. Kovačević, "Wavelets: The Mathematical Background," *Proc. of the IEEE*, vol. 84, pp. 514–522, 1996.

[4] R.R. Coifman and D.L. Donoho, "Translation-invariant denoising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, editors, pp. 125–150, New York: Springer Verlag, 1995.

[5] M. S. Crouse, R. D. Nowak, and R. G. Baranuik, "Wavelet-Based Statistical Signal Processing Using Hidden Markov Models," *IEEE Trans. on Signal Proc.*, vol. 46, pp. 886–902, Apr 1998.

[6] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia: SIAM, 1992.

[7] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, 1995.

[8] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* John Wiley & Sons, 1973.

[9] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. on Pattern Anal. and Machine Intel.*, vol. 6 pp. 721–741, Nov 1984.

[10] T.-C. Hsung, D. P.-K Lun, and W.-C. Siu. "Denoising by singularity detection," *IEEE Trans. on Signal Proc.*, vol. 47, pp. 3139–3144, Nov 1999.

[11] M. Jansen and A. Bultheel, "Geometrical Priors for Noise-free Wavelet Coefficient Configurations in Image De-noising," in *Bayesian inference in wavelet based models*, P. Müller and B. Vidaković, editors, Springer Verlag 1999, pp. 223–242.

[12] M. Jansen and A. Bultheel, "Empirical Bayes approach to improve wavelet thresholding for image noise reduction," *J. of the Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 629-639, 2001.

[13] M. Kivanç Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-Complexity Image Denoising Based on Statistical Modeling of Wavelet Coefficients," *IEEE Signal Proc. Letters*, vol. 6, pp. 300–303, Dec 1999.

[14] S. Z. Li, *Markov Random Field Modeling in Computer Vision,* Springer-Verlag, 1995.

[15] J. Liu and P. Moulin, "Analysis of interscale and intrascale dependencies between image wavelet coefficients", *in Proc. Int. Conf. on Image Proc., (ICIP)*, Vancouver, Canada, Sep. 2000.

[16] D.J.C. MacKay, "Introduction to Monte Carlo Methods," in *Learning in Graphical Models*, M. I. Jordan, editor, MIT Press, 1999.

[17] M. Malfait and D. Roose, "Wavelet-Based Image Denoising Using a Markov Random Field a Priori Model," *IEEE, Trans. on Image Proc.*, vol. 6, pp. 549–565, April 1997.

[18] S. Mallat, "A theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Anal. and Machine Intel.*, vol. 11, no. 7, pp. 674–693, July 1989.

[19] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *IEEE Trans. on Pattern Anal. and Machine Intel.*, vol. 14, pp. 710–732, July 1992.

[20] S. Mallat and W. L. Hwang, "Singularity Detection and Processing with Wavelets," *IEEE Trans on Inform. Theory*, vol. 38, pp. 617–643, Mar 1992.

[21] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.

[22] P. Moulin and J. Liu, "Analysis of Multiresolution Image Denoising Schemes Using Generalized Gaussian and Complexity Priors," *IEEE Trans. Inform. Theory*, vol. 45, pp. 909–919, Apr 1999.

[23] A. Papoulis, *Probability, random variables, and stochastic processes* , McGraw-Hill, 1984.

[24] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian Tree Structured Image Modeling Using Wavelet-Domain Hidden Markov Models," *IEEE Trans. on Image Proc.*, vol. 10, no. 7, pp. 1056–1068, July 2001.

[25] E. P. Simoncelli and E. H. Adelson, "Noise Removal via Bayesian Wavelet Coring," *in Proc. IEEE International Conf. Image Processing (ICIP)*, Lausanne, Switzerland , pp. 379–382, 1996.

[26] E. P. Simoncelli, "Modeling the joint statistics of image in the wavelet domain," *in Proc. SPIE Conf. 3813 on Wavelet Applications in Signal and Image Processing VII*, Denver, CO, July 1999.

[27] H.L. Van Trees, *Detection, Estimation and Modulation Theory*, John Wiley, New York, 1968.

[28] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding* , Prentice-Hall, 1995.

[29] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Caro Methods* , Springer-Verlag, 1995.

[30] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu, "Wavelet Transform Domain Filters: A Spatially Selective Noise Filtration Technique," *IEEE Trans. Image Proc.*, vol. 3, pp. 747–758, Nov 1994.