Image Denoising Algorithms: From Wavelet Shrinkage to Non-local Collaborative Filtering

Aleksandra Pižurica

***Abstract*—This paper presents an overview of image denoising algorithms ranging from wavelet shrinkage to patch-based non-local processing. The focus is on the suppression of additive Gaussian noise (white and coloured). A great attention is devoted to explaining the main underlying ideas and concepts of representative approaches with illustrative examples, accessible also to non-experts in the field. A Bayesian perspective of wavelet shrinkage is given, with different instances of spatial context modelling (including local spatial activity indicators, Markov Random Fields, Hidden Markov Tree models and Gaussian Scale Mixture models). Extensions to other transform domains (curvelets and other generalizations of wavelets) are addressed too, showing the benefits in terms of image quality. Patch-based image denoising is illustrated with principles of non-local means filtering and collaborative filtering, explaining also the connections with dictionary learning. Some general notes on the performance comparison are given, by summarizing the benefits and limitations of various approaches against each other, and pointing to some of the current trends in the field.**

***Index Terms*—Noise reduction, Wavelet shrinkage, Spatial context, Gaussian Scale Mixture, Markov Random Field, Hidden Markov Model, Patch-based processing, Non-local means, Collaborative filtering, Dictionary learning.**

I. Introduction

Image noise manifested as random fluctuations of values of digital picture elements (pixels) arises inevitably during image formation; the origins of such image noise are both in the analogue electronic circuitry in the imaging device and in the physical phenomena characteristic for a given imaging modality (e.g., fluctuations in the emitted photon distributions in X-ray imaging or random scattering of electromagnetic waves in radar imaging).

Image denoising is desired in many applications, not only to improve visual appearance or *user experience* (which is of particular interest for digital camera images) but also to facilitate subsequent automatic processing (segmentation into distinct regions, detection of edges or certain features of interest). Fig. 1 illustrates an example of applying automatic edge detection to a satellite image before and after noise reduction. The image was acquired by the Synthetic Aperture Radar (SAR) [1] and is affected mainly by speckle noise [2]. Note how many spurious short edges are present when edge detection is applied to the noisy image. The actual contours are hidden in the lots of "rubbish" resulting from noise, while some of the actual coastline contours are not detected. After noise reduction, the edge detection reveals clearly even the contours of the smallest and hardly visible islands. This is why noise reduction is so important in practice.

A. Pižurica is with the Department Telecommunications and Information Processing, IPI-TELIN-iMinds, Ghent University, Belgium
E-mail: aleksandra.pizurica@ugent.be

As one of the fundamental problems in digital image processing, image denoising has been extensively studied over several past decades. Development of *locally adaptive spatial filters* in the early eighties was an important breakthrough [3], [4]: the filtering was applied to each pixel independently (hence allowing simple and fast implementation, even parallel processing) and the parameters were automatically tuned at each spatial position according to the local image content. For example, for the case of additive white Gaussian noise, the approach of [3] operates as a locally adaptive Wiener filter [5], where the signal variance is estimated from a local window around each pixel. The local spatial adaptation can also be interpreted as means to achieve *non-linearity* in the filter design; linear filters, like the classical Wiener filter would inevitably oversmooth the image edges. Important classes of nonlinear filters for image denoising, include *order-statistic* [6], *stack* [7], [8], *weighted median* [9], [10], *morphological* [11] and *rational* [12] filters; for a compact overview see [13].

Another class of successful and widely used nonlinear filters are methods based on minimizing *total variation* (TV) [14]–[16]. Inspired by anisotropic diffusion [17], the TV anisotropic diffusion model, known also as the ROF (Rudin-Osher-Fatemi) model, was first introduced in a seminal work [18]. Faster algorithms for TV filtering include [19], [20]. Some of the most efficient current image restoration methods have been constructed using TV model with optimization methods such as *alternating direction method of moments* (ADMM) [21], [22], variations or equivalent formulations thereof, including *split-Bregman* methods [23] and other *forward-backward splitting* schemes [24]. For some of the recent related developments see [25], [26]. *Bilateral* filtering [27] enjoys great popularity as a conceptually simple nonlinear filtering scheme, which can be implemented in a non-iterative manner. The pixel value is simply replaced by a weighted average of the nearby pixels; the weights are influenced both by the spatial proximity to the central pixel and by the similarity in values. The idea can be traced back to nonlinear Gaussian filters [28], and the concept has been applied in various other contexts next to image denoising [29].

An information-theoretic approach to denoising advocated as *universal* denoiser [30] denoises data sequences generated by a discrete source and received over a discrete, memoryless channel assuming no knowledge of the source statistics (hence the name universal). Excellent performance was demonstrated on binary images (halftone images and scanned text). An extension for greyscale images is in [31].

This paper concentrates on the removal of *additive Gaussian noise*, which is most commonly addressed noise type in the image denoising literature. Other important noise models include *impulsive noise* arising often from analogue to digital conversion and due to various transmission errors, [32], [33], *speckle* noise [2] affecting all coherent imaging systems (including ultrasound and radar imaging) and *Poisson noise* that characterizes photon-limited imaging.

From the early nineties until recently, image denoising, and especially Gaussian noise removal, was largely dominated by *wavelet shrinkage* methods [34]–[37]. Currently, *patch-based* and *non-local* techniques [38]–[40], often in combination with wavelet-like representations or with *learned dictionaries* of image atoms [41], provide the state-of-the-art results. The underlying idea behind many of the recent approaches is to collect similar image patches throughout the image and to jointly process them, which is sometimes denoted as *collaborative filtering* [42]. Here we focus on highlighting the main concepts and ideas behind different classes of well known wavelet-shrinkage and patch-based methods, without going into much technical details. This article also aims at providing sufficient background to make the content accessible to non-experts in the field. Recent methods are reported, e,g., in [37], [43]–[45] and comprehensive reviews with performance analyses in [46], [47].

This paper is organized as follows. Section II explains the assumed noise model and places it in a wider context of noise modelling in digital imaging sensors. Transform domain methods (with emphasis on wavelet shrinkage) are presented in Section III, and patch-based non-local methods in Section IV. Some insights into how denoising approaches are evaluated and how different denoising methods are typically compared are in Section V, together with general guidelines on choosing the right denoising method for a given task. The paper ends with some future prospects in Section VI.

II. Noise model

We explain the assumed noise model and how it can be applied more widely in practice, with certain preprocessing steps that are detailed elsewhere. The denoising schemes presented in the following Sections apply to the following additive noise model:

$$\mathbf{d} = \mathbf{f} + \boldsymbol{\vartheta} \tag{1}$$

where $\mathbf{d} = \{d_1, ..., d_n\}$ is the input image, $\boldsymbol{\vartheta} = \{\vartheta_1, ..., \vartheta_n\}$ is noise (a vector of random variables), and $\mathbf{f} = \{f_1, ..., f_n\}$ is unknown degradation-free image; the indices correspond to pixel positions, visited in some raster-scanning order. For zero-mean noise ($E(\boldsymbol{\vartheta}) = \mathbf{0}$), the covariance matrix is

$$Q = E[(\boldsymbol{\vartheta} - E(\boldsymbol{\vartheta}))(\boldsymbol{\vartheta} - E(\boldsymbol{\vartheta}))^T] = E(\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T) \tag{2}$$

On its diagonal are the variances $\sigma_l^2 = E(\vartheta_l^2)$. If the covariance matrix is diagonal, i.e., if $E(\vartheta_l, \vartheta_k) = 0$ for $l \neq k$, the noise is uncorrelated and is called *white*. If all ϑ_l follow the same distribution, they are said to be *identically distributed*. This implies $\sigma_l^2 = \sigma^2$, for all $l = 1, ..., n$. For Gaussian noise, with the probability density function (pdf)

$$p_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(Q)}} e^{-\frac{1}{2}\boldsymbol{\vartheta}^T Q^{-1} \boldsymbol{\vartheta}} \tag{3}$$

it holds that if noise variables are uncorrelated, they are also statistically *independent* $p_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}) = \prod_l p_{\vartheta_l}(\vartheta_l)$. The reverse implication (independent variables are uncorrelated) holds for all densities. A common assumption is that the noise variables are independent, identically distributed (i.i.d.), leading to the commonly employed additive white Gaussian (AWGN) model.

While the conventional AWGN model is typically assumed in the image processing literature, the actual noise in digital imaging sensors is rather *signal-dependent* [50]–[55] and better described by more generic models as [51]:

$$d_i = f_i + \omega(f_i)\phi_i \tag{4}$$

where ϕ_i is zero-mean independent random noise with standard deviation equal to 1, and $\omega(f_i)$ is the standard deviation depending on the signal value f_i. The noise term $\omega(f_i)\phi_i$ is often assumed to be composed of two independent parts [51]: a *Poissonian* signal-dependent component and a *Gaussian* signal independent component. Suppression of this mixed Poisson-Gaussian noise directly by considering the statistics of the two components has been addressed, e.g., in [56] and specialized methods for the suppression of Poisson noise include [57]–[59]. In many cases of practical interest, it is possible to address the above described signal-dependent mixed noise model by adapting the methods aimed for additive white Gaussian noise. The approximation of the Poisson distribution by the normal one

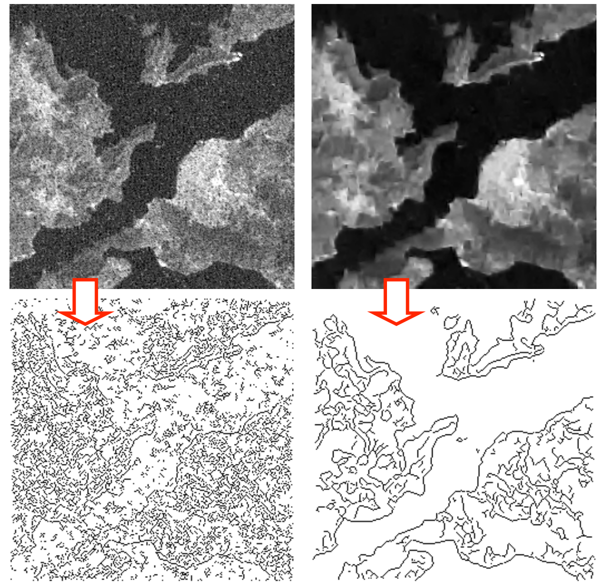

Fig. 1. An example of speckle noise reduction in a SAR image and its favourable effect on the subsequent edge detection. (In this example, Canny's edge detection [48] and denoising method of [49] were used).

$\mathcal{P}(\lambda) = \mathcal{N}(\lambda, \lambda)$, which holds for relatively large λ leads to the common approximation of the generic signal-dependent term in (4) by *heteroscedastic Gaussian* noise [51], [60], [61]: $\omega(f_i)\phi_i = \vartheta_i^h(f_i)$, where $\vartheta_i^h(f_i) \sim \mathcal{N}(0, af_i + b)$, with $a, b > 0$. In general, the mixed noise in (4) can be efficiently treated by applying the methods developed for AWGN after applying the so-called *variance stabilization* [62], [63]. The key idea is to "standardize" the noise, which is then removed by the methods developed for AWGN, and subsequently the inverse of the stabilizing transform is performed. While the generalized Anscombe transform is commonly used as the stabilizing transform, improved solutions are offered in [62], [63]. These and related works are of great value in practice for they enable the widely studied denoising methods for Gaussian noise (as those that will be reviewed next) to be applicable to many real applications with the noise arising from actual digital imaging sensors.

Similarly, certain adaptations are needed in real imaging applications to deal with *clipping*, i.e., censoring [51] of the noisy images. The dynamic range of acquisition, transmission and storage systems is always limited and therefore the applications of standard denoising methods that do not take this aspect specifically into account requires in principle appropriate declipping transformations [60].

Another important aspect in practice is *spatial variability* of the noise characteristics. Electronic sensors in combination with optical components in digital cameras tend to produce spatially non-uniform noise. This is especially interesting for the current *high dynamic range* (HDR) applications [64], [65] that produce noise of different levels in different parts of images. Therefore, denoising not only needs to include noise estimation in local neighborhoods but also to include a kind of soft transitioning of the denoising strength spatially. These cases are outside the scope of the present paper. The reader is referred to studies that treat specifically realistic capture models for imaging devices, see, e.g., [50]–[52] and a recent special issue [55] and articles therein that address different aspects of camera noise modelling [53], [54], [66]. The effect of mismatch between the true and the estimated noise parameters is addressed, e.g., in [67]. Adapting the denoising algorithms to realistic capture models is crucial for alleviating the limitations of current image denoising algorithms in practice.

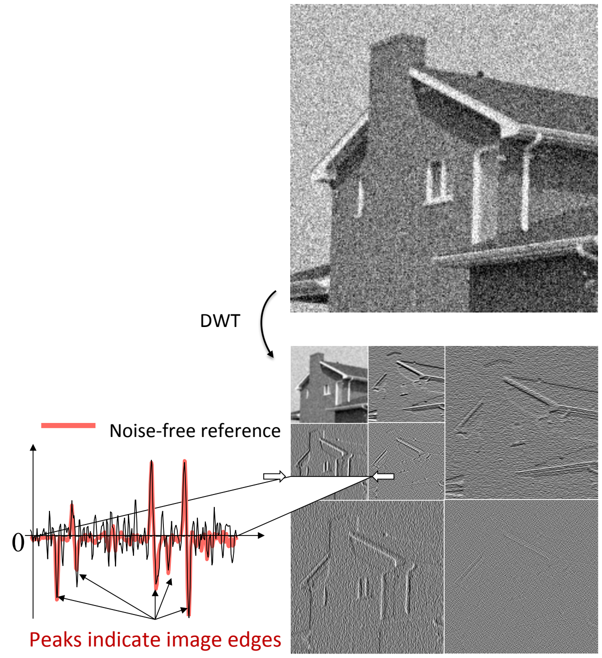

Fig. 2. A wavelet decomposition of a noisy image (DWT – discrete wavelet transform). The coefficients along one line in a vertical subband are shown in comparison with the noise-free reference.

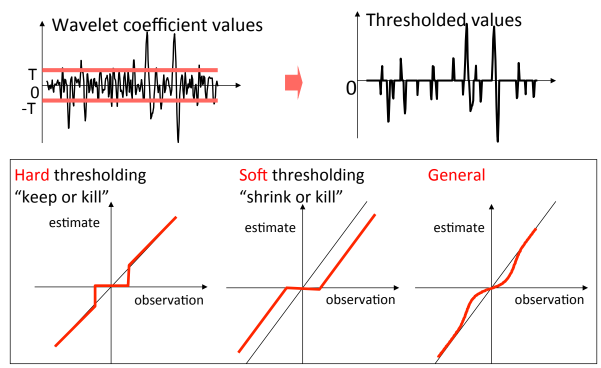

Fig. 3. The effect of wavelet thresholding (top) and representatives of thresholding non-linearities (bottom).

III. Transform Domain Image Denoising

Most of this Section will be devoted to wavelet domain denoising, having on mind that the same principles apply to related transforms that are generalizations of wavelets, such as curvelets [68] and shearlets [69]–[71]. An alternative transform-domain denoising strategy, based on shape-adaptive discrete cosine transform (DCT) is in [72].

The wavelet transform [73]–[76] naturally facilitates construction of *spatially adaptive* image denoising algorithms. The essential information content from a signal or an image is compressed into a relatively few large coefficients, which coincide with the areas of major spatial activity (edges, corners, peaks, etc.). Additive white noise gets spread over all the coefficients and at noise levels that are of practical importance the most important coefficients typically stand out of noise (see an example in Fig. 2). This has motivated noise reduction via *wavelet thresholding* [34], [35], [77], where all the coefficients with magnitudes below a certain threshold are set to zero (see Fig. 3, top). The remaining coefficients are either kept unchanged or shrunk according to some rule. *Hard thresholding* keeps the surviving coefficients unchanged and *soft thresholding* reduces their magnitude by the value of the threshold. A wavelet shrinkage non-linearity can be derived from Bayesian estimation too [78].

Due to linearity of the wavelet transform, the additive model (1) remains additive in the transform domain as well:

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \tag{5}$$

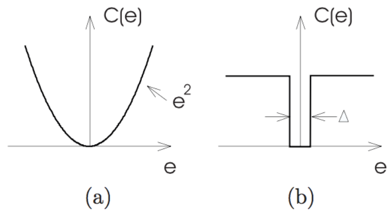

Fig. 4. Typical cost functions: (a) mean-square error and (b) uniform.

where $\mathbf{y} = \mathcal{W}_d\mathbf{d}$ are the observed wavelet coefficients, $\mathbf{x} = \mathcal{W}_d\mathbf{f}$ are the noise-free coefficients, $\mathbf{n} = \mathcal{W}_d\boldsymbol{\vartheta}$ is additive noise, and $\mathcal{W}_d$ is an operator that yields the discretized wavelet coefficients. An *orthogonal* wavelet transform maps the white noise in the input image into a white noise in the wavelet domain, with the same variance σ^2. In the case of *bi-orthogonal* and *non-decimated* [76] transforms, the noise variance differs from one wavelet subband to another, depending on the resolution level and on the subband orientation. The proportionality constant which relates noise variance in a given subband to the input noise variance is computed using the filter coefficients of the wavelet transform [79], [80].

While early work on wavelet-based image denoising was mostly concerned with defining shrinkage nonlinearities and deriving optimal uniform thresholds (fixed per whole subband) [35], [81], [82], it soon became clear that major gains could be achieved by adapting these shrinkage nonlinearities locally, depending on the spatial context around each coefficient [36], [83]–[90]. Many of these locally adaptive wavelet shrinkage methods were derived in a Bayesian framework, under a given prior for *spatial clustering* of important wavelet coefficients [83], [84], [88], [90]–[92], a prior for a local spatial activity indicator [85], [93] or joint statistics of neighbouring wavelet coefficients [36], [85], [87], [89], [94]–[98]. In the following, some of these concepts are outlined after a brief introduction into Bayesian estimation, signal versus noise characterization and subband statistics of natural images.

A. Bayesian shrinkage estimators

In general, Bayes' rules are shrinkers [99]–[101] and their shape in many cases has a desirable property: it can heavily shrink small arguments and only slightly shrink large arguments. The resulting actions on wavelet coefficients can be very close to thresholding. The Bayesian estimate $\hat{x}$ minimizes the *Bayes risk* $\mathcal{R}$, which is the expected value of a cost $C(x, \hat{x})$

$$\mathcal{R} \triangleq E\{C(x,\hat{x})\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(x,\hat{x})p_{X,Y}(x,y)dxdy \quad (6)$$

The cost function is chosen such to measure user satisfaction adequately and to yield a tractable problem [102]. Typically, the cost function depends only on the error of the estimate $x_\epsilon = \hat{x} - x$. Rewriting the joint density as $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y)$, the risk becomes

$$\mathcal{R} = \int_{-\infty}^{\infty} p_Y(y)dy \int_{-\infty}^{\infty} C(x-\hat{x})p_{X|Y}(x|y)dx \quad (7)$$

Two typical cost functions are illustrated in Fig. 4: the *mean-square error* cost accentuates the effects of large errors. Notice that in the resulting risk expression $\mathcal{R}_{ms} = \int_{-\infty}^{\infty} p_Y(y)dy \int_{-\infty}^{\infty}(x-\hat{x})^2 p_{X|Y}(x|y)dx$ the inner integral and $p_Y(y)$ are non-negative. Therefore $\mathcal{R}_{ms}$ is minimized by minimizing the inner integral. Differentiating the inner integral with respect to $\hat{x}$ yields

$$\frac{d}{d\hat{x}}\int_{-\infty}^{\infty}(x-\hat{x})^2 p_{X|Y}(x|y)dx$$
$$= -2\int_{-\infty}^{\infty} x p_{X|Y}(x|y)dx + 2\hat{x}\underbrace{\int_{-\infty}^{\infty} p_{X|Y}(x|y)dx}_{1} \quad (8)$$

Set this result to zero and the minimum mean-square error (MMSE) estimate $\hat{x}_{ms}$ follows as the *conditional mean*:

$$\hat{x}_{ms} = \int_{-\infty}^{\infty} x p_{X|Y}(x|y)dx \quad (9)$$

The *uniform cost* (also called 0-1 loss) function in Fig. 4(b) assigns zero cost to all errors less than $\pm\Delta/2$ where $\Delta > 0$ is arbitrarily small. The corresponding risk is $\mathcal{R}_{unf} = \int_{-\infty}^{\infty} p_Y(y)dy\Big[1 - \int_{-\hat{x}_{unf}-\Delta/2}^{-\hat{x}_{unf}+\Delta/2} p_{X|Y}(x|y)dx\Big]$. To minimize it, we maximize the inner integral. For arbitrarily small nonzero Δ, this is equivalent to choosing the value x at which the posterior density $p_{X|Y}(x|y)$ has its maximum [102, p.57]. Hence the name *maximum a posteriori* (MAP) estimate:

$$\hat{x}_{unf} = \hat{x}_{map} = \arg\max_x p_{X|Y}(x|y) \quad (10)$$

In many cases of interest, the MAP and the MMSE estimates coincide [102]. For a large class of cost functions the optimum estimate is the conditional mean whenever the a posteriori density is symmetric about the conditional mean.

Another class of Bayesian wavelet estimators shrinks a coefficient directly in proportion to the posterior probability that this coefficient contains an important signal component [83], [90], [93], [103]:

$$\hat{x}_{ProbShrink} = P(H_1|y)y \quad (11)$$

where H_1 denotes the hypothesis that the observation y contains a significant noise-free component, i.e., a signal of interest. This estimator can be motivated from the view point of *joint signal detection and estimation* [104] under certain assumptions [90], [105], [106].

Either we follow a classical Bayesian estimation approach or not, in most cases the denoising problem becomes that of minimizing a sum of two terms: an observation cost term, penalizing the estimation to depart form the observation, and a regularization term, penalizing the estimation to depart from the typical image behaviour. To understand this, observe that we can represent the prior model in terms of *a priori energy* $H(x)$, as $p_X(x) \propto \exp(-H(x))$ (see Section III-D for a concrete example, although with binary labels instead of x) and that $p_{Y|X}(y|x) \propto \exp(-D(x,y))$, where $D(x,y)$ is some distance between x and y. For i.i.d. Gaussian noise with standard deviation σ, this is simply $D(x,y) = \frac{(y-x)^2}{2\sigma^2}$. Maximizing the *a posteriori* probability $p_{X|Y}(x|y)$ is equivalent to maximizing its logarithm, which in its turn reduces to minimizing $\lambda H(x) + D(y,x)$, with $\lambda > 0$. Here, $H(x)$ is a regularization term and $D(y,x)$ is the observation cost term, enforcing data fidelity.

B. Signal versus noise characterization

Bayesian shrinkers described above require specification of the probability density functions of the ideal (noise-free) coefficients $p_X(x)$ on the one hand, and the pdf of the noisy coefficients $p_Y(y)$, or the likelihood $p_{Y|X}(y|x)$ on the other hand[1]. In fact, recognizing and modelling statistical differences between noise and uncorrupted images is key to denoising. The distinctive properties of image features versus noise make it possible to separate the noise from the image content in the first place, and all (not only Bayesian) methods make use of this insight, either explicitly or implicitly. While random image noise is typically characterized by a given probability density function, or with a mixture of multiple components with a relatively simple dependency on the signal strength and possibly with spatially varying parameters, depending on some external source, as it was hinted in Section II, the characterization of natural images is far more complex. There is a vast literature on natural image modelling; the reader is referred to [107]–[115].

One can identify at least four distinct paths to characterize typical image properties that noise does not have:

- *Sparsity in the analysis sense*: the image content gets "packed" into a few large coefficients, while the others can largely be neglected. Consequently, the statistics of images in the wavelet- and related domains strongly departs from Gaussian and follows highly kurtotic distributions [78] (see also Sections III-C, III-F).
- *Sparsity in the synthesis sense*: typical noise-free images can be expressed as linear combinations of relatively few spatially shifted atoms from a generic, suitable dictionary. This is related to *matching pursuit* [116] and *basis pursuit denoising* (BPDN) [41], [117]–[122]; we list some related methods in Section IV-C.
- *Spatial clustering*: typical images posses spatially connected geometric primitives (such as edges and contours), which produce spatially connected clusters of large coefficients. Moreover, the regularity of natural image discontinuities (i.e., relatively soft transitions of pixel intensities) guarantee survival of the large coefficients across the scales [123] and their reappearance at the same relative positions in different subbands, which is exploited in the *zerotree* coding scheme [124]. See also the methods in Section III-D and III-E.
- *Spatial repetitiveness*: typical images present various degree of self-similarity subject to spatial translation, which can also be extended to rotation and translation. This observation establishes a link between typical images and texture [125]. Indeed, the non-local means denoising algorithm [38], [39], [43] and derived ones were inspired by a texture synthesis method [126]. We return to this aspect in Sections III-G, IV-A and IV-B.

The seminal work in [123] and many follow-up papers make use of the fact that signal coefficients and noise coefficients *propagate* differently across the scales. The noise coefficients diminish rapidly as the scale increases, while the signal coefficients survive as a consequence of the different *local Lipschitz regularity* of the signal- and noise-induced discontinuities [123], [127].

The empirical probability density functions describing the rate of propagation of the noise-free and noisy wavelet coefficients across the scales have been reported in [90] for the case of AWGN. An intersting observation is that the pdf's of the wavelet estimators of the local Lipschitz regularity, named *average cone ratios* (ACR) in [90], are only slightly affected by the noise level (and for pure noise coefficients practically unaffected by the noise standard deviation). This makes denoising strategy with local Lipschitz regularity especially iteresting at relatively high noise levels, where the pdf's of the magnitudes of the noisy coefficients and pure noise are largely overlapping. This aspect as well as joint distributions of the magnitudes and their propagation across the scales was elaborated on in [90]. A general procedure for estimating the empirical pdf's of the noisy coefficients on the one hand and merely noise on the other hand, for unknown, *arbitrary* noise-types has been presented in [128]. Striking differences in the shape of *joint conditional histograms* of pairs of neighbouring wavelet coefficients of natural images and Gaussian noise have been shown in [109], [112].

C. Subband statistics, inter- and intrascale dependencies

We now look with somewhat more detail into the statistical properties of natural noise-free images in the wavelet domain. This subsection will also categorize the denoising methods to be reviewed next, based on the type of statistical dependencies among the coefficients that they incorporate.

As Fig. 5 (left) illustrates, histograms of noise-free wavelet coefficients of natural images are typically sharply peaked at zero (due to many nearly flat image regions) and long-tailed (because very large coefficients arise at positions of image discontinuities). The marginal pdf of image wavelet coefficients is hence well modelled by *highly kurtotic* models, such as generalized Laplacian (i.e., generalized Gaussian) [78], α-stable [129] and similar. Common marginal prior models for wavelet coefficients are also mixtures of two Gaussians (the *Bernoulli-Gaussian* model), where the mixing parameter is constant in a given subband [130] or estimated for each coefficient [84], [131], [132]. A number of methods to be reviewed in Section III-E and III-F, use *Gaussian scale mixture* models (GSM) [133], where each coefficient is modeled as the product of a positive scalar, and an element of a Gaussian random field.

Fig. 5 (right) illustrates also *inter-* and *intrascale dependencies* among the wavelet coefficients: notice that

[1]Notice that $p_{X|Y}(x|y) \propto p_{Y|X}(y|x)p_X(x)$.

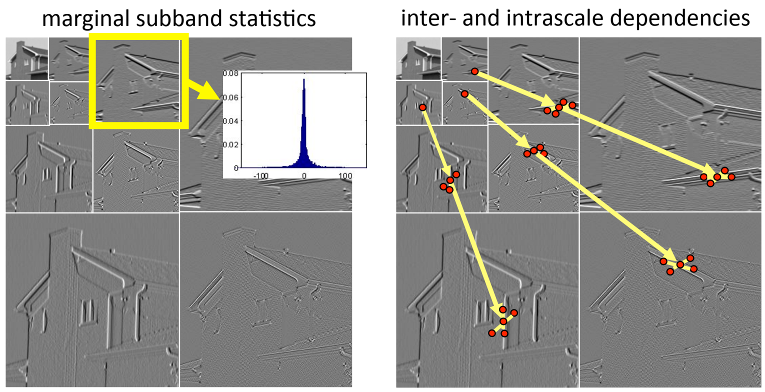

Fig. 5. **Left**: An illustration of the marginal subband statistics in noise-free images, manifesting itself in highly kurtotic histograms of wavelet coefficients. **Right**: An illustration of *intrascale* dependencies of image coefficients (large wavelet coefficients are typically spatially clustered) and *interscale* dependencies (large coefficients appear at relatively same positions in different subbands).

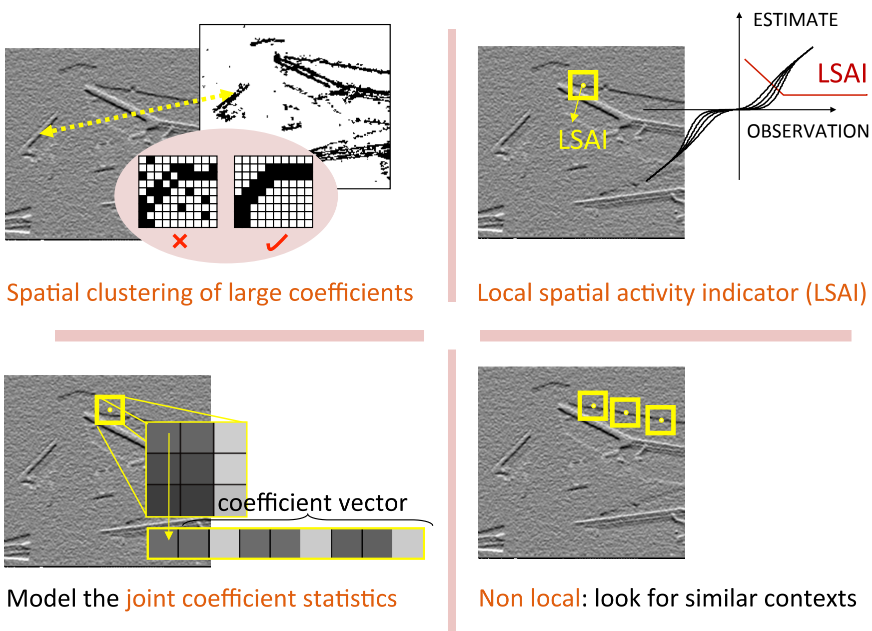

Fig. 6. Different ways of using coefficient dependencies for denoising. The interscale dependencies (although not depicted here) can be employed similarly within each category of the methods.

large coefficients are spatially clustered within the same subband, and appear at relatively the same positions at different scales. Joint conditional histograms of pairs of neighbouring wavelet coefficients (as well as of parent-child pairs) for natural images show a characteristic *bow-tie* shape [78], [109], [112].

Spatially adaptive transform domain methods typically model one of the following phenomena, depicted in Fig. 6:

- *spatial clustering* properties of large coefficients;
- *local spatial activity indicators*;
- *joint statistics* of groups of coefficients;
- *non-local* similarities of the spatial context.

Each of these categories is reviewed briefly next.

D. Methods based on spatial clustering models

In order to model spatial clustering properties, it is convenient to introduce *hidden labels* $l_i \in \{0, 1\}$, which mark each coefficient x_i as significant (if $l_i = 1$) or insignificant (if $l_i = 0$). The label l_i can be seen as a realization of a random variable L_i. By modelling the field $\mathbf{L} = \{L_1, ..., L_n\}$ as a *Markov Random Field* (MRF) [134], one can encode efficiently *a priori* knowledge about clustering of the significant and insignificant wavelet coefficients. The joint probability of a MRF is a Gibbs distribution, where the energy is represented as a sum of the so-called *clique potentials*:

$$P(\mathbf{L} = \mathbf{l}) = \frac{1}{Z} \exp\Big(-\frac{1}{T} \sum_{c \in \mathcal{C}} V_c(\mathbf{l})\Big) \tag{12}$$

Here Z denotes the normalizing constant (partition constant), T the temperature (which controls the peakedness of the distribution), c is a clique (a set of sites that are all neighbors of each other for a particular neighbourhood system), $\mathcal{C}$ the set of all possible cliques, and $V_c(\mathbf{l})$ the

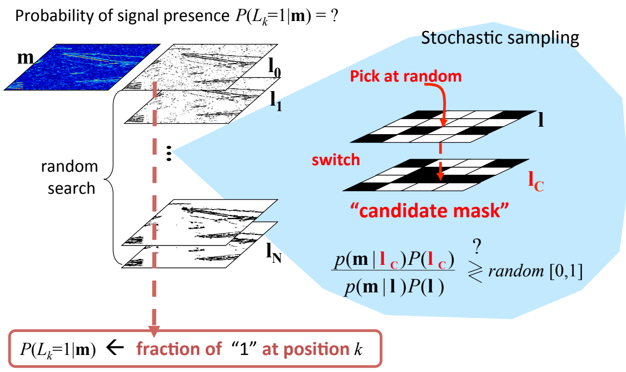

Fig. 7. An illustration of stochastic sampling, employed to estimate the probabilities that wavelet coefficients at different locations contain a significant signal component.

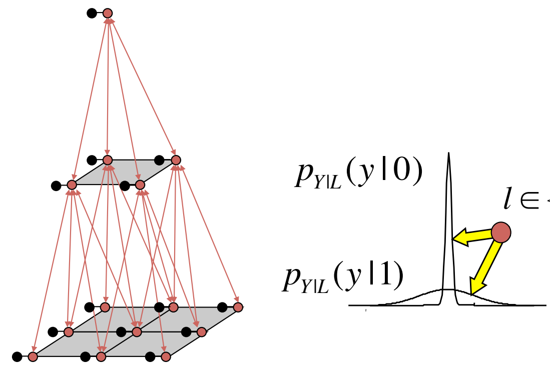

Fig. 8. An illustration of the HMT model. **Left**: a stochastic process on a quadtree, with hidden labels (red dots) attached to wavelet coefficients (black dots). **Right**: the marginal pdf as a mixture of two Gaussians.

clique potential, which is a function of all the labels l_i belonging to the clique c. Typically, the goal is to favour configurations where significant as well as insignificant coefficients form *spatially connected clusters*. Hence, negative potential is assigned to a clique consisting of the labels of the same type and a positive potential to a clique consisting of different labels. Some of the representative approaches [83], [86], [90], [103] apply *stochastic sampling* (e.g., the Metropolis sampler [134]) to estimate the probability that a given coefficient represents a signal of interest, given *all* observed coefficients (within the subband) and the assumed prior on their spatial clustering. Fig. 7 illustrates this stochastic sampling procedure, where the probability of signal presence is conditioned on a given significance measure of the wavelet coefficients, such as their magnitude $\mathbf{m} = \{m_1, ..., m_n\}$, $m_i = |y_i|$. The resulting probabilities $P(L_i = 1|\mathbf{m})$ can be plugged into various Bayesian estimators or used directly as shrinkage factors for the wavelet coefficients:

$$\hat{x}_i = P(L_i = 1|\mathbf{m})y_i \tag{13}$$

This estimator will act as a family of shrinkage nonlinearities, adapting itself to the local spatial context (presence of the detected geometrical structures) within each subband.

A closely related class of methods applies the *Hidden Markov Tree* (HMT) modeling framework [84], [92], [131], [132], [135], which is naturally related to a decimated wavelet representation where the *parent – child* relationships are captured on a *quadtree* structure (see Fig. 8). The pdf of noise-free wavelet coefficients is modelled by a Gaussian mixture model of [130]:

$$p(x_i) = p_i^0 \mathcal{N}(0, \sigma_{i,0}^2) + p_i^1 \mathcal{N}(0, \sigma_{i,1}^2) \tag{14}$$

where $p_i^0 = P(L_i = 0)$, $p_i^1 = 1 - p_i^0$. Each parent-child link has a state transition matrix, characterized with 'persistence probabilities' $p_i^{0\to0}$ and $p_i^{1\to1}$ and 'novelty probabilities" $p_i^{0\to1} = 1 - p_i^{0\to0}$ and $p_i^{1\to0} = 1 - p_i^{1\to1}$. The parameters, grouped into a vector $\boldsymbol{\theta}$ are computed by "upward-downward" algorithms through the tree, and model training detailed in [84]. Once the parameters are estimated, the wavelet coefficients are estimated as

$$\hat{x}_i = E(x_i|y_i, \boldsymbol{\theta}) = \sum_{q\in\{0,1\}} P(L_i = q|y_i, \boldsymbol{\theta}) \frac{\sigma_{q,i}^2}{\sigma_n^2 + \sigma_{q,i}^2} y_i \tag{15}$$

where σ_n is the noise standard deviation. Commonly mentioned problems are a large number of unknown parameters, convergence and lack of spatial adaptation. The spatial context can be introduced via an additional hidden label, leading to a local contextual HMT model [131].

E. Methods based on spatially local signal modelling

Spatial clustering methods from the previous Section model *global* clustering properties based on local interactions. E.g., the MRF-based approach models plausible support configurations (masks indicating the positions of important discontinuities) for the entire subband. Effective, low-complexity wavelet domain image denoisers can be also constructed by adapting a shrinkage non-linearity only locally, according to a value of a given *local spatial activity indicator*, such as the local variance, the locally averaged coefficient magnitude, the value of the "parent" coefficient or a certain combination of these.

The main idea is that such an activity indicator can provide a more reliable information about the presence of actual signal structures at the current position than the coefficient magnitude alone. This concept has been employed within a number of different schemes, some of which are briefly outlined below. In all of these, a family of shrinkage characteristics exists such that the coefficient of a given magnitude becomes less suppressed if the local spatial activity is higher and vice-versa (see an illustration in Fig. 6, upper right).

When the distribution of noise-free wavelet coefficients $p_X(x)$ is modelled as a product of a Gaussian random variable and a multiplier derived from its local surrounding, the MMSE estimate in (9) becomes the *locally adaptive Wiener* filter [85], [87], [89], [136]. For example, it can be observed that the histogram of the wavelet coefficients, in a given subband, *scaled* by their local standard deviations approaches well the Gaussian distribution [85]. By modelling the wavelet coefficients as conditionally independent zero mean Gaussian random variables, *given* their local variances, the MMSE estimator is a locally adaptive Wiener filter: $\hat{x}_l = \sigma_{y,l}^2 y_l/(\sigma_{y,l}^2 + \sigma_n^2)$, which is practically implemented as

$$\hat{x}_l = \frac{\hat{\sigma}_{y,l}^2}{\hat{\sigma}_{y,l}^2 + \sigma_n^2} y_l \tag{16}$$

where $\hat{\sigma}_{y,l}$ is an estimate of $\sigma_{y,l}$, formed based on the coefficients from a local window N_l. The maximum likelihood (ML) estimate of the signal variance is[2]

$$\hat{\sigma}_{y,l(ML)}^2 = \max\Big(0, \frac{1}{n}\sum_{i \in N_l} y_i^2 - \sigma_n^2\Big) \tag{17}$$

It has been observed that histograms of these estimates of the signal variance follow an exponential distribution, which motivated deriving the MAP estimate[3] under the exponential prior $p(\sigma_y) = \lambda e^{-\lambda \sigma_y^2}$, as [85]

$$\hat{\sigma}_{y,l}^2 = \max\Big(0, \frac{n}{4\lambda}\Big(-1 + \sqrt{1 + \frac{8\lambda}{n^2}\sum_{i \in N_l} y_i^2}\Big) - \sigma_n^2\Big) \tag{18}$$

where n denotes the number of the coefficients in a given subband. The resulting approach is known as *LAWMAP*, from locally adaptive window-based denoising using MAP. Related approaches include [136] and [87], [89] which employ Gaussian Scale Mixture models (the spatial activity indicator is a random multiplier for the Gaussian density).

An elegant *BiShrink* estimator employs the parent coefficient as a local spatial activity indicator. Denoting by y_l^p the parent of the coefficient y_l (the wavelet coefficient at the corresponding relative location in the coarser subband of the same orientation), the *BiShrink* estimator of [137] is

$$\hat{x}_l = \frac{\left(\sqrt{y_l^2 + (y_l^p)^2} - \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)_+}{\sqrt{y_l^2 + (y_l^p)^2}} y_l \tag{19}$$

where $(x)_+ = x$ if $x \geq 0$, and is zero otherwise. This estimator was derived under the MAP criterion, assuming an empirically established joint parent-child distribution $p(x_l, x_l^p) = \frac{3}{2\pi\sigma^2}\exp\left(-\frac{\sqrt{3}}{\sigma}\sqrt{x_l^2 + (x_l^p)^2}\right)$.

In essence, a well defined local spatial activity indicator refines estimation of the probability that a given coefficient represents a signal of interest. A spatially adaptive *ProbShrink* estimator of [90] defines the LSAI as a locally averaged coefficient magnitude $z_l = \frac{1}{|N(l)|}\sum_{k \in N(l)} |y_k|$ and estimates the noise-free coefficient as

$$\hat{x}_l = P(H_1|y_l, z_l) y_l = \frac{\eta_l \xi_l \mu}{1 + \eta_l \xi_l \mu} y_l \tag{20}$$

where $\eta_l = p(y_l|H_1)/p(y_l|H_0)$, $\xi_l = p(z_l|H_1)/p(z_l|H_0)$ and $\mu = P(H_1)/P(H_0)$, with H_0 and H_1 denoting the hypotheses that the signal of interest is absent or present in y_l, respectively. The required η_l, ξ_l and μ are estimated from the observed noisy coefficient histogram, under an appropriate prior for the noise-free coefficients, like the generalized Laplacian.

Fig. 9. An illustration of denoising digital camera image with spatially adaptive wavelet shrinkage. **Left**: original image. **Right**: the result of applying *ProbShrink* estimator from Eq (20).

F. Methods using joint statistics of groups of coefficients

As opposed to calculating a single number from a group of wavelet coefficients in order to describe the spatial activity in a local neighbourhood, one can also model groups of wavelet coefficients *jointly*. Such models can also suppress *correlated* noise where the previously described methods fail. The work of [89] is pioneering in this field, and has inspired others that we review here. For the purpose of clarity, we start from a vectorized version of the *ProbShrink* estimator from [98], which replaces (20) by

$$\hat{x}_l = P(H_1|\mathbf{y}_l) y_l \tag{21}$$

where $\mathbf{y}_l$ denotes a vector of wavelet coefficients from a window centred at l. Fig. 10 illustrates characteristic parts of this estimator on a simplified example, where $\mathbf{y}_i$ consists of two neighbouring coefficients only, showing the statistics of the corresponding noise-free vectors $\mathbf{x}_i$ and the resulting estimator from (21). An important aspect here is generalizing the definition of the signal of interest to a multivariate case. Typically, the signal of interest was defined as a noise-free component above a certain threshold $T = k\sigma$, where k is a proportionality constant and σ noise standard deviation. This can be written as: $S(x) = I(|x| \leq k\sigma) = I(|x/\sigma| \leq k)$, where $I(x)$ is the indicator function ($I(x) = 1$ if x is true and zero otherwise). A multivariate extension is:

$$S(\mathbf{x}) = I(\|\mathbf{C_n}^{-1/2}\mathbf{x}\| \leq k) \tag{22}$$

While the estimator (21) in general outperforms its simpler version with LSAI from (20), the advantage becomes especially obvious when it comes to the suppression of *correlated* noise. As Fig. 11 illustrates, *ProbShrink* from (20), which was quite successful in removing white noise from a digital camera image in Fig. 9, now completely fails in removing the correlated noise, while its vector version (21) removes even this difficult noise type remarkably well. This is because the univariate estimator (20) is blind for the noise structure, while its vector version (21) takes the noise structure into accounnt through the noise covariance matrix $\mathbf{C_n}$ appearing in (22). Similar performance is guaranteed by other wavelet domain methods using joint coefficient statistics, such as [36], [89], [95], [96], [98]

[2]This follows from $\hat{\sigma}_{y,l,ML}^2 = \arg\max_{\sigma_y^2 > 0} \prod_{i \in N_l} p(y_i|\sigma_y^2)$ when $p(y_j|\sigma_y^2) = \mathcal{N}(0, \sigma_y^2 + \sigma_n^2)$.

[3]The MAP estimate: $\arg\max_{\sigma_y^2 > 0} \prod_{i \in N_l} P(y_i|\sigma_y^2) p(\sigma_y^2)$.

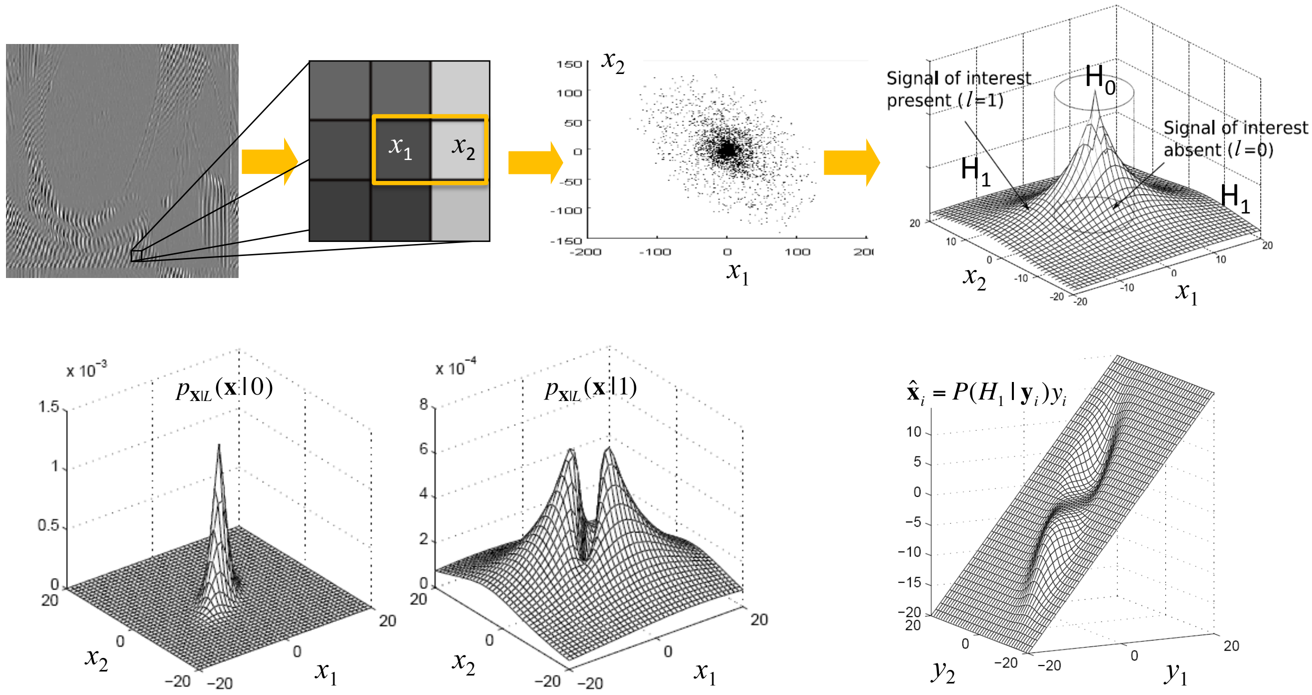

Fig. 10. An illustration of the vector *ProbShrink* estimator from Eq (21) on a simplified example where joint statistics is calculated for pairs of wavelet coefficients. Top row illustrates fitting a joint pdf, where H_0 and H_1 denote the hypotheses corresponding to the absence and the presence of the signal of interest, respectively. The bottom row shows the two joint conditional densities corresponding to H_0 and H_1 (i.e., $p_{\mathbf{X}|L}(\mathbf{x}|0)$ and $p_{\mathbf{X}|L}(\mathbf{x}|1)$, respectively, where the random variable L labels the central coefficient as insignificant or significant) and the resulting estimator.

Best known wavelet shrinkage method based on the joint statistics of the wavelet coefficients is the Bayesian Least Squares estimator using Gaussian Scale Mixture model (BLS-GSM) [36]. The coefficients within each local neighbourhood are modelled by a product of a Gaussian zero mean vector $\mathbf{u}_j$ and an independent positive scalar random variable $\sqrt{z}$, called random multiplier:

$$\mathbf{x}_j \overset{d}{=} \sqrt{z}\mathbf{u}_j \tag{23}$$

where $\overset{d}{=}$ denotes equality in distribution. Under this model, the vector $\mathbf{x}$ is an infinite mixture of Gaussian vectors, whose density is determined by the covariance matrix $\mathbf{C_u}$ and the mixing density $p_Z(z)$: $p_{\mathbf{X}}(\mathbf{x}_j) = \int p(\mathbf{x}_j|z)p_Z(z)dz$, with $p(\mathbf{x}_j|z) = \mathcal{N}(\mathbf{x}_j; 0, z\mathbf{C_u})$. The density of the observed vector $\mathbf{y}_j = \sqrt{z}\mathbf{u}_j + \mathbf{n}_j$ is also zero mean Gaussian, with covariance $z\mathbf{C_u} + \mathbf{C_n}$, which yields

$$\mathbb{E}(\mathbf{x}_j|\mathbf{y}_j, z) = z\mathbf{C_u}(z\mathbf{C_u} + \mathbf{C_n})^{-1}\mathbf{y}_j \tag{24}$$

The BLS estimator seeks the least squares estimate of the *central* coefficient x_j of the vector $\mathbf{x}_j$, which is the conditional mean

$$\hat{x}_j = \mathbb{E}(x_j|\mathbf{y}_j) = \int_0^\infty p_{Z|\mathbf{Y}}(z|\mathbf{y}_j)\mathbb{E}(x_j|\mathbf{y}_j, z)dz \tag{25}$$

In practical calculations, $\mathbb{E}(x_j|\mathbf{y}_j, z)$ is obtained by diagonalizing (24) and restricting it only to the central coefficient x_j, and the posterior distribution of the multiplier $p_{Z|\mathbf{Y}}(z|\mathbf{y}_j)$ is obtained from the Bayes' rule $p_{Z|\mathbf{Y}}(z|\mathbf{y}_j) = p_{\mathbf{Y}|Z}(\mathbf{y}_j|z)p_Z(z)/\int_0^\infty p_{\mathbf{Y}|Z}(\mathbf{y}_j|z)p_Z(z)dz$ under the assumed prior on z and using the normal distribution $\mathcal{N}(\mathbf{y}_j; 0, z\mathbf{C_u} + \mathbf{C_n})$ for $p_{\mathbf{Y}|Z}(\mathbf{y}_j|z)$, as detailed in [36]. This estimator has been since its introduction in 2003 among the most effective wavelet shrinkage methods to date. One limitation is that the signal covariance matrix is assumed to be constant within a subband up to a scaling factor. A richer representation and somewhat improved performance is hence offered by extensions, such as *spatially varying GSM* models [96] and *orientation adaptive GSM* [95]. A true generalization is offered by the *mixture of GSM*, known as MGSM [138], [139]. An efficient implementation of the MGSM model, called mixture of *projected* GSMs (MPGSM) was reported in [97]. Another extension is the so-called *fields of GSM* [140], which combines GSM with MRF modelling achieving thereby another level of adaptation to the actual signal structure. The improved performance comes at the expense of a more complex model.

G. Methods using non-local similarities

The fourth and the last category of spatially adaptive wavelet shrinkage methods from Fig. 6 employs *non local* dependencies among the transform coefficients and their surroundings. Natural images exhibit non-local self-similarities: relatively small image areas reappear often with exactly or nearly the same content at different places in the image. Grouping and processing together these similar image patches leads to the state-of-the-art denoisers.

Fig. 11. An illustration of correlated noise suppression using joint statistics in the wavelet domain. **Left**: Input image affected by correlated noise. **Middle**: the result of applying *ProbShrink* estimator from Eq (20). **Right**: the result of applying vector *ProbShrink* estimator from Eq (21).

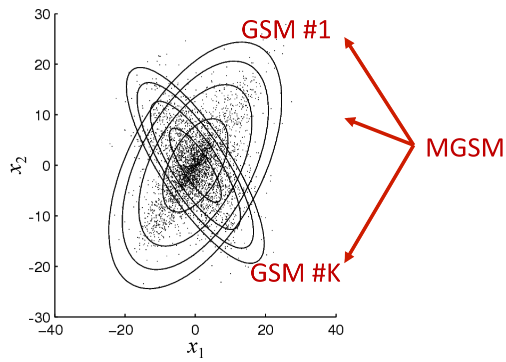

Fig. 12. An illustration of the Mixtures of Gaussian Scale Mixtures (MGSM) model of [138], [139].

Although this approach was originally applied in the image domain [38], [39], it can also be implemented in the wavelet domain or another transform domain, as well as in a hybrid way (combining the image and the transform domains) as in [42]. These methods will be addressed with some more detail in Section IV.

H. Gain from over-complete representations

All the denoising principles and estimators described so far in this Section are also directly applicable in other “wavelet-like” domains and pyramidal decompositions, such as the *dual tree complex wavelet* transform [141], [142] *steerable pyramids* [36], *curvelets* [68], *shearlets* [71] and *contourlets* [143]. These and related transforms have been proposed to improve the directional selectivity and approximation properties when it comes to analysing images and other multi-dimensional signals, which is also reflected in an improved denoising performance. Applying the same wavelet shrinkage rule in a redundant, non-decimated wavelet transform instead of the critically sampled (orthogonal) one, improves easily the signal-to-noise-ratio (SNR) of the denised image by 1dB or even more (see, e.g., [93], [144]). In the same way, porting the wavelet shrinkage estimator to a highly redundant and orientation selective transform, such as the steerable pyramid, curvelet or shearlet transform yields yet another significant improvement in terms of SNR as well as in terms of visual quality [71], [145], [146]. Figure 13 illustrates this with an example where the same estimator (in this particular case, *ProbShrink* from (20)) is applied in a non-decimated wavelet transform and in the curvelet domain. The artefacts due to unsuppressed noisy coefficients are now much less disturbing; these artefacts always reflect the shape of the base functions, resembling chekerboard patterns in the case of wavelets and faint lines with curvelets.

Fig. 13. The effect of using different multiscale transforms. **Top**: an original image and its noisy version. **Bottom**: the results of applying the same shrinkage rule (see text) in the wavelet domain (left) and in the curvelet domain (right). Notice different artefacts in these two results, each reflecting the particular base functions.

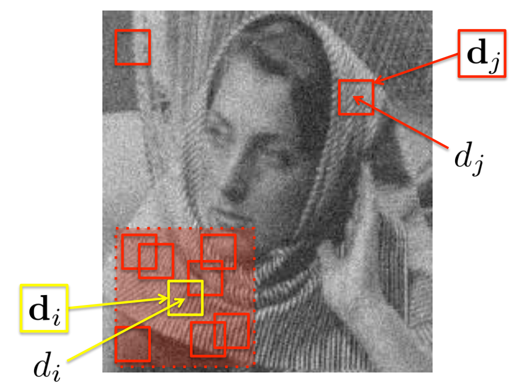

Fig. 14. An illustration of NLM from Eq (26), (27). In practice, it runs within a sliding window, as marked by the shaded block in this image.

IV. Patch-based methods

Patch based methods manipulate small image blocks (patches). By grouping similar contexts throughout the image, significant improvements are achieved over the more traditional pixel-based and local techniques.

A. Non-local means and generalizations

Averaging over a number of small areas that are equal up to an added white noise component, suppresses the noise while ideally preserving the underlying structure. Instead of searching explicitly for nearly identical neighbourhoods, one can allow merging *all* neighbourhoods with weighting factors that depend on their proximity to the neighbourhood of the processed pixel. This way, non-local methods estimate every pixel intensity based on information from the *whole* image, making use of the non-local pattern similarities. The idea comes from texture synthesis [126].

The *Non-Local Means* (NLM) filter [38], [39] considers the AWGN model (1) and estimates a noise-free pixel intensity as a weighted average of all pixel intensities in the image, where the weights are proportional to the similarity between the local neighbourhood of the pixel being processed and local neighbourhoods of surrounding pixels (see Fig. 14). Denote by $d_i = f_i + \vartheta_i$ a noisy pixel at position i and by $\mathbf{d}_i$ a small window of noisy pixels centered at i. The non-local means estimate of the noise-free pixel is

$$\hat{f}_i = \frac{\sum_j w(i,j) d_j}{\sum_j w(i,j)} \tag{26}$$

where the weights $w(i,j)$ measure mutual similarity between the two local neighbourhoods $\mathbf{d}_i$ and $\mathbf{d}_j$:

$$w(i,j) = \exp\Big(-\frac{||\mathbf{d}_i - \mathbf{d}_j||^2}{2h^2}\Big) \tag{27}$$

with a parameter $h > 0$. The complexity of NLM is quadratic in the number of pixels in the image, which makes the technique computationally intensive and even impractical in some real applications. The processing is typically confined to sliding windows rather than running over the whole image. Various improvements have been reported for enhancing the visual quality and reducing the computation time. These include better similarity measures [147], adaptive local neighbourhoods [148], and various algorithmic acceleration techniques [149]. With a pre-whitening filter, the NLM filter can be applied to correlated noise as well (see Fig. 15).

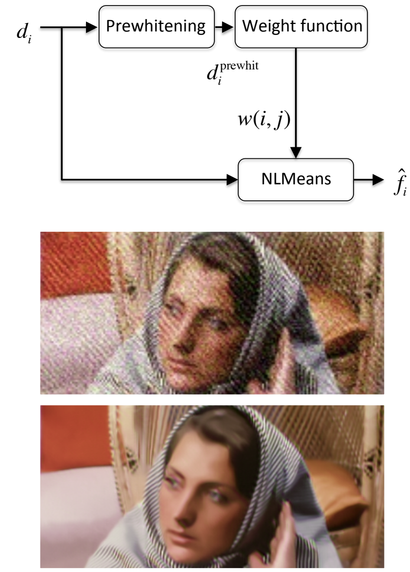

Fig. 15. A pre-whitening filter can be employed to adjust the computation of the weight coefficients of the NLM filter, such to enable suppression of correlated noise. Top-to bottom: an adjustment of NLM for correlated noise [149], an input image with coloured noise and the filtering result.

B. Collaborative filtering

The so-called *collaborative filtering* approach [42] groups similar 2D neighbourhoods, applies a 3D transformation (like 3D wavelet or discrete cosine transform) to these groups of similar 2D patches and removes noise from the 3D coefficients. After the inverse transform, the patches are returned to their corresponding locations. Fig. 16 illustrates this principle. Searching for similar 2D patches involves block matching; hence this approach was named in [42] BM3D (referring to block matching and 3D transformation). The noise reduction in each of the 3D cubes is performed by a shrinkage operation (like hard- or soft-thresholding or by Wiener filtering). Practically, the BM3D method is implemented in two stages: in the first stage, patch grouping and collaborative filtering with simple hard thresholding are applied to produce a basic estimate (i.e., an initial estimate of the noise-free image). This basic estimate is then exploited in the second stage to find more reliably self-similar noisy patches and to obtain an estimate of the true energy spectrum needed for the Wiener filtering of the noisy 3D cubes. The denoising superiority of BM3D has been attributed to *enhanced sparsity* in the transform domain: a 3D wavelet or a related transform applied to stacks of similar patches will yield much sparser coefficients than a collection of 2D transforms applied to each patch separately. A related method was introduced in [150] using both geometrically and photometrically similar patches to optimize the parameters of the Wiener filter.

Even though these methods provide current state-of-the-art in denoising, they are inherently limited by the efficiency of the underlying patch matching. The search for similar patches is most often restricted to a relatively small portion of the image, because allowing the search over the entire image tends to be prohibitively expensive. Hence, the denoising performance gets occasionally considerably degraded locally due to the lack of similar patches that could effectively be found. In response to these shortcomings, it has been studied how to make use of information from the entire image more efficiently, without having to visit all the spatial locations when searching for similar patches. E.g., a *global denoising* method of [45] employs the Nÿstrom method, to exploit the global image information while sampling only a fraction of the total pixels. The actual denoising performance was similar to BM3D, but the oracle performance showed promise for further research.

C. Dictionary learning based methods

Recent image processing methods often employ over-complete dictionaries of learned image atoms for sparsely

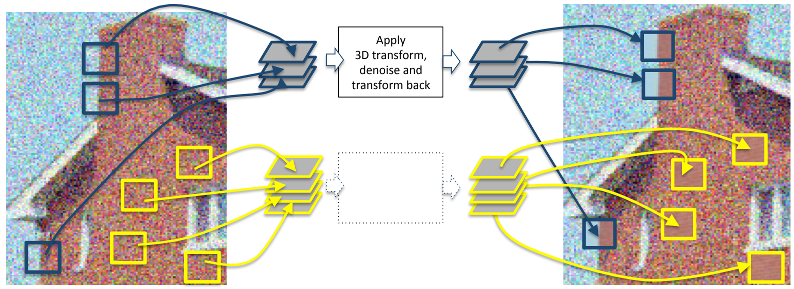

Fig. 16. An illustration of the BM3D method [42].

representing image patches. Instead of employing a predefined set of basis functions, these approaches learn a dictionary from examples. Triggered by the pioneering work [151] on learning a *sparse code* of natural images, numerous dictionary learning methods have been designed, K-SVD [41] being one of the most frequently used ones. Comprehensive tutorials on dictionary learning for sparse image representation include [120], [152].

The main idea of denoising methods based on dictionary learning is to present each image patch (of a predefined size) as a linear combination of relatively few atoms from the corresponding learned dictionary $\mathbf{\Gamma}$, while respecting the proximity with the degraded, observed image version. In practice, it means that the estimate of the underlying noise-free image $\hat{\mathbf{f}}$ and the sparse coefficients $\hat{\boldsymbol{\alpha}}_i$ for each of its patches $\mathbf{f}_i$ need to be simultaneously estimated. Denoting by $\mathbf{r}_i$ a vector that extracts i-th image patch of a given size from $\mathbf{f}$ (in a raster-scan fashion), the problem can be formulated as follows [118]:

$$\{\{\hat{\boldsymbol{\alpha}}_i\}_i, \hat{\mathbf{f}}\} = \arg\min_{\mathbf{f}} \Big(\lambda ||\mathbf{f} - \mathbf{g}||_2^2 + \sum_i \mu_i ||\boldsymbol{\alpha}_i||_0 + \sum_i ||\mathbf{\Gamma}\hat{\boldsymbol{\alpha}}_i - \mathbf{r}_i \mathbf{f}||_2^2 \Big) \qquad (28)$$

where λ and μ_i are positive constants. In this expression, the first term imposes the proximity between the measured image $\mathbf{d}$ and its denoised (and unknown) version $\mathbf{f}$. The second and the third terms make sure that in the constructed image every patch $\mathbf{f}_i = \mathbf{r}_i \mathbf{f}$ in every location i has a sparse representation with bounded error. In general, there are two options for training the dictionary $\mathbf{\Gamma}$ [118]: 1) using patches from the corrupted image $\mathbf{d}$ itself or 2) training on a corpus of patches taken from a high-quality set of images.

V. Choosing the Right Method

The ultimate objective of image denoising is to produce an estimate $\hat{\mathbf{f}}$ of the unknown noise-free image $\mathbf{f}$, which approximates it best, under given evaluation criteria. Like in any estimation problem, an important objective goal is to minimize the error of the result as compared to the unknown, uncorrupted data. A common criterion is minimizing the *mean squared error* (MSE)

$$\text{MSE} = \frac{1}{N} ||\mathbf{f} - \hat{\mathbf{f}}||^2 = \frac{1}{N} \sum_{i=1}^{N} (f_i - \hat{f}_i)^2 \qquad (29)$$

One can express the *signal to noise ratio* (SNR) in terms of the mean squared error as

$$\text{SNR} = 10 \log_{10} \frac{||\mathbf{f}||^2}{||\mathbf{f} - \hat{\mathbf{f}}||^2} = 10 \log_{10} \frac{||\mathbf{f}||^2 / N}{MSE} \qquad (30)$$

where SNR is in dB. In image processing, the *peak signal to noise ratio* (PSNR) is more common. For grey scale images it is defined in dB as

$$\text{PSNR} = 10 \log_{10} \frac{255^2 / N}{MSE} \qquad (31)$$

The aforementioned performance measures treat an image simply as a matrix of numbers and as such do not reflect faithfully *perceived visual quality*. For example, it is a well established fact that people tolerate certain amount of noise in an image better than a reduced sharpness and this cannot be captured by a mean squared error. Our visual system is also highly intolerant to various artifacts, like "blips" and "bumps" [34] in the reconstructed image. The importance of avoiding those artifacts is not only cosmetic; in certain applications (like astronomy, or medicine) such artifacts may give rise to wrong data interpretation. Moreover, the visual quality is highly subjective [153], and it is difficult to express it in objective numbers. There exist certain objective criteria, for expressing the degree of edge preservation (e.g., [154]). The *structural similarity index* (SSIM) [155] is a metric that was specifically designed for predicting the perceived quality of digital images. Image denoising methods are typically compared based on objective metrics (PSNR) and perceptual metrics, such as SSIM. For an overview of recent visual quality metrics and databases for image quality evaluation, see, e.g., [156].

Each of the different denoising approaches presented above may have certain advantages over the others, either in terms of the achievable quality of results, preservation of particular type of details or in terms of complexity, performance guarantees or capability to be extended to more general and mixed noise types, etc. Based on the many reported studies, including studies on the performance bounds [157] and recent comprehensive technical reviews [47], it can be concluded that *non-local* methods yield in most cases better results than the local ones in terms of PSNR and SSIM measures. Their performance is typically improved by using *multiresolution* representations. The use of *overcomplete* (redundant) representations offers improved performance for both local and non-local techniques, and the *adaptive bases* usually provide better performance than the fixed (non-adaptive) ones.

The improvements from non-local processing and from using adaptive bases typically come at a price of a considerably increased complexity and computation time. Moreover, even though non-local patch based methods yield superior quality in most cases, they are also prone to increased risks of deleting tiny non-repetitive image details and introducing falsely a non-exiting detail, due to combining the image content non-locally. Hence, the advantages and limitations of the different approaches should be evaluated in view of the priorities for a particular application. When denoising is applied to digital camera images or television images, methods that yield sharp and visually pleasing results should be preferred. In other applications, where the goal is to facilitate feature extraction or content interpretation (as it is often the case in remote sensing and medical imaging), some other priorities may be set, among which avoiding artefacts that could interfere with the image content.

VI. Future prospects

Patch-based processing exploiting self-similarity in the image is becoming a widely accepted approach in image processing in general, including image restoration and denoising. However, patch-based methods are strictly dependant on patch matching and their ability is hamstrong by the ability to reliably find sufficiently similar patches [45]. Hence, development of efficient strategies for finding self-similar patches (e.g., using smart indexing schemes and innovative hashing methods) is expected to have a huge impact on future developments in image processing in general, including denoising. A very interesting recent research line is incorporating the characteristics of the human visual system to measure the patch similarity [158].

Further on, dictionary learning methods in combination with non-local processing, offer a great potential for denoising. However, recent studies like [47] underline the problems arising from the *lack of structure* in the dictionaries employed by the current methods and identify the use of more structured dictionaries as one of the main research lines for advancing further image denoising. Some ideas in this respect can be drawn from the research on locally learned dictionaries [157], structured dictionary learning [159] and clustering-based sparse representations [160].

Interestingly, there is no formal model yet for expressing pattern repetitiveness in images and image redundancy, which is a surprising fact considering the importance and success of non-local patch based methods. Perhaps this is a promising line for future research.

Learning dictionaries of image atoms already makes a bridge between image processing and *machine learning* disciplines. Other connections with machine learning appear in recent image denoising works too [161]. *Deep learning* [162]–[164] as an emerging approach within the machine learning community is currently being explored as one of the new avenues in image denoising. In general, learning algorithms for deep architectures are centred around the learning of useful representations of data, which are better suited to the task at hand, and are organized in a hierarchy with multiple layers [165]. Representatives of deep learning methods for image denoising include [166], [167] based on *convolutional neural networks* (CNN) [168] and [169] based on *deep belief networks* (stacked autoencoders) [162], [170]. These approaches are already challenging in performance the state-of-the-art in image denoising and are rapidly progressing.

Acknowledgement

The author is grateful to the anonymous Reviewers and the Associate Editor for the constructive comments and suggestions that helped to improve substantially the quality of this paper.

References

[1] A. W. Doerry and F. M. Dickey, "Synthetic aperture radar," *Optics and Photonics News*, vol. 15, no. 11, pp. 28–33, 2004.
[2] J. W. Goodman, "Some fundamental properties of speckle," *J. Opt. Soc. Am.*, vol. 66, pp. 1145–1150, 1976.
[3] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 2, no. 2, pp. 165–168, Mar. 1980.
[4] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel, "Noise smoothing filter for images with signal-dependent noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 2, pp. 165–177, Mar. 1985.
[5] A. Papoulis, *Probability, random variables, and stochastic processes*. McGraw-Hill, 1984.
[6] I. Pitas and A. N. Venetsanopoulos, *Nonlinear Digital Filters, Principles and Applications*. Kluwer Academic Publishers, 1990.
[7] M. Gabbouj and E. Coyle, "Minimum mean absolute error stack filtering with structural constraints and goals," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 38, pp. 955–968, June 1990.
[8] I. Tabus, D. Petrescu, and M. Gabbouj, "A training framework for stack and Boolean filtering fast optimal design procedures and robustness case study," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 809–826, June 1996.
[9] R. Yang, L. Yin, M. Gabbouj, J. Astola, and Y. Neuvo, "Optimal weighted median filters under structural constraints," *IEEE Trans. Signal Process.*, vol. 43, pp. 591–604, Mar. 1995.
[10] L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo, "Weighted median filters: A tutorial," *IEEE Trans. Circuits Syst. II*, vol. 43, no. 3, pp. 15–192, Mar. 1996.
[11] J. Serra, *Image Analysis and Mathematical Morphology*. Academic Press, 1988.
[12] G. Ramponi, "The rational filter for image smoothing," *IEEE Signal Process. Lett.*, vol. 3, no. 3, pp. 63–65, 1996.
[13] M. Gabbouj and J. Astola, "Nonlinear order statistic filter design: Methodologies and challenges," in *Proc. of EUSIPCO.*, Tampere, Finland, 2000.
[14] T. F. Chan, S. Osher, and J. Shen, "The digital TV filter and nonlinear denoising," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 231–241, Feb. 2001.
[15] A. Chambolle, "An algorithm for total variation minimization and applications," *J. of Math. Imaging and Vision*, vol. 20, pp. 89–97, 2004.
[16] ——, "Total variation minimization and a class of binary MRF models," in *Energy Minimization Methods in Computer Vision and Pattern Recognition,*, ser. Lecture Notes in Computer Sciences, vol. 3757, Springer, 2005, pp. 136–152.
[17] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, pp. 629–639, 1990.
[18] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, 1992.

[19] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Proc.*, vol. 18, no. 11, pp. 2419–2434, 2009.

[20] M. Zhu, S. J. Wright, and T. F. Chan, "Duality-based algorithms for total-variation-regularized image restoration," *J. Comp. Optimization and Applications*, vol. 47, pp. 377–400, 2010.

[21] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, 2011.

[22] J. M. Bioucas-Dias and M. A. T. Figueiredo, "Multiplicative noise removal using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1720–1730, 2010.

[23] T. Goldstein and S. Osher, "The split Bregman method for l1-regularized problems," *SIAM J. Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[24] J. M. Fadili and G. Peyré, "Total variation projection with first order schemes," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 657–669, Mar. 2011.

[25] T. Goldstein, B. O'Donoghue, S. Setzer, and R. G. Baraniuk, "Fast alternating direction optimization methods," *SIAM J. Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2011.

[26] J. Liu, T.-Z. Huang, I. W. Selesnick, X.-G. Lv, and P.-Y. Chen, "Image restoration using total variation with overlapping group sparsity," *Information Sciences*, vol. 295, pp. 232–246, 2015.

[27] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision*, 1998, pp. 839–846.

[28] V. Aurich and J. Weule, "Non-linear Gaussian filters performing edge preserving diffusion," in *Proceedings of the DAGM Symposium*, 1995, pp. 538–545.

[29] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications," *Foundations and Trends in Computer Graphics and Vision*, vol. 4, no. 1, pp. 1–73, 2008.

[30] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, 2005.

[31] G. Motta, E. Ordentlich, I. Ramirez, G. Seroussi, and M. J. Weinberger, "The DUDE framework for continuous tone image denoising," in *IEEE Internat. Conference on Image Processing (ICIP)*, 2005, pp. 345–348.

[32] P. S. Windyga, "Fast impulsive noise removal," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 173–179, 2001.

[33] B. Smolka and A. Chydzinski, "Fast detection and impulsive noise removal in color images," *Real Time Imaging*, vol. 101, pp. 389–402, 2005.

[34] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.

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

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

[37] B. Rajaei, "An analysis and improvement of the BLS-GSM denoising method," *Image Processing On Line*, vol. 4, pp. 44–70, 2014.

[38] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in *Proc. Computer Vision and Pattern Recognition CVPR*, vol. 2, 2005, pp. 60–65.

[39] ——, "A review of image denoising methods, with a new one," *Multiscale Modeling and Simulation*, vol. 4, no. 2, pp. 490–530, 2006.

[40] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, "From local kernel to nonlocal multiple-model image denoising," *International Journal of Computer Vision*, vol. 86, no. 1, pp. 1–32, 2010.

[41] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[42] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Trans. Image Proc.*, vol. 16, no. 8, pp. 2080–2095, 2007.

[43] M. Lebrun, A. Buades, and J. M. Morel, "A nonlocal Bayesian image denoising algorithm," *SIAM Journal Image Science*, vol. 6, no. 3, pp. 1665–1688, 2013.

[44] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Image-specific prior adaptation for denoising," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5469–5478, Dec. 2015.

[45] H. Talebi and P. Milanfar, "Global image denoising," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 755–768, Feb. 2014.

[46] P. Chatterjee and P. Milanfar, "Is denoising dead?" *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 895–911, Apr. 2010.

[47] L. Shao, R. Yan, R. Li, and Y. Liu, "From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1001–1013, July 2014.

[48] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, pp. 679–698, Nov. 1986.

[49] A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy, "Despeckling SAR images using wavelets and a new class of adaptive shrinkage estimators," in *Proc. IEEE Internat. Conf. on Image Proc.*, Thessaloniki, Greece, 2001, pp. 233–236.

[50] C. Liu, R. Szeliski, S. Kang, C. Zitnick, and W. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 30, no. 2, pp. 299–314, 2008.

[51] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian, "Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1737–1754, Oct. 2008.

[52] S. Hasinoff, F. Durand, and W. Freeman, "Noise-optimal capture for high dynamic range photography," in *IEEE 23rd Conference on Computer Vision and Pattern Recognition, CVPR2010*, 2010, pp. 553–560.

[53] B. Aiazzi, L. Alparone, S. Baronti, M. Selva, and S. L., "Unsupervised estimation of signal-dependent CCD camera noise," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 231, 2012.

[54] B. Goossens, H. Luong, J. Aelterman, A. Pižurica, and W. Philips, "Realistic camera noise modeling with application to improved HDR synthesis," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 171, 2012.

[55] K. Egiazarian, K. Hirakawa, A. Pižurica, and J. Portilla, "Advanced Statistical Tools for Enhanced Quality Digital Imaging with Realistic Capture Models," *Special Section, EURASIP Journal on Advances in Signal Processing*, 2013.

[56] F. Luisier, T. Blu, and M. Unser, "Image denoising in mixed Poisson–Gaussian noise," *IEEE Trans. Image Processing*, vol. 20, no. 3, pp. 696–708, Mar. 2011.

[57] R. Zanella, P. Boccacci, L. , Zanni, and M. Bertero, "Efficient gradient projection methods for edge-preserving removal of Poisson noise," *Inverse Problems,*, vol. 25, no. 4, p. 045010, 2009.

[58] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," in *IEEE Internat. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 1109–1112.

[59] M. R. P. Homem, M. R. Zorzan, and N. D. A. Mascarenhas, "Poisson noise reduction in deconvolution microscopy," *Journal of Computational Interdisciplinary Sciences*, vol. 2, no. 3, pp. 173–177, 2011.

[60] A. Foi, "Clipped noisy images: Heteroskedastic modeling and practical denoising," *Signal Processing*, vol. 89, no. 12, pp. 2609–2629, Dec. 2009.

[61] T. H. Thai, R. Cogranne, and F. Retraint, "Camera model identification based on the heteroscedastic noise model," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 250–263, Jan. 2014.

[62] M. Mäkitalo and A. Foi, "Optimal inversion of the generalized Anscombe transformation for Poisson-Gaussian noise," *IEEE Trans. Image Processing*, vol. 22, no. 1, pp. 91–103, Jan. 2013.

[63] L. Azzari and A. Foi, "Variance stabilization for noisy+estimate combination in iterative Poisson denoising," *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1086–1090, 2016.

[64] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000, pp. 2272–2279.

[65] J. Li, C. Bai, Z. Lin, and J. Yu, "Penrose high-dynamic-range imaging," *Journal of Electronic Imaging*, vol. 25, no. 3, pp. 033 024:1–13, 2016.

[66] X. Jin and K. Hirakawa, "Analysis and processing of pixel binning for color image sensor," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 125, 2012.

[67] M. Mäkitalo and A. Foi, "Noise parameter mismatch in variance stabilization, with an application to Poisson-Gaussian noise esti-

mation," *IEEE Trans. Image processing*, vol. 23, no. 12, pp. 5348–5359, Dec. 2014.

[68] E. J. Candès and D. L. Donoho, "Curvelets, multiresolution representation, and scaling laws," in *SPIE, Wavelet Applications in Signal and Image Processing VIII*, vol. 4119, 2000.

[69] G. Kutyniok and T. Sauer, "Adaptive directional subdivision schemes and shearlet multiresolution analysis," *SIAM J. Math. Anal.*, vol. 41, pp. 1436–1471, 2009.

[70] G. R. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, July 2008.

[71] G. R. Easley and D. Labate, "Image processing using shearlets," in *Shearlets*, ser. Applied and Numerical Harmonic Analysis, G. Kutyniok and D. Labate, Eds. Springer, 2012, pp. 283–325.

[72] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1395–1411, 2007.

[73] M. S. Crouse, R. D. Nowak, and R. G. Baranuik, "Orthonormal bases of compactly supported wavelets," *Comm. Pure & Appl. Math.*, vol. 41, pp. 909–996, 1988.

[74] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

[75] 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.

[76] ——, *A wavelet tour of signal processing*. Academic Press, London, 1998.

[77] J. Weaver, Y. Xu, D. Healy, and J. Driscoll, "Filtering MR images in the wavelet transform domain," *Magn. Reson. Med*, vol. 21, pp. 288–295, 1991.

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

[79] S. Foucher, G. B. Bénié, and J. M. Boucher, "Multiscale MAP filtering of SAR images," *IEEE Trans. Image Proc.*, vol. 10, no. 1, pp. 49–60, Jan. 2001.

[80] A. Pižurica, "Image denoising using wavelets and spatial context modeling," Ph.D. dissertation, Ghent University, 2002.

[81] M. L. Hilton and R. T. Ogden, "Data analytic wavelet threshold selection in 2-D signal denoising," *IEEE Trans. Signal Process.*, vol. 45, pp. 496–500, 1997.

[82] M. Jansen, M. Malfait, and A. Bultheel, "Generalized cross validation for wavelet thresholding," *Signal Processing*, vol. 56, pp. 33–44, Jan. 1997.

[83] M. Malfait and D. Roose, "Wavelet-based image denoising using a Markov random field a priori model," *IEEE Trans. Image processing*, vol. 6, no. 4, pp. 549–565, Apr. 1997.

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

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

[86] M. Jansen and A. Bultheel, "Geometrical priors for noisefree wavelet coefficients in image denoising," in *Bayesian inference in wavelet based models*, ser. Lecture Notes in Statistics, P. Müller and B. B. Vidakovic, Eds., vol. 141. Springer Verlag, 1999, pp. 223–242.

[87] V. Strela, J. Portilla, and E. P. Simoncelli, "Image denoising using a local Gaussian scale mixture model in the wavelet domain," in *Proc. SPIE, 45th Annual Meeting*, San Diego, 2000.

[88] M. Jansen, *Noise Reduction by Wavelet Thresholding*. New York: Springer-Verlag, 2001.

[89] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain," in *Proc. IEEE Internat. Conf. on Image Proc.*, Thessaloniki, Greece, 2001.

[90] A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy, "A joint inter- and intrascale statistical model for wavelet based Bayesian image denoising," *IEEE Trans. Image Proc*, vol. 11, no. 5, pp. 545–557, May 2002.

[91] ——, "Image de-noising in the wavelet domain using prior spatial constraints," in *Proc. IEE Conf. on Image Proc. and its Applications IPA*, Manchester, UK, 1999, pp. 216–219.

[92] G. Fan and X. G. Xia, "Wavelet-based image denoising using hidden Markov models," in *Proc. IEEE International Conf. on Image Proc. ICIP*, Vancouver, BC, Canada, 2000.

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

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

[95] D. K. Hammond and E. P. Simoncelli, "Image modeling and denoising with orientation-adapted Gaussian scale mixtures," *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2089–2101, Nov. 2008.

[96] J. A. Guerrero-Colon, L. Mancera, and J. Portilla, "Image restoration using space-variant Gaussian scale mixtures in overcomplete pyramids," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 27–41, 2008.

[97] B. Goossens, A. Pižurica, and W. Philips, "Removal of correlated noise by modeling the signal of interest in the wavelet domain," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1153–1165, June 2009.

[98] ——, "Image denoising using mixtures of projected Gaussian scale mixtures," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1689–1702, Aug. 2009.

[99] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a Bayesian approach," *J. of the Royal Statist. Society B*, vol. 60, pp. 725–749, 1998.

[100] M. Clyde, G. Parmigiani, and B. Vidakovic, "Multiple shrinkage and subset selection in wavelets," *Biometrika*, vol. 85, no. 2, pp. 391–401, 1998.

[101] B. Vidakovic, "Wavelet-based nonparametric Bayes methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, ser. Lecture Notes in Statistics, D. D. Dey, P. Müller, and D. Sinha, Eds., vol. 133. Springer Verlag, New York, 1998, pp. 133–155.

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

[103] M. Jansen and A. Bultheel, "Empirical Bayes approach to improve wavelet thresholding for image noise reduction," *J. Amer. Stat. Assoc.*, vol. 96, no. 454, pp. 629–639, June 2001.

[104] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 434–443, May 1968.

[105] T. Aach and D. Kunz, "Spectral estimation filters for noise reduction in X-ray fluoroscopy imaging," in *Proc. European Conf. on Signal Process. EUSIPCO*, Trieste, 10-13 Sept., 1996.

[106] ——, "Anisotropic spectral magnitude estimation filters for noise reduction and image enhancement," in *Proc. IEEE Internat. Conf. on Image Proc. ICIP*, Lausanne, Switzerland, 16-19 Sept., 1996.

[107] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[108] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, p. 517548, 1996.

[109] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conference on Signals Systems, and Computers*, Pacific Grove, CA, 1997, pp. 673–678.

[110] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, 2001.

[111] E. P. Simoncelli, "Statistical modeling of photographic images," in *Handbook of Image and Video Processing*, A. Bovik, Ed. Academic Press, 2005, pp. 431–441.

[112] M. J. Wainwright, E. P. Simoncelli, and A. S. Wilsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, pp. 89–123, 2001.

[113] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. C. Zhu, "On advances in statistical modeling of natural images," *J. Math. Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.

[114] S.-C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 691–712, 2003.

[115] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer, 2009.

[116] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[117] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
[118] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
[119] J. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[120] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
[121] P. R. Gill, A. Wang, and A. Molnar, "The in-crowd algorithm for fast basis pursuit denoising," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4595–4605, Oct. 2011.
[122] Y. Kuang, L. Zhang, and Z. Yi, "Image denoising via sparse dictionaries constructed by subspace learning," *Circuits, Systems, and Signal Processing*, vol. 33, no. 7, pp. 2151–2171, July 2014.
[123] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Information Theory*, vol. 38, no. 8, pp. 617–643, Mar. 1992.
[124] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
[125] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Journal of Computer Vision*, vol. 40, no. 1, pp. 49–71, Oct. 2000.
[126] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1033–1038.
[127] S. Jaffard, "Pointwise smoothness, two-microlocalization and wavelet coefficients," *Publicacions Matemàtiques*, vol. 35, pp. 155–168, 1991.
[128] A. Pižurica, W. Philips, I. Lemahieu, and M. Acheroy, "A versatile wavelet domain noise filtration technique for medical imaging," *IEEE Trans. Medical Imaging*, vol. 22, no. 3, pp. 323–331, Mar. 2003.
[129] A. Achim and E. E. Kuruoğlu, "Image denoising using bivariate α-stable distributions in the complex wavelet domain," *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 17–20, Jan. 2005.
[130] 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.
[131] G. Fan and X. G. Xia, "Image denoising using local contextual hidden Markov model in the wavelet domain," *IEEE Signal Processing Letters*, vol. 8, no. 5, pp. 125–128, May 2001.
[132] J. Romberg, H. Choi, and R. Baraniuk, "Bayesian tree structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1056–1068, July 2001.
[133] D. Andrews, , and C. Mallows, "Scale mixtures of normal distributions," *J. Royal. Statist. Soc. B*, vol. 36, no. 1, pp. 99–102, 1974.
[134] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. Springer, 2009.
[135] J. Romberg, H. Choi, R. Baraniuk, and N. Kingsbury, "Multiscale classification using complex wavelets and hidden Markov tree models," in *Proc. IEEE Internat. Conf. on Image Proc. ICIP*, Vancouver, Canada, 2000.
[136] X. Li and M. Orchard, "Spatially adaptive denoising under overcomplete expansion," in *Proc. IEEE Internat. Conf. on Image Proc.*, Vancouver, Canada, Sept. 2000.
[137] L. Şendur and I. W. Selesnick, "Bivariate shrinkage with local variance estimation," *IEEE Signal Proc. Letters*, vol. 9, no. 12, pp. 438–441, Dec. 2002.
[138] J. Portilla and J. A. Guerrero-Colon, "Image restoration using adaptive Gaussian scale mixtures in overcomplete pyramids," in *Proc. SPIE Wavelets XII*, 2007, pp. 67 011F–1–15.
[139] J. A. Guerrero-Colon, E. P. Simoncelli, and J. Portilla, "Image denoising using mixtures of Gaussian scale mixtures," in *Proc. IEEE Internat. Conf. on Image Process. ICIP*, San Diego, CA, USA, 12-15 Oct., 2008, pp. 565–568.
[140] S. Lyu and E. P. Simoncelli, "Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 693–706, Apr. 2009.
[141] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, May 2001.
[142] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, Nov. 2005.
[143] A. L. Da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, Oct. 2006.
[144] R. R. Coifman and D. L. Donoho, "Translation-invariant denoising," in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim, Eds., Springer Verlag, New York, 1995, pp. 125–150.
[145] J. L. Starck, E. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, pp. 670–684, June 2002.
[146] L. Tessens, A. Pižurica, A. Alecu, A. Munteanu, and W. Philips, "Context adaptive image denoising based on joint image statistics in the curvelet domain," *Journal of Electronic Imaging*, vol. 17, no. 3, Sept. 2008.
[147] C. Kervrann, J. Boulanger, , and P. Coupé, "Bayesian Non-Local Means filter, image redundancy and adaptive dictionaries for noise removal," in *Proc. Int. Conf. on Scale Space and Variational Methods in Computer Visions (SSVM07)*, Ischia, Italy, 2007, pp. 520–532.
[148] C. Kervrann and J. Boulanger, "Optimal spatial adaptation for patch-based image denoising," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2866–2878, 2006.
[149] B. Goossens, H. Luong, A. Pižurica, and W. Philips, "An improved non-local denoising algorithm," in *Proc. Int'l Workshop on Local and Non-Local Approximation in Image Processing, LNLA2008*, Aug 23-24, Lausanne, Switzerland, 2008.
[150] P. Chatterjee and P. Milanfar, "Patch-based near-optimal image denoising," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1635–1649, Apr. 2012.
[151] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.
[152] I. Tošić and F. P., "Dictionary learning: What is the right representation for my signal?" *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.
[153] P. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. Bellingham, WA: SPIE Press, 1999.
[154] F. Sattar, L. Floreby, G. Salomonsson, and B. Lövström, "Image enhancement based on a nonlinear multiscale method," *IEEE Trans. Image Proc.*, vol. 6, no. 6, pp. 888–895, June 1997.
[155] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[156] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, and M. Carli, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 55–57, Jan. 2015.
[157] P. Chatterjee and P. Milanfar, "Clustering-based denoising with locally learned dictionaries," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1438–1451, July 2009.
[158] A. Foi and G. Boracchi, "Foveated nonlocal self-similarity," *Int. J. Computer Vision*, 2016, in press.
[159] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE Internat. Conf. Comput. Vision*, 2009, pp. 2272–2279.
[160] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit. (CVPR)*, Colorado, USA,, 2011, p. 457 464.
[161] I. Frosio, K. Egiazarian, and K. Pulli, "Machine learning for adaptive bilateral filtering," in *Proc. SPIE Image Processing: Algorithms and Systems XIII*, vol. 9399, 2005.
[162] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *IEEE Trans. Signal Process.*, vol. 18, pp. 1527–1554, 2006.
[163] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
[164] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[165] H. Li, “Deep learning for image denoising,” *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 7, no. 3, pp. 171–180, 2014.

[166] V. Jain and S. H. Seung, “Natural image denoising with convolutional networks,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 21, 2008.

[167] X. Wang, Q. Tao, L. Wang, D. Li, and M. Zhang, “Deep convolutional architecture for natural image denoising,” in *International Conference on Wireless Communications and Signal Processing (WCSP)*, 2015.

[168] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[169] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” in *Advances in Neural Information Processing Systems 25*, 2012, pp. 350–358.

[170] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.