Local adaptivity to variable smoothness for exemplar-based image regularization and representation

Charles KERVRANN & Jérôme BOULANGER

IRISA / INRIA Rennes, Campus Universitaire de Beaulieu, F-35042 Rennes Cedex
INRA, UR341 Mathématiques et Informatique Appliquées, F-78352 Jouy-en-Josas

Abstract. A novel adaptive and exemplar-based approach is proposed for image restoration (denoising) and representation. The method is based on a pointwise selection of similar image patches of fixed size in the variable *neighborhood* of each pixel. The main idea is to associate with each pixel the weighted sum of data points within an adaptive neighborhood. We use small image patches (e.g. 7×7 or 9×9 patches) to compute these weights since they are able to capture local geometric patterns and texels seen in images. In this paper, we mainly focus on the problem of adaptive neighborhood selection in a manner that balances the accuracy of approximation and the stochastic error, at each spatial position. The proposed pointwise estimator is then iterative and automatically adapts to the degree of underlying smoothness with minimal *a priori* assumptions on the function to be recovered. The method is applied to artificially corrupted real images and the performance is very close, and in some cases even surpasses, to that of the already published denoising methods. The proposed algorithm is demonstrated on real images corrupted by non-Gaussian noise and is used for applications in bio-imaging.

KEYWORDS: exemplar-based methods, estimation, bias-variance trade-off, restoration, denoising, nonlinear filtering, detection, energy minimization, bio-imaging.

1. Introduction

Traditionally, the problem of image recovering is to reduce undesirable distortions and noise while preserving important features such as homogeneous regions, discontinuities, edges and textures. This can be accomplished by taking into account local geometries and statistics during the filtering process. Popular image restoration algorithms are therefore nonlinear to reduce the amount of smoothing near abrupt changes:

— Most of the more efficient regularization methods are based on discrete [34, 7] or continuous [64, 72] energy functional minimization since they are designed to explicitly account for the image geometry, involving the adjustment of global weights that balance the contribution of prior smoothness terms and a fidelity term to data. Related partial differential equations (PDE) and variational methods, including anisotropic diffusion [67, 87, 9] and total variation (TV) minimization [72, 17], have shown impressive results to tackle the problem of edge-preserving smoothing [67, 16, 87, 84] and image decomposition into geometric, textured and noise components [60, 66, 1].

— For reasons of performance in computer vision, other smoothing algorithms aggregate information according to a spatial criterion and a brightness criterion. The *neighbor-*

hood filters [53, 40, 76], including the so-called *bilateral filtering* [83, 5], use this generic principle, involving the local weighted averaging of input data over a spatial neighborhood.

In most cases, the global amount of smoothing being performed is controlled by a relatively small number of parameters. Nevertheless, when local characteristics of the data differ significantly across the domain, setting these control parameters is probably not satisfying. Some efforts have been initiated in this direction to determine local scales of significant image features and detect non-stationarities in images for better spatial adaptation and regularization [58, 9, 20, 38, 48].

But, what makes image restoration a difficult task, is that natural images often contain many irrelevant objects. This type of "noise" is sometimes referred to as "clutter". To develop better image enhancement algorithms that can deal with structured noise, we need non-parametric models to capture all the regularities and geometries seen in local patterns. In contrast to the usual *neighborhood filters* and PDEs-based filters [17], a line of work consists then in modeling non-local interactions from training data or a library of natural images. The idea is to improve the traditional Markov random field (MRF) models by learning potential functions over extended neighborhoods as initiated in [93]. In [71], the authors also proposed to use a library of examples in the training step. In the meanwhile, it has been experimentally confirmed that non-parametric patch-based approaches are very effective 2D texture synthesis [27], image inpainting [22], image reconstruction and super-resolution [32] or image-based rendering [31]. Similarities between image patches have been also used in the early 90's for texture segmentation [35, 47]. More recently, the redundancy property observed in many images has been successfully exploited by Buades, Coll and Morel [14, 15] and Awate and Whitaker [3] for image denoising, and is a key ingredient at the origin of fractal denoising methods [36]. Also, this idea was early and independently suggested for Gaussian noise reduction in [24] and for impulse noise removal in [86, 92].

In our non-parametric estimation framework, we will also assume that small image patches selected in the variable neighborhood of a pixel capture the essential information required for local restoration. Unlike most existing exemplar-based MRF methods that use training sets for learning and denoising [32, 71], the proposed restoration approach is unsupervised and based on the key idea of iteratively increasing a window at each pixel and adaptively weighting the input data. As in [14, 15], the data points with a similar patch to the reference patch will have larger weights in the average. However, unlike Buades *et al.* [14, 15], we address the central problem of choosing the smoothing window (neighborhood) which can be different at each pixel to cope with spatial inhomogeneities across the image domain. Our main contribution is then to use a change-point detection procedure, initiated by Lepskii for 1D signals [55]. The Lespkii's principle, also rooted in the ideas of wavelet representation [25, 26, 45], is a procedure which aims at minimizing the pointwise L_2 risk of the estimator. The local optimization amounts to balancing the accuracy of approximation and the stochastic error, at each spatial position. This *pointwise adaptive estimation* approach has been described in its general form and in great details in [55, 56], and the interested readers should of course have a look at these milestone papers. Throughout this paper, we will show how our approach based on this idea of *pointwise adaptive estimation* improves the results obtained using the *Non-Local means filter* [14, 15]. Other related works to our approach are *neighborhood filters* [40, 83, 5, 89,

63], information-theoretic adaptive filtering [3] and statistical regularization schemes [77, 69, 46], enhanced via incorporating either a variable window scheme or exemplar-based weights.

The remainder of the paper is organized as follows. Related studies are presented in Section 2. In Section 3, we introduce the image modeling and some notations. In Section 4, we formulate the problem of the selection of the "best" possible window and present the adaptive estimation procedure. We describe a practical algorithm with well calibrated parameters for image denoising. In Section 5, technical arguments for this iterative procedure are presented: given a window size, we perform a one-step of a fixed-point iteration to solve the optimality conditions of an energy functional involving non-local interaction terms. In Section 6, theoretical and statistical properties of the pointwise adaptive estimator are briefly presented. In Section 7, we demonstrate the ability of the method to restore corrupted images with artificial additive white Gaussian noise (WGN). We compare our results to the published results and show our approach compares favorably to very competitive methods, included the recent wavelet-based methods [26, 81, 70, 68, 54], when applied to a commonly-used image dataset [70]. Also, the method is applied to denoise real images with artifacts and corrupted by non-Gaussian noise, and fluorescence microscopy images in bio-imaging.

2. Neighborhood filters and related methods

In this section, the relationships between the proposed technique and other image restoration methods are discussed. We focus on iterative and non-iterative neighborhood filters introduced in statistical and variational frameworks. This includes nonlinear Gaussian filters, M-estimators from robust statistics, nonlinear diffusion and some related energy minimization methods since they share some common points with the proposed method (see also [14] for a recent survey).

First, it is well known that fixed-window methods yield good results when all the pixels in the window come from the same population as the central pixel. However, difficulties arise when the square (or circular) window overlaps a discontinuity. Filtering with a window (or spatial kernel) that is symmetric around the central pixel results in averaging information from different regions in the vicinity of edges. In such a case, a possible strategy is to substitute the border pixel with a pixel inside the object. Nitzberg & Shiota [65] (see also [30]) proposed an offset term that displaces spatial kernel centers away from the presumed edge location, thus enhancing the contrast between adjacent regions without blurring their boundaries. Intuitively, the displacement direction is calculated from the dominant local gradient direction or, more robustly, given by the *structure tensor* [88, 84]. Sensitivity to noise is then reduced, but the user-defined setting of a global scale of features to be preserved, is required [88]. An other class of nonlinear filters aims at estimating a connected component of "homogeneous" pixels, that can be of arbitrary shape, containing the point of interest [91, 11]. One primarily and typical filter based on this principle, is the sigma filter [53] and continuous versions are the Lee's [53], Saint-Marc's [73] and Susan [76] filters. More recently, Buades *et al.* proposed the so-called *Non-Local means*

filter [14, 15] of the following form

$$\widehat{u}(\mathbf{x}_i) \;=\; \frac{\sum_{\mathbf{x}_j \in \Omega} L_g(\mathbf{Y}_i - \mathbf{Y}_j)\, K_h(\mathbf{x}_i - \mathbf{x}_j)\, Y_j}{\sum_{\mathbf{x}_j \in \Omega}\; L_g(\mathbf{Y}_i - \mathbf{Y}_j) K_h(\mathbf{x}_i - \mathbf{x}_j)} \tag{1}$$

where $K_h(\cdot) = (1/h)K(\cdot/h)$ and $L_g(\cdot) = (1/g)L(\cdot/g)$ are rescaled versions of non-negative kernel functions, $\mathbf{Y}_j$ denotes a vector of pixel values taken in the neighborhood of a point $\mathbf{x}_i$ belonging to the image domain Ω. The similarity between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is measured by the Euclidean distance $\|\mathbf{Y}_i - \mathbf{Y}_j\|_2$ between two vectorized image *patches*. We can control the spatial support of the filter by varying h and the level of blurring by varying g. Both parameters are set manually according to the image contents and the signal-to-noise ratio. In [14, 15], it has been clearly demonstrated that the *NL-means filter* for denoising improves the state-of-the-art (see also [51]).

The *NL-means filter* may be regarded as a generalization of previous *neighborhood filters*. Indeed, if the size of the patch is reduced to one pixel and $L_g(\cdot)$ and $K_h(\cdot)$ are Gaussian kernels, the *NL-means filter* is then equivalent to *bilateral filtering* [83]. In the meanwhile, other connections between bilateral filtering and better understood methods have been investigated. First, emphasizing the importance of extended neighborhoods, Barash & Comaniciu [6] have showed that bilateral filtering represents a weighted averaging algorithm which turns out to be a special implementation of nonlinear diffusion [67, 87], controlled by a global scale parameter. Elad [28] established further how the bilateral filter is algorithmically related to anisotropic diffusion [67] and robust estimation [8] in terms of minimizing functionals. The bilateral filter can also be viewed as an Euclidean approximation of the Beltrami flow and originates from image manifold area minimization [79, 80]. Barash & Comaniciu showed that kernel density estimation applied into the joint spatial-range domain yields a powerful processing paradigm - the *mean-shift* procedure [21] - also related to bilateral filtering but having additional flexibility [6]. The link between iterative *mean-shift* algorithm, local mode filtering, clustering, local M-estimators, nonlinear diffusion, regularization approaches were already analyzed in [18, 89, 85, 28, 63, 6]. Also, all these methods have been casted into a unified framework for functional minimization combining nonlocal data and nonlocal smoothness terms in [63]. In particular, Mrazek *et al.* emphasized the large amount of structural similarities between the iterated bilateral filter and the local M-smoother [19]. It is confirmed that local M-smoothing uses the initial image in the averaging procedure and searches for the minimum of a local criterion whereas iterated bilateral filtering uses the evolving image and has to stop after a certain number of iterations in order to avoid a flat image. To complete the state of the art related to neighborhood filters, we mention the *digital Total Variation (TV) filter* [17], inspired by the ROF (Rudin-Osher-Fatemi) minimizing process which simultaneously computes a piecewise smooth solution and estimates the image discontinuities [72]. The ROF method essentially penalizes image derivatives and is one of the most successful tool for image restoration. We will also compare our results to those obtained with this famous method [72] in Section 7.

Finally, all previous neighborhood filters [72, 17, 83, 14] have a relatively small number of smoothing parameters that control the global amount of smoothing being performed. In implementing these filters, the first question to be asked is how should the smoothing parameters be chosen ? A number of authors have turned to various statistical tools such as bandwidth selection, to be exploited for global parameter selection in the diffusion

process [18, 89, 85, 28, 63, 6, 38, 37]. However, it is also theoretically confirmed that automatically determining a bandwidth for spherical kernels is a difficult problem [40], and the bandwidths involved in bilateral filtering are usually chosen to give a good visual impression and are heuristically chosen [83, 5]. Furthermore, when local characteristics of the data differ significantly across the domain, selecting optimal and global smoothing parameters is probably not satisfying. There is seldom a single scale that is appropriate for a complete image. These difficulties motivated the development of more adaptive methods to cope with inhomogeneities in images. First, Aurich *et al.* proposed to use a chain of nonlinear filters for which the bandwidths vary at each iteration according to specific and deterministic rules to balance edge detection and noise removal [2]. In [42], the local amount of Gaussian smoothing is computed in terms of variance in a space-scale framework, through the minimal description length criterion (MDL). The local variance is actually useful for localization of significant image features as explained in [9, 38, 42]. An alternative way to select the local scale is to maximize a measure of edge strength over scale space [58] but the resulting scale computed from image derivatives, is sensitive to signal-to-noise ratio. More recently, the TV flow has been suggested since it includes a non-explicit scale parameter useful for detecting the scale of image features [12].

In this paper, we will also focus on this problem which is an open research issue, and propose a stable scheme to select the "best" possible neighborhood for local restoration and smoothing as in [69]. In addition, we propose to use image patches to take into account complex spatial interactions in images. Our adaptive smoothing works in the joint spatial-range domain as the *Non-Local means* filter [15]. However, it has a more powerful adaptation to the local structure of the data since the neighborhood sizes and the control parameters are estimated from local image statistics as presented in the remainder of the paper.

3. Image redundancy and basic idea

In order to describe our estimation method, let us first introduce some useful notations. Consider the following image model

$$Y_i = u_{\text{true}}(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \ldots, |\Omega| \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^d$, $d \geq 2$, represents the spatial coordinates of the discrete image domain Ω of $|\Omega|$ pixels, and $Y_i \in \mathbb{R}_+$ is the observed intensity at location $\mathbf{x}_i$. We suppose the errors ϵ_i to be iid (independent identically distributed) Gaussian zero-mean random variables with unknown variance σ^2. Our aim is then to recover $u_{\text{true}} : \mathbb{R}^d \to \mathbb{R}_+$ from noisy observations Y_i. In what follows, we will restrict to functions having a minimal regularity: for $0 < C_0 < \infty$ and $0 < C_1 < \infty$, we assume that $\sup_{\mathbf{x}_i \in \Omega} |u_{\text{true}}(\mathbf{x}_i)| < C_0$, and $\forall \mathbf{x}_i, \mathbf{x}_j \in \Omega$, $|u_{\text{true}}(\mathbf{x}_j) - u_{\text{true}}(\mathbf{x}_i)| \leq C_1 |\mathbf{x}_j - \mathbf{x}_i|$ (u_{true} is Lipschitz continuous).

In order to recover u_{true} from noisy observations, we need additional prior assumptions on the structure of the image. In particular, we will assume that the unknown image $u_{\text{true}}(\mathbf{x}_i)$ can be approximated by the weighted average of input data over a variable neighborhood Δ_i around that pixel $\mathbf{x}_i$. The points $\mathbf{x}_j \in \Delta_i$ with a similar patch $\mathbf{u}_j$ to the reference image patch $\mathbf{u}_i$ will have larger weights in the average as in [14, 15]. This amounts to supposing that there exists some stationarity in the neighborhood of a point

$\mathbf{x}_i$ which can help to recover $u(\mathbf{x}_i)$. In what follows, $\mathbf{u}_i$ at $\mathbf{x}_i$ will denote indifferently a fixed size square window of $\sqrt{p} \times \sqrt{p}$ pixels or a vector of p elements where the pixels are concatenated along a fixed lexicographic ordering. As with all exemplar-based techniques, the size p of image patches must be specified in advance according to how stochastic the user believes the image to be [27, 22]. Note that our ambition is not to learn generic image priors from a database of image patches as proposed in [93, 32, 44, 71]. We only focus on image patches as non-local image features able to capture local geometries, and adapt non-parametric regression techniques for image restoration.

An important question we address in this paper is how to determine the size and shape of the variable neighborhood Δ_i at each pixel, from image data. The selected neighborhood must be different at each pixel to take into account the inhomogeneous regularity of the image. For the sake of parsimony and computational efficiency, we assume that the set $\mathcal{J}$ of admissible neighborhoods is finite and will be arbitrarily chosen as a geometric grid of nested square windows

$$\mathcal{J} = \{\Delta_{i,n} : |\Delta_{i,n}| = (2^n+1) \times (2^n+1), n=1, \ldots, N\},$$

where $|\Delta_{i,n}| = \#\{\mathbf{x}_j \in \Delta_{i,n}\}$ is the cardinality of $\Delta_{i,,n}$ and N is the number of elements of $\mathcal{J}$. For technical reasons, we will require the following conditions: $\Delta_{i,n}$ is centered at $\mathbf{x}_i$ and $\Delta_{i,n} \subset \Delta_{i,n+1}$. In the next sections, we will precise the adaptive estimation procedure and describe a local window selector which achieves two objectives: spatial adaptivity and computational efficiency. We will introduce the notion of *local L_2 risk* as an objective criterion to guide the optimal selection of the smoothing window for constructing the "best" possible estimator. This optimization will be mainly accomplished by starting, at each pixel, with a small window $\Delta_{i,0}$ and a *pilot* estimator $\widehat{u}_{i,0}$, and increasing $\Delta_{i,n}$ with n. The "performance" of the estimator $\widehat{u}_{i,n}$ is then improved at each iteration of the procedure while the estimation window is not too large, according to the so-called *"bias-variance trade-off"* explained in Section 5.2.

The proposed approach requires no training step and may be then considered as *unsupervised*. This makes the method very attractive for computer vision applications.

4. Adaptive estimation procedure

The proposed procedure is iterative and works as follows [69, 48].

At the initialization, we choose a local window $\Delta_{i,0}$ containing only the point of estimation $\mathbf{x}_i$. A first estimate $\widehat{u}_{i,0}$ (and its variance $v^2(\widehat{u}_{i,0})$) is then

$$\widehat{u}_{i,0} = Y_i \quad \text{and} \quad v^2(\widehat{u}_{i,0}) = \widehat{\sigma}^2 \tag{3}$$

where an estimated variance $\widehat{\sigma}^2$ has been plugged in place of σ^2 since the variance of errors is supposed to be unknown. At the next iteration, a larger window $\Delta_{i,1}$ with $\Delta_{i,0} \subset \Delta_{i,1}$ centered at $\mathbf{x}_i$ is considered. Every point $\mathbf{x}_j \in \Delta_{i,1}$ gets a weight[1] $\pi_{i \sim j,1}$ defined by comparing pairs of previous estimated patches $\widehat{\mathbf{u}}_{i,0} = (\widehat{u}_{i,0}^{(1)}, \cdots, \widehat{u}_{i,0}^{(p)})^T$ and $\widehat{\mathbf{u}}_{j,0} = (\widehat{u}_{j,0}^{(1)}, \cdots, \widehat{u}_{j,0}^{(p)})^T$, obtained at the first iteration. As usual, the points $\mathbf{x}_j$ with a similar patch $\widehat{\mathbf{u}}_{j,0}$ to the reference patch $\widehat{\mathbf{u}}_{i,0}$ will have weights close to 1 and 0 otherwise. Then, we recalculate

---

[1] The subscript $i \sim j$ means "$\mathbf{x}_j \in \Delta_{i,\cdot}$ *and the index j runs through the neighborhood of* $\mathbf{x}_i$".
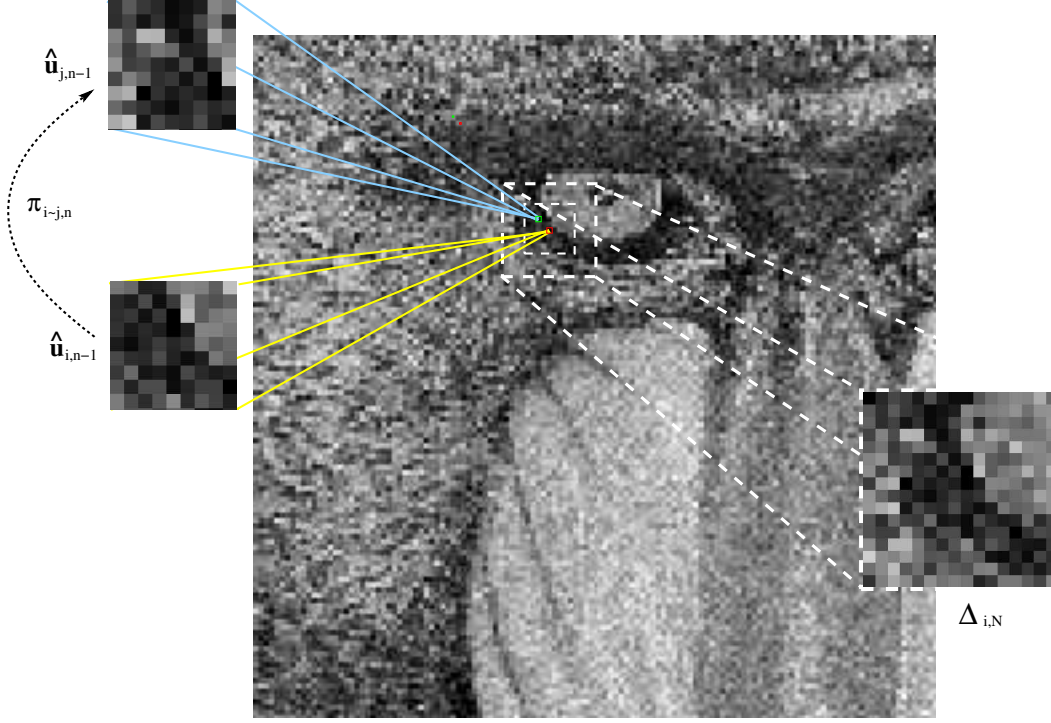
Figure 1. Description of the exemplar-based restoration approach at a pixel $\mathbf{x}_i$. The largest window $\Delta_{i,N}$ (17 × 17 pixels when $N = 4$) and the patch $\widehat{\mathbf{u}}_{j,n-1}$ at $\mathbf{x}_j \in \Delta_{i,n}$ to be compared to the reference patch $\widehat{\mathbf{u}}_{i,n-1}$, have been selected and zoomed. The square boxes in dashed lines indicate the intermediate windows $\Delta_{i,n}$ tested at the previous iterations $n \leq N$.

the estimate $\widehat{u}_{i,1}$ as the *weighted average* of data points lying in the neighborhood $\Delta_{i,1}$. We continue this way, increasing with n the window $\Delta_{i,n}$ while $n \leq N$ where N denotes the maximal number of iterations of the algorithm and the cardinality of $\mathcal{J}$.

For each $n \in \{1, \cdots, N\}$, the iterative estimator and its variance are defined as

$$\widehat{u}_{i,n} = \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i \sim j,n} \, Y_j, \quad \text{and} \quad v^2(\widehat{u}_{i,n}) = \widehat{\sigma}^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\pi_{i \sim j,n})^2 \tag{4}$$

where the weights $\pi_{i \sim j,n}$ are continuous and assumed deterministic variables that satisfy the usual conditions $0 \leq \pi_{i \sim j,n} \leq 1$ and $\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i \sim j,n} = 1$. Here, we suggest to compute the weights from pairs of previous restored patches $\widehat{\mathbf{u}}_{i,n-1}$ and $\widehat{\mathbf{u}}_{j,n-1}$ obtained at iteration $n - 1$ as

$$\pi_{i \sim j,n} = \frac{G_{\lambda_\alpha}(\text{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{j,n-1}))}{\sum_{\mathbf{x}_k \in \Delta_{i,n}} G_{\lambda_\alpha}(\text{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{k,n-1}))} \tag{5}$$

where $G_s(\cdot)$ denotes an exponential function, i.e. $G_s(z) = \exp(-z/2s)$. Due to the fast decay of the exponential kernel, large distances between estimated patches lead to nearly

zero weights, and λ_α acts a scale parameter. Besides, the use of weights enables to relax the structural assumption that the neighborhood is roughly modeled by a square window, and is an alternative strategy to a more natural geometric strategy which consists in estimating the anisotropy and direction of a rectangular or elliptic window (e.g. see [90]). In equation (5), the weights $\pi_{i\sim j,n}$ do not directly depend on input data, but are calculated from neighboring restored patches to improve robustness to noise. This contrasts with the traditional local M-estimators [19] and the *Non-Local means* filter [14] (see (1)).

To complete the description, the estimator (4) requires to define an appropriate distance to compare patches, and an objective criterion to determine the "best" window adapted for each pixel. These problems are addressed in the next sections.

4.1. Adaptive weights In order to compute $\widehat{u}_{i,n+1}$ at the next iteration, we need an objective distance to measure the similarity between two patches $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$. In [27, 22, 14, 15], several authors showed that the Euclidean distance is a reliable measure to compare image patches. To take into account the estimator variance, we have rather used the following normalized and symmetrized distance to compare two restored patches

$$\text{dist}(\widehat{\mathbf{u}}_{i,n}, \widehat{\mathbf{u}}_{j,n}) = (\widehat{\mathbf{u}}_{i,n} - \widehat{\mathbf{u}}_{j,n})^T \, \widehat{\mathbf{V}}_{ij,n}^{-1} \, (\widehat{\mathbf{u}}_{i,n} - \widehat{\mathbf{u}}_{j,n}) \qquad (6)$$

where $\widehat{\mathbf{V}}_{ij,n}^{-1} = \frac{1}{2} \text{diag}\left[v^{-2}(\widehat{u}_{i,n}^{(\ell)}) + v^{-2}(\widehat{u}_{j,n}^{(\ell)}) \right]$ is a $p \times p$ diagonal matrix where $v^2(\widehat{u}_{\cdot,n}^{(\ell)}), \ell = 1, \cdots, p$, is the local variance of $\widehat{u}_{\cdot,n}^{(\ell)}$ (see (4)), and the index ℓ is used to denote a spatial position in an image patch $\widehat{\mathbf{u}}_{\cdot,n} = (\widehat{u}_{\cdot,n}^{(1)}, \cdots, \widehat{u}_{\cdot,n}^{(\ell)}, \cdots, \widehat{u}_{\cdot,n}^{(p)})^T$. Accordingly, the hypothesis $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$ are similar, is accepted if the distance is small, i.e. $\text{dist}(\widehat{\mathbf{u}}_{i,n}, \widehat{\mathbf{u}}_{j,n}) \leq \lambda_\alpha$. Since $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$ are Gaussian vectors from (4), it follows that $\widehat{\mathbf{u}}_{i,n} - \widehat{\mathbf{u}}_{j,n}$ is Gaussian distributed: $\widehat{\mathbf{u}}_{i,n} - \widehat{\mathbf{u}}_{j,n} \sim \mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}_{ij,n})$. Hence, the parameter $\lambda_\alpha \in \mathbb{R}_+$ will be chosen as a quantile of a $\chi^2_{p,1-\alpha}$ distribution with p degrees of freedom. This parameter controls the probability of "type I error" defined as

$$\mathbb{P}\left(\text{dist}(\widehat{\mathbf{u}}_{i,n}, \widehat{\mathbf{u}}_{j,n}) \leq \lambda_\alpha\right) = 1 - \alpha, \qquad (7)$$

that is the probability of incorrectly rejecting the hypothesis $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$ are similar. All these tests ($|\Delta_{i,n}|$ tests) have to be performed at a very high significance level, our experience suggesting to use a $1 - \alpha = 0.99$-quantile. If $\text{dist}(\widehat{\mathbf{u}}_{i,n}, \widehat{\mathbf{u}}_{j,n})$ exceeds this critical threshold λ_α, then we have a significant difference between $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$ and we reject the hypothesis that $\widehat{\mathbf{u}}_{i,n}$ and $\widehat{\mathbf{u}}_{j,n}$ are coming from the same texture "class".

4.2. Adaptive windows In the iterative procedure, n coincides with the iteration and we will use $\widehat{n}(\mathbf{x}_i)$ to designate the index of the estimated window at $\mathbf{x}_i$ chosen among all the non-rejected windows $\Delta_{i,n}$, such as

$$\widehat{n}(\mathbf{x}_i) = \sup\left\{ n = 1, \cdots, N \; : \; |\widehat{u}_{i,n} - \widehat{u}_{i,n'}| \leq \varrho\, v(\widehat{u}_{i,n'}) \;\; \text{for all} \;\; n' < n \right\} \qquad (8)$$

where ϱ is a positive constant and $v(\widehat{u}_{i,n'})$ is the standard deviation of $\widehat{u}_{i,n'}$. This rule means that we select the largest window such that the estimators $\widehat{u}_{i,n}$ and $\widehat{u}_{i,n'}$ are not too different, in some sense, for all $n' < n$. Intuitively, incorporating more input data points in the computation of the weighted average as the window increases, tends to reduce noise in the resulting image. However, the window needs to be not too large since the remote data

---

Algorithm Exemplar-based image restoration algorithm

Let $\{p, \lambda_\alpha, N\}$ be the parameters.

Initialization: $\widehat{u}_{i,0} = Y_i$ and $v^2(\widehat{u}_{i,0}) = \widehat{\sigma}^2$ for each $\mathbf{x}_i \in \Omega$ where

$$\widehat{\sigma} = 1.4826 \, \mathrm{med}(|\, \mathbf{r} - \mathrm{med}|\mathbf{r}|\,|)$$

where $\mathbf{r} = \{r_1, r_2, \ldots, r_{|\Omega|}\}$ is the set of local residuals of the entire image defined as

$$r_i = (2Y_{i_1,i_2} - (Y_{i_1+1,i_2} + Y_{i_1,i_2+1}))/\sqrt{6}.$$

and Y_{i_1,i_2} is the observation Y_i at point $\mathbf{x}_i$ (see [33, 48]).

Repeat

- for each $\mathbf{x}_i \in \Omega$

  - compute

$$\pi_{i \sim j, n} = \frac{G_{\lambda_\alpha}(\mathrm{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{j,n-1}))}{\displaystyle\sum_{\mathbf{x}_k \in \Delta_{i,n-1}} G_{\lambda_\alpha}(\mathrm{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{k,n-1}))}$$

$$\widehat{u}_{i,n} = \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i \sim j, n} Y_j, \quad \text{and} \quad v^2(\widehat{u}_{i,n}) = \widehat{\sigma}^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\pi_{i \sim j, n})^2.$$

  - test the window using

$$\widehat{n}(\mathbf{x}_i) = \sup\{n = 1, \cdots, N \; : \; |\widehat{u}_{i,n} - \widehat{u}_{i,n'}| \le \varrho \, v(\widehat{u}_{i,n'}) \text{ for all } n' < n\}.$$

    If this rule is violated at iteration n, we do not accept $\widehat{u}_{i,n}$ and keep the estimate $\widehat{u}_{i,n-1}$ as the final estimate at $\mathbf{x}_i$, i.e. $\widehat{u}(\mathbf{x}_i) = \widehat{u}_{i,n-1}$ and $\widehat{n}(\mathbf{x}_i) = n - 1$. This estimate is unchanged at the next iterations and $\mathbf{x}_i$ is "frozen".

- increment n

while $n \le N$.

Figure 2. Exemplar-based image restoration algorithm.

points are likely less significant and can originate from other spatial contexts; this results in reducing the accuracy of the estimation, i.e. the local bias of the estimator increases. For each pixel, the detection of this transition enables to determine the critical size of the window. In the next section, we will show that, rejecting $\widehat{u}_{i,n}$ in favor of $\widehat{u}_{i,n'}$, $n' < n$, as the procedure prescribes, would result in a reduction of the local bias more substantial than the increasing of the local variance of the estimator. Out of all the windows which have not been thus rejected, the one corresponding to the smallest variance is used in the construction of the final estimator $\widehat{u}(\mathbf{x}_i) := \widehat{u}_{i,\widehat{n}(\mathbf{x}_i)}$.

Note that the rule (8) relies on the assumption that the variance decreases mono-
tonically as the window increases as explained in Section 5. Accordingly, an additional
test to validate the current window, should be introduced in the algorithm to ensure
$v(\widehat{u}_{i,n+1}) \leq v(\widehat{u}_{i,n})$. Nevertheless, the denoising results are not altered if this test is not
introduced in the iterative procedure, and may be considered as optional in the algorithm.

4.3. Algorithm We describe a practical algorithm (see Fig. 1 and Fig. 2), with a minimal
number of calibrated parameters, based on the previous construction of the adaptive win-
dow and the corresponding estimator. The key ingredient of the procedure is an increasing
sequence of nested square windows, centered at $\mathbf{x}_i$, of size $|\Delta_{i,n}| = (2^n + 1) \times (2^n + 1)$
pixels with $n = 1, \ldots, N$. At the initialization, we naturally choose $|\Delta_{i,0}| = 1$ and set
the fixed size of $\sqrt{p} \times \sqrt{p}$ patches and the parameter λ_α involved in the image patch
comparison. In addition, the estimation procedure relies on the preliminary estimation of
the noise variance $\widehat{\sigma}^2$ robustly estimated from input data (In Fig. 2, the constant $1/\sqrt{6}$
is used to ensure $\mathbb{E}[r_i^2] = \widehat{\sigma}^2$ in homogeneous regions). We manually set the number N
(typically $N = 4$) of iterations to bound the numerical complexity which is of the order
$p \times N \times |\Delta_{\cdot,N}| \times |\Omega|$ if an image contains $|\Omega|$ pixels. As expected, increasing N allows for
additional variance reduction in homogeneous regions.

In our experimental results reported in Section 7, we demonstrate that the use of vari-
able and overlapping windows contributes to the restoration performance with no block
effect in natural images.

5. Minimization problem and optimality conditions

Throughout this section, we shall see the rational behind the pointwise statistical rule (8)
for choosing the optimal estimation window, and motivate the iterative algorithm that
updates the estimator when the window increases at each iteration.

We show that the "best" possible estimator $\widehat{u}(\mathbf{x}_i) := \widehat{u}_{i,\widehat{n}(\mathbf{x}_i)}$, computed from the whole
path of values $\{\widehat{u}_{i,n}\}$, is solution of the following original minimization problem

$$
\begin{cases}
\widehat{u} &= \arg\min_u \left\{ J(u, Y) = \sum_{\mathbf{x}_i \in \Omega} \sum_{\mathbf{x}_j \in \Delta_{i,n}} \phi_s \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right) \right\}, \\
\\
\widehat{n}(\mathbf{x}_i) &= \arg\min_n \mathbb{E}\left[|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)|^2 \right],
\end{cases}
\tag{9}
$$

where n and $\widehat{n}(\mathbf{x}_i)$ are respectively the index of the current window $\Delta_{i,n}$ and the index of
the estimated window $\Delta_{i,\widehat{n}(\mathbf{x}_i)}$ at pixel $\mathbf{x}_i$. In the first equation in (9), $\phi_s : \mathbb{R}^+ \to \mathbb{R}$ is a
non-convex function of the form $\phi_s(z) = 1 - e^{-z/2s}$, and we define

$$
\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} = (\mathbf{u}_i - \widetilde{\mathbf{u}}_j)^T \widetilde{\mathbf{V}}_{ij}^{-1} (\mathbf{u}_i - \widetilde{\mathbf{u}}_j)
\tag{10}
$$

with $\mathbf{u}_i = (u_i^{(1)}, \cdots, u_i^{(c)}, \cdots, u_i^{(p)})^T$ and $\widetilde{\mathbf{u}}_j = (u_j^{(1)}, \cdots, Y_j, \cdots, u_j^{(p)})^T$ where the super-
script $^{(c)}$ is used to denote the central pixel in the reference patch $\mathbf{u}_i$, and $\widetilde{\mathbf{V}}_{ij}$ is a $p \times p$
diagonal covariance matrix (see Section 4.1). In the definition of $\widetilde{\mathbf{u}}_j$, Y_j is substituted to
$u_j^{(c)}$ and, accordingly, the minimization of $J(u, Y)$ does not lead to a constant image.

The functional $J(u, Y)$ is a non-local energy functional similar to those already studied in [63] since $\Delta_{i,n}$ can be large, but is enhanced here by introducing image patches in the definition. Intuitively, minimizing $J(u, Y)$ amounts to estimating an image for which neighboring patches are similar and, at the same time, the estimated value at the central position in the reference patch $\mathbf{u}_i$ must be as close as possible to the input data Y_j observed at the central positions in the neighboring patches $\{\mathbf{u}_j\}$, included in the reference patch $\mathbf{u}_i$. The non-local and complex interactions in spatially varying neighborhoods are thus taken into account in this image modeling.

In the second equation in (9), the index $\widehat{n}(\mathbf{x}_i)$ of the estimated window at position $\mathbf{x}_i$ minimizes the local L_2 risk involving the true and unknown function $u_{\text{true}}(\mathbf{x}_i)$.

5.1. Non-local energy minimization. According to the variation calculus method (see also [52, 39]), we have

$$J(u + \delta u, Y) - J(u, Y) = \sum_{\mathbf{x}_i \in \Omega} \sum_{\mathbf{x}_j \in \Delta_{i,n}} \left[\phi_s \left(\|(\mathbf{u}_i + \delta\mathbf{u}_i) - (\widetilde{\mathbf{u}}_j + \delta\widetilde{\mathbf{u}}_j)\|^2_{\widetilde{\mathbf{V}}_{ij}} \right) - \phi_s \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right) \right]$$

with the abbreviations $\delta\mathbf{u}_i = (\delta u_i^{(1)}, \cdots, \delta u_i^{(c)}, \cdots, \delta u_i^{(p)})^T$ and $\delta\widetilde{\mathbf{u}}_j = (\delta u_j^{(1)}, \cdots, 0, \cdots, \delta u_j^{(p)})^T$. A straightforward calculation shows that the first variation of $J(u, Y)$ is given by (high-order terms are neglected)

$$J(u + \delta u, Y) - J(u, Y) \approx 2 \sum_{\mathbf{x}_i \in \Omega} \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\delta\mathbf{u}_i - \delta\widetilde{\mathbf{u}}_j)^T \widetilde{\mathbf{V}}_{ij}^{-1} (\mathbf{u}_i - \widetilde{\mathbf{u}}_j) \phi_s' \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right).$$

Since we are only interested in the local variation at point $\mathbf{x}_i$, we set $\delta u_j = 0, \forall j \neq i$. Now, let us assume that if $\mathbf{x}_j \in \Delta_{i,n}$ then $\mathbf{x}_i \in \Delta_{j,n}$ and vice-versa (i.e. $\mathbf{x}_j$ and $\mathbf{x}_i$ are mutually neighbors), even this does not always hold true. It follows that

$$\frac{J(u_1, \cdots, u_i + \delta u_i, \cdots, u_{|\Omega|}, Y) - J(u_1, \cdots, u_i, \cdots, u_{|\Omega|}, Y)}{\delta u_i} \approx$$
$$\frac{2}{\left[\widetilde{\mathbf{V}}_{ij}\right]_i} \sum_{\mathbf{x}_j \in \Delta_{i,n}} (u_i - Y_j) \, \phi_s' \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right).$$

where $\left[\widetilde{\mathbf{V}}_{ij}\right]_i$ denotes the element of the $p \times p$ diagonal matrix $\widetilde{\mathbf{V}}_{ij}$ at point $\mathbf{x}_i$ (see Section 4.1). If u is a stationary point of $J(u, Y)$, then we have

$$\frac{J(u_1, \cdots, u_i + \delta u_i, \cdots, u_{|\Omega|}, Y) - J(u_1, \cdots, u_i, \cdots, u_{|\Omega|}, Y)}{\delta u_i} = 0, \quad \forall i \in \{1, \cdots, |\Omega|\}.$$

This can be transformed into fixed-point form as

$$u_i = \frac{\sum_{\mathbf{x}_j \in \Delta_{i,n}} \phi_s' \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right) Y_j}{\sum_{\mathbf{x}_j \in \Delta_{i,n}} \phi_s' \left(\|\mathbf{u}_i - \widetilde{\mathbf{u}}_j\|^2_{\widetilde{\mathbf{V}}_{ij}} \right)}. \tag{11}$$

If we impose $\phi_s(z) = 1 - e^{-z/2s}$ and perform a one-step of a fixed-point iteration, we get

$$u_{i,\text{new}} = \frac{\displaystyle\sum_{\mathbf{x}_j \in \Delta_{i,n}} G_s\left(\|\mathbf{u}_{i,\text{old}} - \widetilde{\mathbf{u}}_{j,\text{old}}\|^2_{\widetilde{\mathbf{V}}_{ij,\text{old}}}\right) Y_j}{\displaystyle\sum_{\mathbf{x}_j \in \Delta_{i,n}} G_s\left(\|\mathbf{u}_{i,\text{old}} - \widetilde{\mathbf{u}}_{j,\text{old}}\|^2_{\widetilde{\mathbf{V}}_{ij,\text{old}}}\right)}. \tag{12}$$

since $\phi_s'(z) = G_s(z)/2s$. We use this equation to build up an iterative method to minimize $J(u, Y)$. As usual, we could iterate the updating steps to minimize $J(u, Y)$ given $\Delta_{i,n}$, and then increase the window size. To speed-up the iterative procedure, a single fixed-point iteration is performed, which does not, surprisingly, alter the results in our experiments. Accordingly, the index n of the window size and the iteration will be the same and we substitute n to the subscript "new" (and $n-1$ to "old") in the previous equation. Moreover, if we use the following approximation $\|\mathbf{u}_{i,\text{old}} - \widetilde{\mathbf{u}}_{j,\text{old}}\|^2_{\widetilde{\mathbf{V}}_{ij,\text{old}}} \approx \text{dist}(\mathbf{u}_{i,\text{old}}, \mathbf{u}_{j,\text{old}})$ which holds for p large (e.g $p \geq 25$) and set the scale parameter s to λ_α, the iterative estimator then reads as

$$\widehat{u}_{i,n} = \frac{\displaystyle\sum_{\mathbf{x}_j \in \Delta_{i,n}} G_{\lambda_\alpha}\left(\text{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{j,n-1})\right) Y_j}{\displaystyle\sum_{\mathbf{x}_j \in \Delta_{i,n}} G_{\lambda_\alpha}\left(\text{dist}(\widehat{\mathbf{u}}_{i,n-1}, \widehat{\mathbf{u}}_{j,n-1})\right)}. \tag{13}$$

and corresponds to the solution already given in Section 4.

In [49], we proved that this estimator fulfills the *extremum principle* and the *average gray level invariance*.

5.2. An oracle window for smoothing. In this section, we demonstrate that minimizing $\mathbb{E}\left[|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)|^2\right]$ in (9) is equivalent to satisfying the pointwise statistical rule (8).

First, image smoothing from noisy data can be regarded as a 2D statistical kernel regression problem and the bandwidths are locally adapted to local image features [20, 75]. In [75], the authors determine local bandwidths using Parzen windows to mimic local density for image segmentation. This is a variant of the *plug-in* idea usually used in the statistics literature, which is fast and easy to compute. However the plug-in approach is problematic since it is known to be highly sensitive to noise in images and to the choice of a global initial bandwidth. In [82], the authors propose to control bootstrap processes using the covariance of the transformation estimate for retinal image registration. In [37], the local neighborhoods are adapted to the local smoothness of the image; a limitation of this approach is the use of a global bandwidth.

In this section, we adopt the *adaptive estimation* framework [55, 56, 57] to address the problem of automatic selection of the window adapted for each pixel. It is well understood that the local smoothness varies significantly from point to point in the image and usual *global risks* cannot wholly reflect the performance of estimators at a point. Then, a classical way to measure the performance of the estimator $\widehat{u}_{i,n}$ to its target value $u_{\text{true}}(\mathbf{x}_i)$ is to choose the *local L_2 risk*, which is explicitly decomposed into the sum of the squared bias $b^2(\widehat{u}_{i,n})$ and the variance $v^2(\widehat{u}_{i,n})$:

$$\mathbb{E}[|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)|^2] = b^2(\widehat{u}_{i,n}) + v^2(\widehat{u}_{i,n}). \tag{14}$$

Our goal is to minimize this local L_2 risk with respect to the size of the window $\Delta_{i,n}$, and for each pixel in the image. The closed-form solution that minimizes the L_2 risk represents the "best" possible estimator (also called "oracle" estimator) which unfortunately cannot be used in practice. Actually, this optimal solution explicitly depends on the smoothness of the "true" function $u_{\text{true}}(\mathbf{x}_i)$ which is unknown, and so, of less practical interest (see [74, 43, 46, 78]). A natural way to bring some further understanding of the situation is then to individually analyze the behavior of the bias and variance terms when $\Delta_{i,n}$ increases or decreases with n as follows:

- The bias term $b(\widehat{u}_{i,n}) = \mathbb{E}\left[\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)\right]$ is nonrandom and characterizes the accuracy of approximation of the function u_{true} at $\mathbf{x}_i$ by the smoothing window. As it explicitly depends on the unknown function $u_{\text{true}}(\mathbf{x}_i)$, it is usually not very useful by itself. Nevertheless, approximations to the bias can be derived if we assume u_{true} is Lipschitz continuous, that is $|u_{\text{true}}(\mathbf{x}_j) - u_{\text{true}}(\mathbf{x}_i)| \leq C_1|\mathbf{x}_j - \mathbf{x}_i|$, $0 < C_1 < \infty$. If we use the geometric inequality $|\mathbf{x}_j - \mathbf{x}_i| \leq \frac{\sqrt{2}}{2}|\Delta_{i,n}|^{1/2}$ for 2D images, it follows that

$$
\begin{aligned}
|b(\widehat{u}_{i,n})| \;=\; \left| \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n} \left[\mathbb{E}[Y_j] - u_{\text{true}}(\mathbf{x}_i)\right] \right| \;&\leq\; \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n} \left|u_{\text{true}}(\mathbf{x}_j) - u_{\text{true}}(\mathbf{x}_i)\right| \\
&\leq\; C_1 \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n} \left|\mathbf{x}_j - \mathbf{x}_i\right| \\
&\leq\; \frac{C_1|\Delta_{i,n}|^{1/2}}{\sqrt{2}}
\end{aligned}
$$

  and so $b^2(\widehat{u}_{i,n})$ is of the order $O(|\Delta_{i,n}|)$. Thus the squared bias is small when $|\Delta_{i,n}|$ is small and typically increases when $\Delta_{i,n}$ increases. In other words, incorporating data points from irrelevant image contexts tends to decrease the accuracy of the estimator.

- The behavior of the variance term is just opposite. The errors are independent and the stochastic term $v^2(\widehat{u}_{i,n})$ can be computed on the basis of observations. Since $0 \leq \pi_{i\sim j,n} \leq 1$ and $\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n} = 1$ and are assumed to be deterministic variables, it follows that

$$
\frac{\sigma^2}{|\Delta_{i,n}|} \;\leq\; v^2(\widehat{u}_{i,n}) \;\leq\; \sigma^2.
$$

  In addition, we can reasonably assume that there exits a constant $0 \leq \gamma^2(\mathbf{x}_i) \leq 1$ such that $v^2(\widehat{u}_{i,n}) \approx \sigma^2|\Delta_{i,n}|^{-\gamma^2(\mathbf{x}_i)}$. Accordingly, as $\Delta_{i,n}$ increases, more data is used to construct the estimate $\widehat{u}_{i,n}$, and so $v^2(\widehat{u}_{i,n})$ decreases.

In short, the bias and standard deviation are then assumed to be monotonous functions with opposite behavior. To ensure this monotonicity, we decide to reject the current window $\Delta_{i,n}$ at point $\mathbf{x}_i$ if $v^2(\widehat{u}_{i,n}) > v^2(\widehat{u}_{i,n-1})$ at iteration n, and to continue the estimation process with a larger window. Image gradient points are especially affected by this strategy.

In order to approximately minimize the local L_2 risk of the estimator with respect to $|\Delta_{i,n}|$, a natural idea is to minimize an upper bound of the form

$$
\mathbb{E}[|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)|^2] \;\leq\; \frac{C_1^2}{2}|\Delta_{i,n}| + \frac{\sigma^2}{|\Delta_{i,n}|^{\gamma^2(\mathbf{x}_i)}}.
$$

This equation summarizes the well-known bias-variance trade-off and the size of the expected window can be easily calculated as

$$|\Delta_{\text{oracle}}(\mathbf{x}_i)| = \left[\frac{2\sigma^2\gamma^2(\mathbf{x}_i)}{C_1^2}\right]^{\frac{1}{\gamma^2(\mathbf{x}_i)+1}}.$$

This solution cannot be used in practice since C_1 and $\gamma^2(\mathbf{x}_i)$ are unknown, but it can be shown, for the solution $\Delta_{\text{oracle}}(\mathbf{x}_i)$, that (see [46, 78])

$$\frac{|b(u_{\text{oracle}}(\mathbf{x}_i))|}{v(u_{\text{oracle}}(\mathbf{x}_i))} \leq \gamma(\mathbf{x}_i)$$

where $u_{\text{oracle}}(\mathbf{x}_i)$ represents the estimated value if we would know $\Delta_{\text{oracle}}(\mathbf{x}_i)$. Hence, a good choice of the window is then the largest window $\Delta_{i,n}$ such that $|b(\widehat{u}_{i,n})|$ is not larger than $\gamma(\mathbf{x}_i)v(\widehat{u}_{i,n})$, for some real value $\gamma(\mathbf{x}_i) \in \mathbb{R}_+$, i.e.

$$n_{\text{oracle}}(\mathbf{x}_i) = \sup\{n = 1, \cdots N \; : \; |b(\widehat{u}_{i,n})| \leq \gamma(\mathbf{x}_i)v(\widehat{u}_{i,n})\}.$$

Since $n_{\text{oracle}}(\mathbf{x}_i)$ is based on the full knowledge of the bias (and then on the unknown function u_{true}), $u_{\text{oracle}}(\mathbf{x}_i)$ is called the *"oracle"* estimate[2] since it represents a value we cannot expect to attain. In practice, the bias is not observable and the bias-variance trade-off cannot be obtained by sweeping the measured bias $b(\widehat{u}_{i,n})$ and variance $v(\widehat{u}_{i,n})$ indexed by the smoothing window $\Delta_{i,n}$ at point $\mathbf{x}_i$. Therefore, we need more precise characterizations to derive a selection procedure. We propose then the following standard decomposition of the estimator $\widehat{u}_{i,n}$ [57]

$$\widehat{u}_{i,n} = u_{\text{true}}(\mathbf{x}_i) + b(\widehat{u}_{i,n}) + w(\mathbf{x}_i) \tag{15}$$

where $w(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbb{E}[(w(\mathbf{x}_i))^2])$, $\mathbb{E}[\widehat{u}_{i,n}] = u_{\text{true}}(\mathbf{x}_i) + b(\widehat{u}_{i,n})$ and $\mathbb{E}[(w(\mathbf{x}_i))^2] = \mathbb{E}[|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i) - b(\widehat{u}_{i,n})|^2] := v^2(\widehat{u}_{i,n})$. Therefore, the following inequality

$$|\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)| \leq |b(\widehat{u}_{i,n})| + \varkappa\, v(\widehat{u}_{i,n}) \tag{16}$$

holds with a high probability and $0 < \varkappa < \infty$. Finally, since $|b(\widehat{u}_{i,n})| \leq \gamma(\mathbf{x}_i)v(\widehat{u}_{i,n})$, we modify correspondingly the definition of the *"oracle"* window as

$$\sup\{n = 1, \cdots N \; : \; |\widehat{u}_{i,n} - u_{\text{true}}(\mathbf{x}_i)| \leq (\gamma(\mathbf{x}_i) + \varkappa)\, v(\widehat{u}_{i,n})\}. \tag{17}$$

The crucial point is that this inequality depends no longer on $b(\widehat{u}_{i,n})$, but is yet related to the unknown function $u_{\text{true}}(\mathbf{x}_i)$. Nevertheless, in the next section, we shall see that a *data-driven* window selector based on this definition can actually be derived.

5.3. A data-driven local window selector. In the pointwise estimation approach, we strongly suppose $v^2(\widehat{u}_{i,n})$ decreases as n increases and the ordering relation $\widehat{u}_{i,n'} \preccurlyeq \widehat{u}_{i,n}$ that implies $v^2(\widehat{u}_{i,n}) \leq v^2(\widehat{u}_{i,n'})$, can be introduced. If this assumption is not fulfilled for the original set $\mathcal{J}$, i.e. there is $\Delta_{i,n'} \subset \Delta_{i,n}$ with the property $v(\widehat{u}_{i,n'}) > v(\widehat{u}_{i,n})$, then we simply exclude the window $\Delta_{i,n'}$ from $\mathcal{J}$ at point $\mathbf{x}_i$. The collection of estimators $\{\widehat{u}_{i,1}, \ldots, \widehat{u}(\mathbf{x}_i)\}$ is then naturally ordered in the direction of increasing $|\Delta_{i,n}|$ where $\widehat{u}(\mathbf{x}_i)$

---

[2] $\Delta_{\text{oracle}}(\mathbf{x}_i)$ denotes the *"oracle"* window.

can be thought as the best possible estimator. Accordingly, we propose a selection procedure based on pairwise comparisons of an essentially one-dimensional family of competing estimators $\widehat{u}_{i,n}$ as described below.

Actually, the random variables $(\widehat{u}_{i,n} - \widehat{u}_{i,n'})$ are Gaussian random variables with expectations equal to the bias differences $b(\widehat{u}_{i,n}) - b(\widehat{u}_{i,n'})$ and variances $v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'}) \leq v^2(\widehat{u}_{i,n'})$ (see the proof in Appendix A.1.). From (16) and (17), it follows that

$$
\begin{aligned}
|\widehat{u}_{i,n'} - \widehat{u}_{i,n}| &\leq |b(\widehat{u}_{i,n'}) - b(\widehat{u}_{i,n}) + \varkappa\, v(\widehat{u}_{i,n} - \widehat{u}_{i,n'})| \qquad\qquad (18) \\
&\leq |b(\widehat{u}_{i,n'})| + |b(\widehat{u}_{i,n})| + \varkappa v(\widehat{u}_{i,n'}) \\
&\leq \gamma(\mathbf{x}_i) v(\widehat{u}_{i,n'}) + \gamma(\mathbf{x}_i) v(\widehat{u}_{i,n}) + \varkappa \widehat{v}_{i,n'} \\
&\leq (2\gamma(\mathbf{x}_i) + \varkappa) v(\widehat{u}_{i,n'}) \\
&\leq (2\sup_{\mathbf{x}_i \in \Omega} \gamma(\mathbf{x}_i) + \varkappa) v(\widehat{u}_{i,n'}).
\end{aligned}
$$

Now if a threshold $\varrho := (2\sup_{\mathbf{x}_i \in \Omega} \gamma(\mathbf{x}_i) + \varkappa)$ is properly chosen, none of the variables $|\widehat{u}_{i,n'} - \widehat{u}_{i,n}|$ will exceed the value $\varrho v(\widehat{u}_{i,n'}), n' < n$, with a high probability. Among all the candidates satisfying $|\widehat{u}_{i,n'} - \widehat{u}_{i,n}| \leq \varrho v(\widehat{u}_{i,n'})$, one naturally choose the one with the smallest variance. The exact choice of the threshold ϱ then becomes a balancing act based on large deviation calculations: ϱ should be large enough to guarantee a sufficiently high probability and, at the same time, small enough to provide a good control of $\widehat{u}_{i,n} - \widehat{u}(\mathbf{x}_i)$. Following the above discussion, a window selector is then based on the rule [56, 59, 45, 57, 41]:

$$
\widehat{n}(\mathbf{x}_i) = \sup\{n = 1, \cdots, N \ : \ |\widehat{u}_{i,n} - \widehat{u}_{i,n'}| \leq \varrho\, v(\widehat{u}_{i,n'}) \text{ for all } n' < n\}. \qquad (19)
$$

From this definition, it seems that the pixels are treated in an independent fashion, but we recall that spatial correlations are indirectly introduced in (4). Now, setting the threshold ϱ remains an open issue in practice as already mentioned in [46] and in a recent reliability study in signal processing applications [78]. In the next section, we shall see how the threshold ϱ can be estimated from local image statistics.

5.4. Estimation of ϱ Clearly, the choice of the parameter ϱ plays an important role in the adaptation and must be carefully chosen. In order to calibrate this threshold, we need to evaluate the probability of the event $\{\widehat{n}(\mathbf{x}_i) = n\}$ at $\mathbf{x}_i$ and prove the proposition:

PROPOSITION 1. *The event $\{\widehat{n}(\mathbf{x}_i) = n\}$ occurs at $\mathbf{x}_i$ with a probability*

$$
\mathbb{P}(\widehat{n}(\mathbf{x}_i) = n) \leq \sum_{n'=1}^{n} 2\exp\left(-\frac{\varrho^2}{2}\right).
$$

Proof: see Appendix A.2.

From PROPOSITION 1, it follows

$$
1 - \mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N) = \sum_{n=1}^{N-1} \mathbb{P}(\widehat{n}(\mathbf{x}_i) = n) \leq \sum_{n=1}^{N-1}\sum_{n'=1}^{n} 2\exp\left(-\frac{\varrho^2}{2}\right) \leq N(N-1)\exp\left(-\frac{\varrho^2}{2}\right).
$$

Hence, if we fix the probability $\mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N)$, an upper bound can be calculated as

$$\varrho \leq \sqrt{2 \log \frac{N(N-1)}{1 - \mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N)}}. \tag{20}$$

Finally, to estimate $\mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N)$, we point out that most images are piecewise smooth and then contain a small number of pixels with significant discontinuities. Therefore, the probability $\mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N)$ is high at $\mathbf{x}_i \in \Omega$. The threshold ϱ can be then adapted to image contents using the following approximation

$$\mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N) \approx \mathbb{P}(|r_i| \leq \widehat{\sigma}) = \frac{\#\{\mathbf{x}_i : |r_i| \leq \widehat{\sigma}\}}{|\Omega|}, \tag{21}$$

i.e. by using the empirical distribution of pseudo-residuals $\{r_i\}$ and $\widehat{\sigma}^2$ defined in Fig. 2. In the ideal situation, the estimation windows should very large for pixels belonging to homogeneous (or flat) regions in the image (i.e. $\widehat{n}(\mathbf{x}_i) \geq N$) and small for pixels corresponding to edges. Accordingly, the proportion of pixels with small pseudo-residual magnitudes compared to the noise level gives an approximation of the probability $\mathbb{P}(\widehat{n}(\mathbf{x}_i) \geq N)$. This simple measurement encodes the intuition that these pixels are likely to belong to homogeneous regions. In what follows, we adopt the upper bound (20) (and the approximation (21)) to derive a data-driven estimation of ϱ in our experiments (see Section 7).

6. Theoretical accuracy of the pointwise adaptive estimation

The window $\Delta_{\text{oracle}}(\mathbf{x}_i)$ exactly balances the bias and variance terms, i.e. $|b(u_{\text{oracle}}(\mathbf{x}_i))| = \gamma(\mathbf{x}_i) v(u_{\text{oracle}}(\mathbf{x}_i))$ and the corresponding *"oracle" risk* is then of the form

$$\mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2] = (1 + \gamma^2(\mathbf{x}_i)) \, v^2(u_{\text{oracle}}(\mathbf{x}_i)). \tag{22}$$

For adaptive estimation at a point $\mathbf{x}_i$, one must pay a price for adaptation. The extra factor is a function of $\varkappa$ and $\gamma(\mathbf{x}_i)$, and under fair assumptions, the following proposition can be proved:

PROPOSITION 2. *If and only if $n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)$, there exists an adaptive estimate $\widehat{u}(\mathbf{x}_i)$ with the inaccuracy of order $\left[\frac{(2\gamma(\mathbf{x}_i)+\varkappa)}{\sqrt{1+\gamma^2(\mathbf{x}_i)}} + 1\right]^2$ times the "oracle" risk, i.e.*

$$\mathbb{E}[|\widehat{u}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]\mathbf{1}(n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)) \leq \left[\frac{(2\gamma(\mathbf{x}_i) + \varkappa)}{\sqrt{1 + \gamma^2(\mathbf{x}_i)}} + 1\right]^2 \mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2].$$

Proof: see Appendix A.3.

It is clear that the *"oracle" risk* is unattainable in general but, iif $n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)$, our estimator has a risk proportional to the desired risk. To complete this analysis, it remains to evaluate $\mathbb{P}(\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i))$ and to prove that we nearly never under-estimate the window, i.e.

PROPOSITION 3. *The event $\{\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)\}$ occurs at $\mathbf{x}_i$ with a low probability*

$$\mathbb{P}(\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)) \ \leq \ n_{\text{oracle}}(\mathbf{x}_i)(n_{\text{oracle}}(\mathbf{x}_i) - 1) \exp\left(-\frac{\varrho^2}{2}\right).$$

Proof: see Appendix A.4.

The probability of the event $\{\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)\}$ is small, provided that the value $n_{\text{oracle}}(\mathbf{x}_i)$ $(n_{\text{oracle}}(\mathbf{x}_i) - 1) \exp(-\varrho^2/2)$ is sufficiently small. Accordingly, if $2.5 \leq \varrho \leq 3$ (as shown in our experiments) and if we choose $N = 4$, the probability of the event $\{\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)\}$ occurs with a low probability. i.e. $\mathbb{P}(\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)) < 0.133$.

7. Experimental results

The proposed methodology is used for image denoising in various contexts as presented in this section. Our results were measured by the peak signal-to-noise ratio (PSNR) in decibels (db) defined as

$$\text{PSNR} = 10 \log_{10} \frac{255^2}{\text{MSE}}, \quad \text{MSE} = \frac{\sum_{\mathbf{x}_i \in \Omega}(u_o(\mathbf{x}_i) - \widehat{u}(\mathbf{x}_i))^2}{|\Omega|}$$

where u_0 is the noise-free original image.

In our experiments, the noise variance $\widehat{\sigma}^2$ and the threshold ϱ are automatically estimated from image data (see Fig. 2 and Section 5.4). We have used 9×9 image patches ($p = 81$) and set $N = 4$. In all the experiments, we shall see $\varrho \leq 3$, which implies the adaptive estimate $\widehat{u}(\mathbf{x}_i)$ is with the inaccuracy of order 9 times the *"oracle" risk* if $\gamma(\mathbf{x}_i) \leq 1$ (see PROPOSITION 2.). The choice of the critical values λ_α is also important. Large values improve stability of the model under homogeneity, while too small critical values lead to a large "false alarm" probability. In our experiments, this "false alarm" probability should not exceed the given level α set to 0.01 yielding to $\lambda_{0.01} = \chi^2_{81,0.99} = 113.5$. The processing of a 256×256 image required typically 67s ($p = 9^2$) on a PC (2.6 Ghz, Pentium IV) and less than 1s with a fast (block-based and spatial sub-sampling) implementation of the algorithm (C-C++ language). In both cases, the results are visually similar and the PSNR values are very close, even higher with a block-based implementation but with no spatial sub-sampling. In table I, we give the performance of the algorithm when applied to a commonly-used set of images available at `http://decsai. ugr.es/∼javier/denoise/test_images/` and described in [70]. In this experiment, the images are corrupted by additive white-Gaussian noise (WGN) (PSNR = 22.13 db, $\sigma = 20$). We tested the block-based implementation in two cases: The top row in Table I presents the results for $N = 3$ and no spatial sub-sampling ; The bottom row in Table I presents the results for $N = 5$ and a sub-sampling by a factor of 5. The results of the method using the algorithm (Fig. 2) are presented in Tables II-IV ($N = 4$). In any cases, our method outperforms most of competitive image denoising methods as shown in the next Section.

7.1. Denoising of artificially noisy images and comparison to the state-of-the-art. The potential of the estimation method is mainly illustrated with the 512×512 *Lena* image

Table I. Computation times and PSNR values obtained with an optimized block-based implementation of the algorithm applied to images corrupted by additive white-Gaussian noise (PSNR = 22.13 db, $\sigma = 20$). top row: $N = 3$ and sub-sampling by a factor of 5 ; bottom row: $N = 5$ and no sub-sampling.

Lena 512×512	Barbara 512×512	Boat 512×512	House 256×256	Pepper 256×256
1.99s / 32.42db	1.90s / 30.17db	2.02s / 30.18db	0.53s / 32.39db	0.48s / 30.70db
481.0s /32.82db	428.6s /30.86db	444.7s /30.31db	119.8s / 33.38db	100.9s / 30.79db

corrupted by an additive white-Gaussian noise (WGN) (Fig. 3a, PSNR = 22.13 db, $\sigma = 20$). In this experiment, we found $\mathbb{P}(|r_i| \leq \widehat{\sigma}) = 0.829$ from image data and computed $\varrho = 2.91$ from (20). In Fig. 3b, the noise is reduced in a natural manner and significant geometric features, fine textures, and original contrasts are visually well recovered with no undesirable artifacts (PSNR = 32.64 db). The noise component is shown in Fig. 3c (magnified $\times 2$) and has been estimated by calculating the difference between the noisy image (Fig. 3a) and the recovered image (Fig. 3b). The estimated noise component contains few geometric structures and is similar to a simulated white Gaussian noise. To better appreciate the accuracy of the restoration process, the variance of the pointwise estimator is shown in Fig. 3d where dark values correspond to high-confidence estimates. As expected, pixels with a low level of confidence are located in the neighborhood of image discontinuities. Figure 3e shows the probability of a patch $\widehat{\mathbf{u}}(\mathbf{x}_i)$ occurring in $\Delta_{i,\widehat{n}(\mathbf{x}_i)}$, i.e

$$\mathbb{P}\left(\widehat{\mathbf{u}}(\mathbf{x}_i) \text{ occurring in } \Delta_{i,\widehat{n}(\mathbf{x}_i)}\right) := \frac{\#\{\mathbf{x}_j \in \Delta_{i,\widehat{n}(\mathbf{x}_i)} : \text{dist}(\widehat{\mathbf{u}}(\mathbf{x}_i), \widehat{\mathbf{u}}(\mathbf{x}_j)) \leq \lambda_\alpha\}}{|\Delta_{i,\widehat{n}(\mathbf{x}_i)}|}.$$

Dark values correspond low probabilities of occurrence and, it is confirmed that repetitive patterns in the neighborhood of image discontinuities are mainly located along image level lines. Figure 3f shows the locations of the most "exceptional" patches by thresholding a continuous form of the probability map $\mathbb{P}(\widehat{\mathbf{u}}(\mathbf{x}_i)$ occurring in$\Delta_{i,\widehat{n}(\mathbf{x}_i)})$. The "rare" elements essentially correspond here to local distinctive features for which the intensity abruptly changes in the image, to various curvature maxima or not repeated patterns in a local neighborhood. Besides, we have compared the performance of our method to several competitive methods: Total Variation (TV) minimizing process [72], bilateral filtering [83], anisotropic diffusion (AD) using a diffusivity function of the type $(1 + |\nabla u|^2/g^2)^{-1}$ [67] and Wiener filtering (WF) (Matlab function `wiener2`). Figures 4a-d shows the results of the four tested methods. We stopped anisotropic diffusion after 150 iterations in order to avoid a over-smoothed image but a decorrelation criterion could be used to stop the diffusion process [62]. The TV minimizing method [72] completely eliminates small textures but also blurs edges when the Lagrange multiplier is set to 0.01. If we set the balance Lagrange multiplier to 0.05, the image is denoised but smooth parts are not completely recovered. Accordingly, the global control parameters of these algorithms were tuned (we have to try several values) to both eliminate noise and simultaneously to get the best PSNR value, and to give a good visual impression (Fig. 4). Additionally, this noisy image has been restored using pointwise adaptive estimation methods [48, 69] which are not patch-based.

Table II. Performance of denoising algorithms when applied to test noisy (WGN) images.
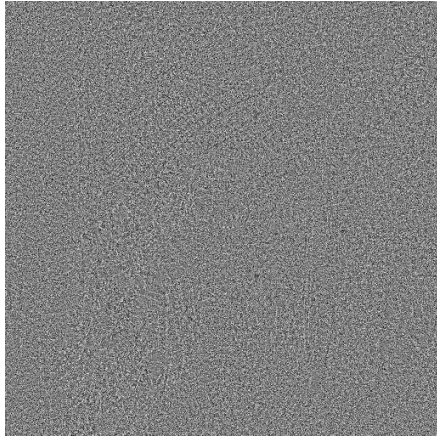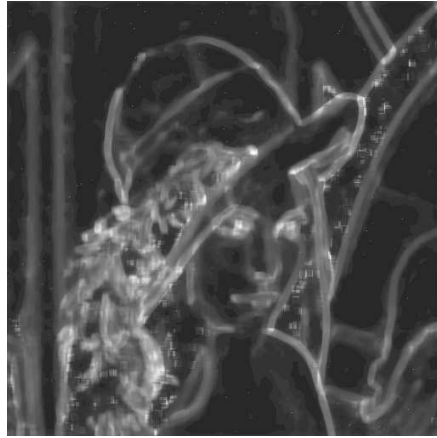
Image σ/PSNR	*Lena* 20 / 22.13	*Barbara* 20 / 22.18	*Boat* 20 / 22.17	*House* 20 / 22.11	*Peppers* 20 / 22.19
Our method (9×9 patch)	**32.64**	**30.37**	**30.12**	**32.90**	**30.59**
Buades *et al.* [14]	31.09	29.38	28.60	31.54	29.05
Ghazel *et al.* [36]	28.50	25.64	26.34	-	-
Kervrann [48]	30.54	26.50	28.01	30.70	28.23
Pizurica *et al.* [68]	32.20	29.53	29.93	-	30.30
Polzehl *et al.* [69]	29.74	26.05	27.74	30.31	28.40
Portilla *et al.* [70]	32.66	30.32	30.38	32.39	30.31
Roth *et al.* [71]	31.92	28.32	29.85	32.17	30.58
Rudin *et al.* [72]	30.48	27.07	29.02	31.03	28.51
Starck *et al.* [81]	31.95	-	-	-	-
Tomasi *et al.* [83]	30.26	27.02	28.41	30.01	28.88
Wiener filering	28.51	26.99	27.97	28.74	28.10

Figures 4e-f provides a visual comparison of image denoising with these two algorithms: the AWS algorithm [69] tends to oversmooth the image and to generate some artificial planar segments in homogeneous regions (Fig. 4f), whereas a variant of this approach [48] yields a similar result (Fig. 4e) to the image regularized with the TV method [72] (see Fig. 4a). In Fig. 4g-l, the corresponding recovered noise components are shown and most of them contain undesirable geometric structures. Moreover, our approach is also compared to the *Non-Local means* algorithm [14, 15] using 7×7 image patches and a fixed search window of 21×21 pixels: the visual impression and the numerical results are improved using our algorithm (see Figs. 5-6). Our approach is also compared to another and recent patch-based approach applied to image denoising [71], that exploits ideas from sparse image coding and training images for learning Markov random field image priors. The PSNR values are reported in table II for the test images; In most cases, our *unsupervised* and simple method produces the best PSNR values. Finally, we reported the best PSNR results obtained using these methods in table II. Both visually and in terms of PSNR, our method outperforms any of the tested methods.

Moreover, we have also examined some complementary aspects of our approach using the artificially corrupted *Barbara* image (WGN, $\sigma = 20$). In this experiment, we found $\mathbb{P}(|r_i| \leq \widehat{\sigma}) = 0.806$ using (21) and derived $\varrho = 2.871$ from (20) and the results are shown in Fig. 7 (zooming view). Finally, we varied the patch size and Fig. 8 shows that taking too small image patches can generate some visually undesirable flat zones during the restoration process. Note, that taking one point every two pixels (in both two directions) in a $\sqrt{p} \times \sqrt{p}$ patch ($k = p/4$) can be applied to produce natural regularized images and reduce the time computing. Table **??** reports the PSNR values obtained by varying the patch size and the sub-sampling (factor 2) for different test images. Note the PSNR values

(a) noisy 512×512 (WGN) image ($\sigma = 20$)

(b) denoised image

(c) noise component ($\times 2$)

(d) variance of the estimator

(e) $\mathbb{P}\left(\widehat{\mathbf{u}}(\mathbf{x}_i) \text{ occurring in } \Delta_{i,\widehat{n}(\mathbf{x}_i)}\right)$

(f) the most 271 "rare" patches

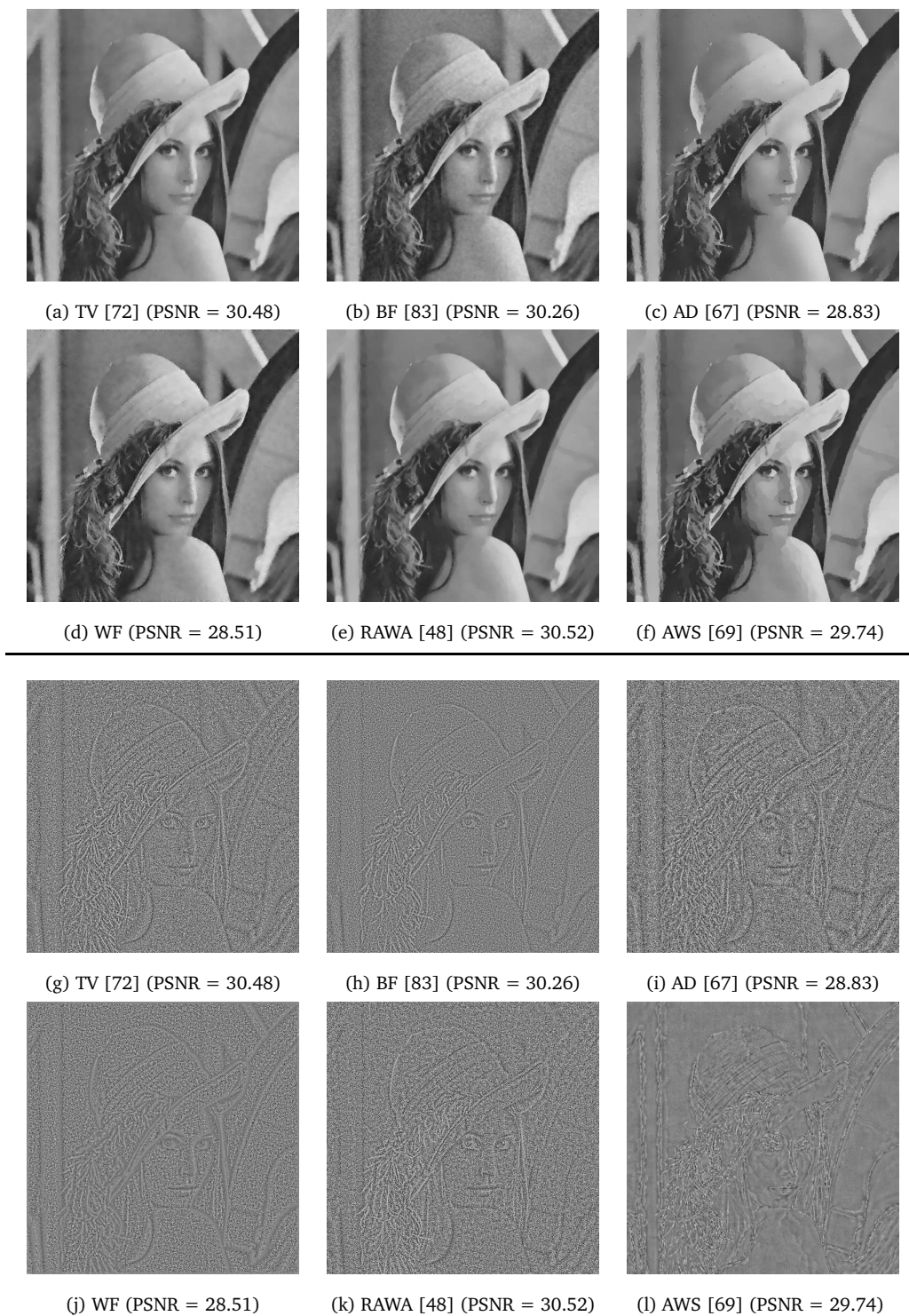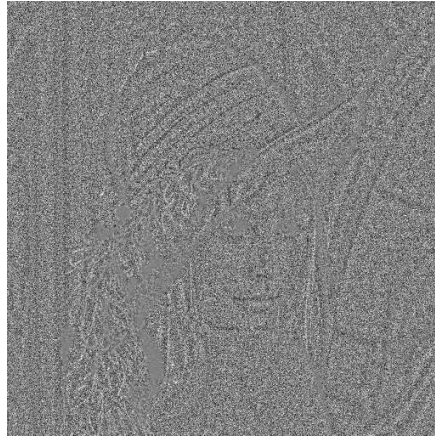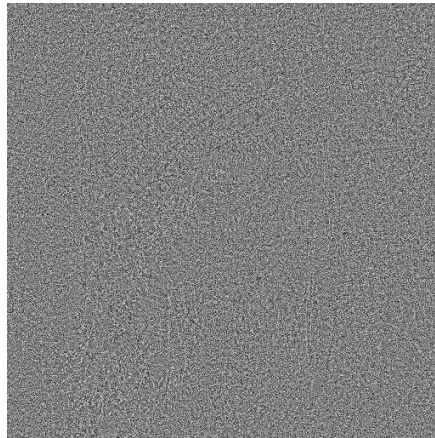*Figure 3.* Denoising of the noisy (WGN) *Lena* 512×512 image ($\sigma = 20$).

(a) TV [72] (PSNR = 30.48) (b) BF [83] (PSNR = 30.26) (c) AD [67] (PSNR = 28.83)

(d) WF (PSNR = 28.51) (e) RAWA [48] (PSNR = 30.52) (f) AWS [69] (PSNR = 29.74)

(g) TV [72] (PSNR = 30.48) (h) BF [83] (PSNR = 30.26) (i) AD [67] (PSNR = 28.83)

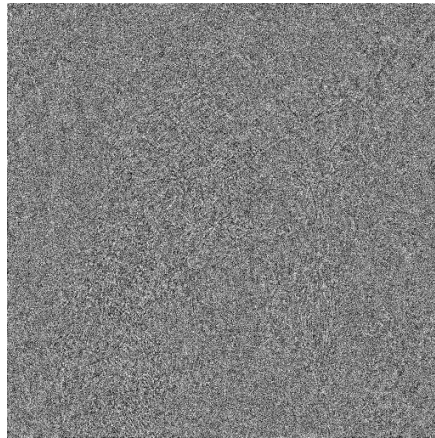(j) WF (PSNR = 28.51) (k) RAWA [48] (PSNR = 30.52) (l) AWS [69] (PSNR = 29.74)

Figure 4. Comparison with restoration methods applied to the noisy (WGN) *Lena* image ($\sigma = 20$): (a) Total Variation (TV) minimizing process [72], (b) bilateral filtering (BF) [83], (c) anisotropic diffusion (AD) [67], (d) Wiener filtering (WF), (e) robust adaptive window approach (RAWA) [48], (f) adaptive weights smoothing (AWS) [69], (g)-(l) estimated noise components for each method.

(a) *Non-Local means* denoising method [14] (PSNR=31.09)

(b) our exemplar-based denoising method (PSNR=32.64)

(c) wavelet-based denoising method [68] (PSNR=32.20)

Figure 5. Comparisons with the *Non-Local means* algorithm [14] and a wavelet-based denoising method [68] when applied to the noisy (WGN) *Lena* image ($\sigma = 20$).
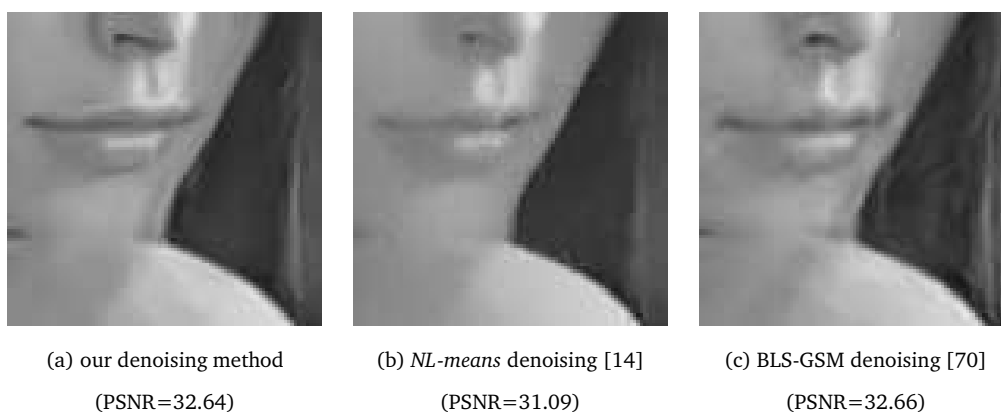
(a) our denoising method   (b) *NL-means* denoising [14]   (c) BLS-GSM denoising [70]

(PSNR=32.64)   (PSNR=31.09)   (PSNR=32.66)

Figure 6. Comparisons with the *Non-Local means* filter [14] and a wavelet-based denoising method [70] when applied to the noisy (WGN) *Lena* image ($\sigma = 20$).
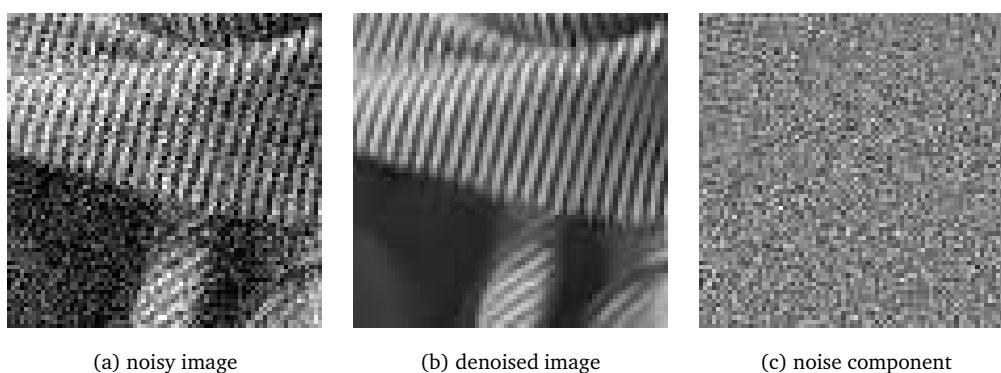

(a) noisy image   (b) denoised image   (c) noise component

Figure 7. Results with patch sizes of 9×9 pixels when the algorithm is applied to the noisy (WGN) *Barbara* 512×512 image ($\sigma = 20$).

(a) 3×3 patch   (b) 5×5 patch   (c) 9×9 patch

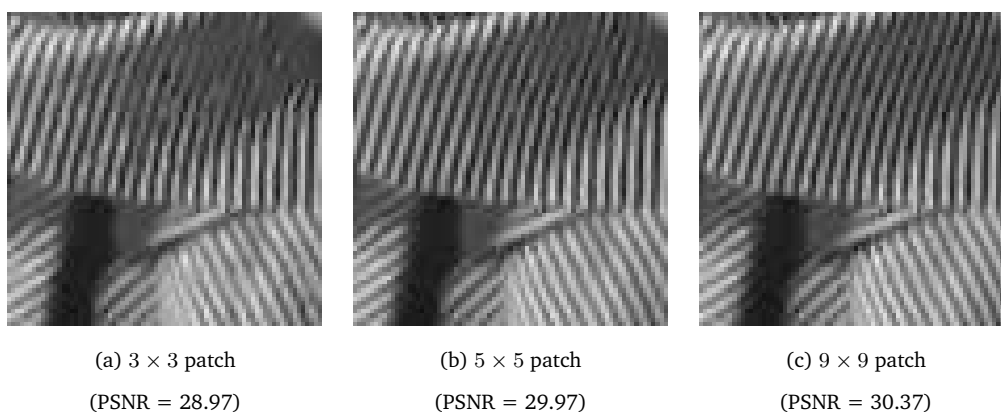(PSNR = 28.97)   (PSNR = 29.97)   (PSNR = 30.37)

Figure 8. Results with different patch sizes of $\sqrt{p} \times \sqrt{p}$ pixels when the algorithm is applied to the noisy (WGN) *Barbara* 512×512 image ($\sigma = 20$).

Table III. PSNR values when our exemplar-based restoration method ($N = 4, \alpha = 0.01$) with different patch size and sub-sampling (factor 2) is applied to noisy (WGN) images ($\sigma = 20$).

# points / $\sqrt{p} \times \sqrt{p}$ patch	$\lambda_{p,0.99}$	Lena 512×512	Barbara 512×512	Boat 512×512	House 256×256	Peppers 256×256
9 points / 3×3	21.67	32.13	28.97	29.86	32.69	30.86
25 points / 5×5	44.31	32.52	29.97	30.15	33.05	30.98
49 points / 7×7	74.92	32.63	30.27	30.17	33.03	30.80
81 points / 9×9	113.5	32.64	30.37	30.12	32.90	30.59
25 points / 9×9	44.31	32.27	29.84	29.64	32.46	30.26
36 points / 11×11	58.62	32.26	29.84	29.51	32.57	29.52

are close for every patch sizes and the optimal patch size depends on the image contents; a 9×9 patch seems appropriate in most cases and a smaller patch can be considered for processing piecewise smooth images.

To demonstrate the spatial adaptation, we present in Fig. 9, a real image corrupted by WGN ($\sigma = 20$) and denoised by our method (Fig. 9 - left). The spatial distribution of window sizes is shown in Fig. 9 (right) for a largest set of window sizes. For this experiment, the threshold ϱ is set to a very low value for demonstration. Clearly, the largest windows are located in the smoother parts in the image.

Finally, the robustness to noise is illustrated on the 512×512 *mandrill* image by varying σ from 5 to 50 (Fig. 10).

We have also compared our method to the best available published results when very competitive methods [71, 70, 68] were applied to the same image dataset [70]. These results were taken from the corresponding publications. We point out that, visually and quantitatively, our very simple and unsupervised algorithm method favorably compares to any of these denoising algorithms, including the more sophisticated wavelet-based denoising methods (see Fig. 5c-6c). Note that our method yields an improved PSNR for a wide range of variance as compared to existing methods. If the PSNR gains are marginal for some images, the visual difference can be significant as shown Fig. 5 where less artifacts are visible using our method. To complete the experiments, Table IV shows the PSNR values using our exemplar-based restoration method when applied to this set of test images for a wide range of noise variances as in [70, 71].

During the reviewing period of this manuscript, other patch-based methods for denoising have been published. The best results (in terms of PSNR) have been recently obtained by filtering in 3D transform domain and combining sliding-window transform processing with block-matching [23]. More recently, Elad et al. [29] and Mairal et al. [61] have proposed to estimate simultaneously the patch dictionary and the image by using the *so-called* K-SVD algorithm. The results are very close to those obtained by Dabov et al. [23] and slightly higher than our results (e.g. see Table I). Finally, Azzabou et al. have considered a variational approach [4] inspired by the method presented in this paper; they have introduced an alternative energy functional and obtained comparable experimental results on the same image dataset. More recently, Brox and Cremers have proposed an

Table IV. Performance of our exemplar-based restoration method ($p = 9^2$, $N = 4$, $\alpha = 0.01$) when applied to test noisy (WGN) images.

σ/PSNR	Lena 512 × 512	Barbara 512 × 512	Boat 512 × 512	House 256 × 256	Peppers 256 × 256
5 / 34.15	37.91	37.12	36.14	37.62	37.34
10 / 28.13	35.18	33.79	33.09	35.26	34.07
15 / 24.61	33.70	31.80	31.44	34.08	32.13
20 / 22.11	32.64	30.37	30.12	32.90	30.59
25 / 2017	31.73	29.24	29.20	32.22	29.73
50 / 14.15	28.38	24.09	25.93	28.67	25.29
75 / 10.63	25.51	22.10	23.69	25.49	22.31
100 / 8.13	23.32	20.64	21.78	23.08	20.51

iterated non-local means algorithm and functionals which share some common properties with the proposed method, for texture restoration [13] (see also [39]).
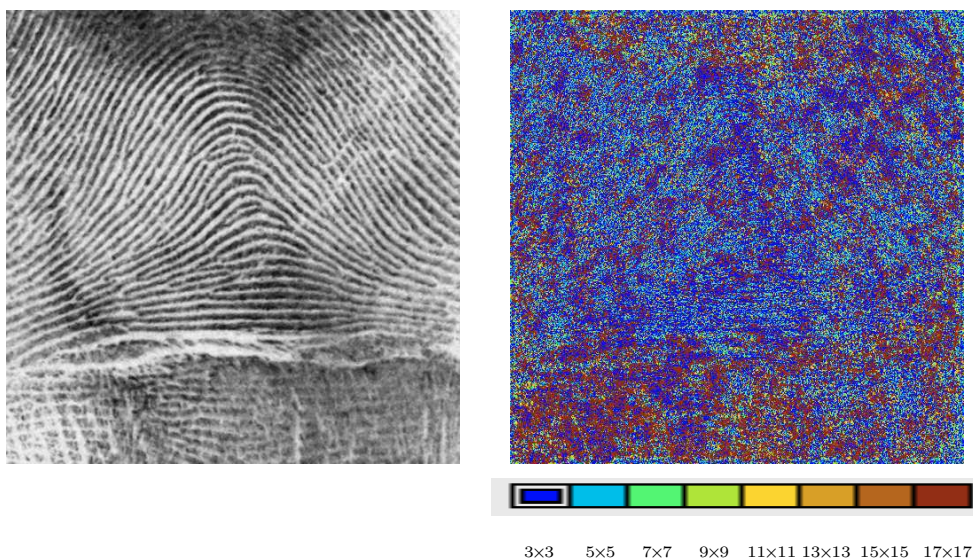

3×3   5×5   7×7   9×9   11×11 13×13 15×15 17×17

Figure 9. Denoising of a 512×512 image (WGN, $\sigma = 20$). left: denoised image (PSNR=28.17); right: spatial distribution of window sizes.

7.2. Denoising of real noisy images with artifacts. In the second part of experiments, the effects of the exemplar-based restoration is approach are illustrated on corrupted images with assumed additive non-Gaussian noise. The set of parameters is unchanged for processing all these test images: $p = 9^2, N = 4, \alpha = 0.01$. In most cases, a good compromise between the amount of smoothing and preservation of edges and textures is automatically
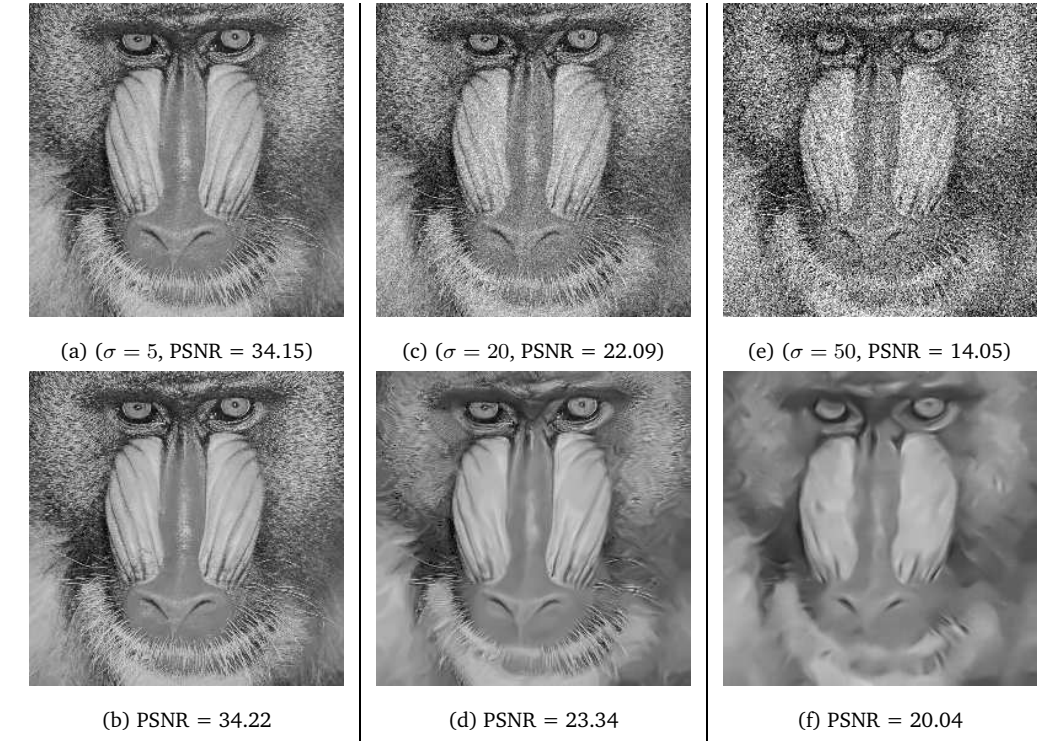
(a) ($\sigma = 5$, PSNR = 34.15)   (c) ($\sigma = 20$, PSNR = 22.09)   (e) ($\sigma = 50$, PSNR = 14.05)

(b) PSNR = 34.22   (d) PSNR = 23.34   (f) PSNR = 20.04

Figure 10. Results on a the noisy 512×512 *mandrill* image artificially corrupted with different signal-to-ratio levels (WGN) (top: noisy images, bottom: denoised images).

reached. In that case, the noise variance $\widehat{\sigma}^2$ is automatically estimated from image data. For illustration , Fig. 11 shows the four intermediate results obtained at each iteration of the algorithm ($N = 4$). The estimated noise component corresponding to the difference between the denoised image (Fig. 11e) and the noisy image (Fig. 11a) is shown in Fig. 11f. Note that edges and geometric structures are well preserved and scan effects are mostly removed. In Fig. 12, we demonstrate that the algorithm is able to remove JPEG artifacts and block effects due to the DCT compression. Finally, we have extended the method to restore color images and an example of an old painting is shown in Fig. 13. The image has been smoothed and cracks are visually partially removed.

Additional examples that demonstrate the performance of the method can be also found in [49, 50].

7.3. Image denoising in bio-imaging. We have also tested the algorithm on 2D and 3D confocal fluorescence microscopy images. Some of the current applications in biological studies are in neuron research. The 271×238 confocal image depicts neural cells (Fig. 14 (top)). The image, denoised using the set of parameters used in the previous experiments, contains larger homogeneous areas than the original 2D image and can be more easily segmented. Finally, the same denoising process has been applied to a 360×372 image showing nuclei in a embryo specimen (Fig. 14 (bottom)). In both cases, spatially-varying noise is reduced and structure is preserved in the restored images.
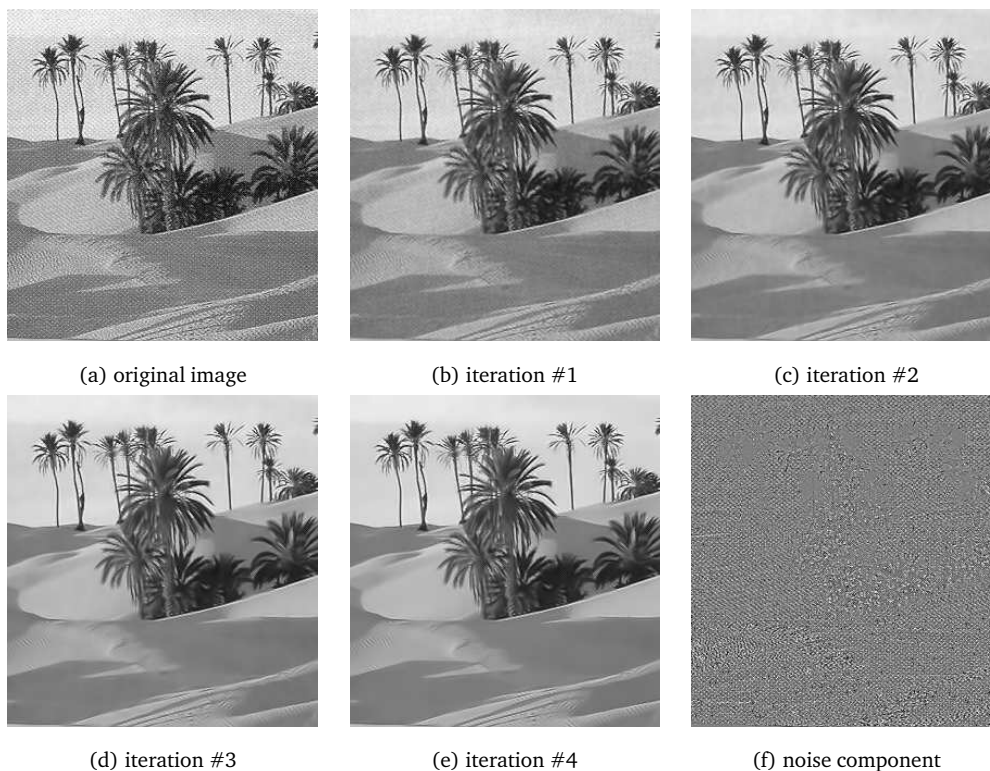
(a) original image   (b) iteration #1   (c) iteration #2

(d) iteration #3   (e) iteration #4   (f) noise component

Figure 11. Results obtained at each iteration of the algorithm ($N = 4$) when applied to a real 286×305 image with artifacts (courtesy of D. Tschumperlé - `http://www.greyc.ensicaen.fr/~dtschump/greycstoration`).

8. Conclusion

We have described a novel feature-preserving adaptive restoration algorithm where local image patches and variable window sizes are jointly used. The proposed smoothing scheme provides an alternative method to anisotropic diffusion, bilateral filtering and *NL-means filtering* [14, 15]. Our straightforward and unsupervised method yields a significant improvement in image denoising, and achieves performances almost always superior to the best wavelet-based denoising algorithms. We believe this method represents an important step forward for the use of neighborhood design that captures spatial dependencies in images. Unlike previous most exemplar-based methods that use learning algorithms, our method is unsupervised and fully automatic since control parameter are easily calibrated with statistical arguments. Experimental results demonstrate its potential for a large variety of images, included in bio-imaging (we refer the readers to [49] and to the following web page `http://www.irisa.fr/vista/Themes/Demos/Debruitage/ImageDenoising.html` to have more visual elements). More recently, the method has been extended to spatio-temporal data and used for video denoising [10].
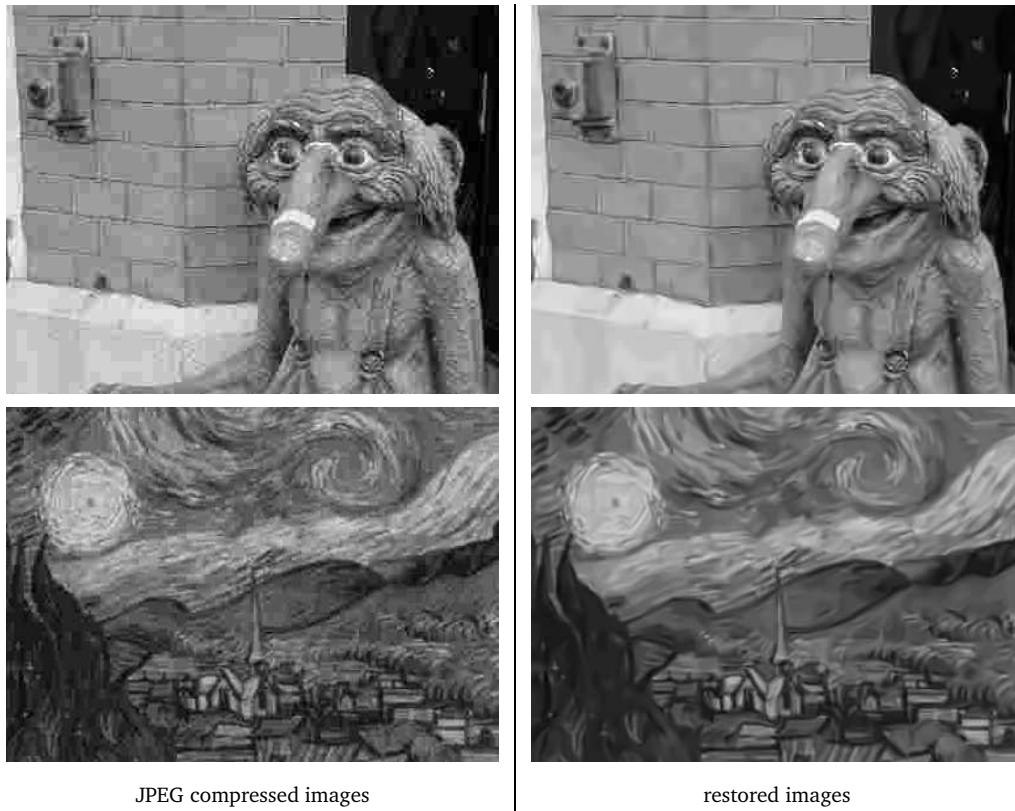
JPEG compressed images   restored images

Figure 12. Restoration of real images with artifacts (JPEG compression) (courtesy of D. Tschumperlé - `http://www.greyc.ensicaen.fr/~dtschump/greycstoration`).

Figure 13. Patch-based restoration (crack removal) and smoothing of a 391×384 color image (`http://restoreinpaint.sourceforge.net/`).

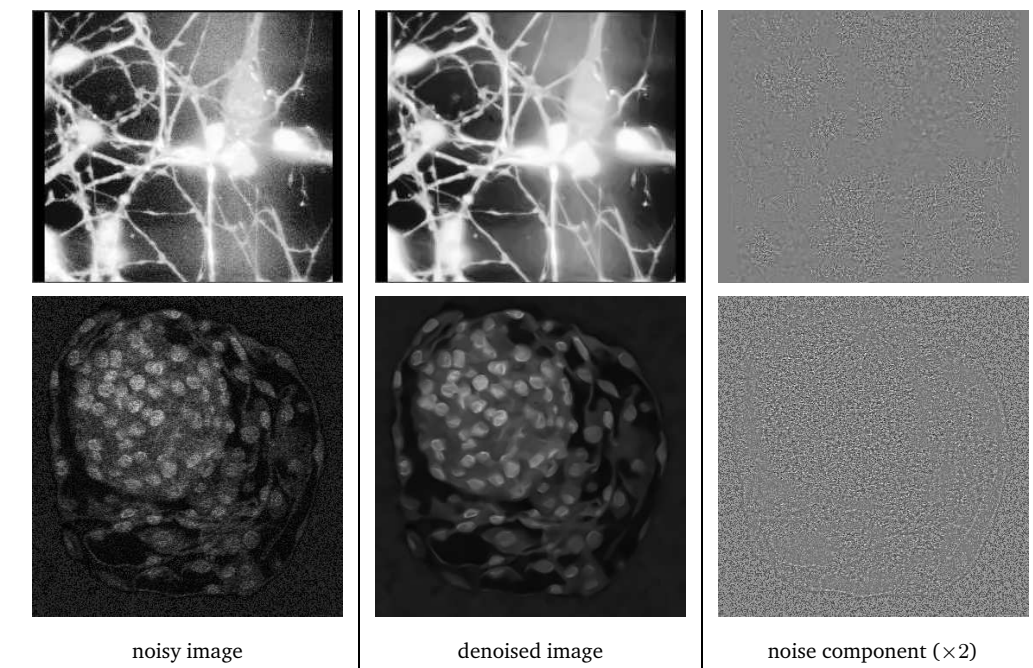| noisy image | denoised image | noise component ($\times 2$) |

Figure 14. Results on 2D images: top: neural cells; bottom: nuclei in a embryology specimen.

References

1. J.F. Aujol, G. Aubert, L. Blanc-Féraud, A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *J. Math. Imag. Vis.*, 22(1): 71-88, 2005.
2. V. Aurich, J. Weule. Nonlinear Gaussian filters performing edge preserving diffusion. In *Proc. 17th DAGM Symposium*, pp. 538-545, Bielefeld, Germany, 1995.
3. S. P. Awate, R. T. Whitaker. Higher-order image statistics for unsupervised, information-theoretic, adaptive image filtering. In *Proc. Comp. Vis. Patt. Recog. (CVPR'05)*, Vol. 2, pp. 44-51, San Diego, CA, 2005.
4. N. Azzabou, N. Paragios, F. Cao, F. Guichard. Variable bandwidth image denoising using image-based noise models. In *Proc. Conf. Comp. Vis. Patt. Recog. (CVPR'07)*, pp. 1-7, Minneapolis, Minnesota, 2007.
5. D. Barash. A fundamental relationship between bilateral filtering, adaptive smoothing and the nonlinear diffusion equation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 24(6): 844-847, 2002.
6. D. Barash, D. Comaniciu. A Common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean-shift. *Image and Video Computing*, 22(1): 73-81, 2004.
7. A. Blake, A. Zisserman. *Visual Reconstruction*, MIT Press, 1987.
8. M.J. Black, G. Sapiro, D.H. Marimont, D. Heeger. Robust anisotropic diffusion. *IEEE Trans. Image Process.*, 7(3): 421-432, 1998.
9. M.J. Black, G. Sapiro. Edges as outliers: Anisotropic smoothing using local image statistics. In *Proc. Scale-Space Th. Comp. Vis. (Scale-Space'99)*, LNCS 1682, pp. 259-270, Kerkyra, Greece, 1999.
10. J. Boulanger, K. Kervrann, P. Bouthemy. Space-time adaptation for patch based image sequence restoration. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(6):1096-1102, 2007.

11. Y. Boykov, O. Veksler, R. Zabih. A variable window approach to early vision. *IEEE Trans. Patt. Anal. Mach. Intel.*, 20(12): 1283-1294, 1998.
12. T. Brox, J. Weickert. A TV flow based local scale measure for texture discrimination. In *Proc. Eur. Conf. Comp. Vis. (ECCV'04)*, Vol. 2, pp. 578-590, Prague, Czech Republic, 2004.
13. T. Brox, D. Cremers. Iterated non-local means for texture restoration. In *Proc. Conf. Scale-Space and Variational Meth. (SSVM' 07)*, Ischia, Italy, 2007.
14. A. Buades, B. Coll, J.-M. Morel, A review of image denoising algorithms, with a new one. *SIAM J. Multiscale Model. Simul.*, 4(2): 490-530, 2005.
15. A. Buades, B. Coll, J.-M. Morel. A non local algorithm for image denoising. In *Proc. Comp. Vis. Patt. Recogn. (CVPR'05)*, Vol. 2, pp. 60-65, San Diego, CA, 2005.
16. F. Catte, P.-L. Lions, J.-M. Morel, T. Coll. Image selective smoothing and edge-detection by nonlinear diffusion. *SIAM J. Numer. Anal.*, 29(1): 182-193, 1992.
17. T.F. Chan, S. Osher, J. Shen. The digital TV filter and nonlinear denoising. *IEEE Trans. Image Process.*, 10(2): 231-241, 2001.
18. I. Cheng. Mean-shift, mode seeking, and clustering. *IEEE Trans. Patt. Anal. Mach. Intel.*, 17(8): 790-7999, 1995.
19. C.K. Chu, K. Glad, F. Godtliebsen, J.S. Marron. Edge-preserving smoothers for image processing. *J. Am. Stat. Ass.*, 93(442): 526-555, 1998.
20. D. Comaniciu, V. Ramesh, P. Meer. The variable bandwidth mean-shift and data-driven scale selection. In *Proc. Int. Conf. Comp. Vis. (ICCV'01)*, Vol 1, pp. 438-445, Vancouver, Canada, 2001.
21. D. Comaniciu, P. Meer. Mean-shift: a robust approach toward feature space analysis. *IEEE Trans. Patt. Anal. Mach. Intel.*, 24(5): 603-619, 2002.
22. A. Criminisi, P. Pérez, K. Toyama. Region filling and object removal by exemplar-based inpainting. *IEEE Trans. Image Process.*, 13(9): 1200-1212, 2004.
23. K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8), 2007.
24. J.S. De Bonet. Noise reduction through detection of signal redundancy. Rethinking Artificial Intelligence, MIT AI Lab, 1997.
25. D.L. Donoho, I.M. Johnston. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81: 425-455, 1994.
26. D.L. Donoho, I.M. Johnston. Denoising by soft-thresholding. *IEEE Trans. Inf. Th.*, 41: 613-627, 1995.
27. A. Efros, T. Leung. Texture synthesis by non-parametric sampling. In *Proc. Int. Conf. Comp. Vis. (ICCV'99)*, pp. 1033-1038, Kerkyra, Greece, 1999.
28. M. Elad. On the bilateral filter and ways to improve it. *IEEE Trans. Image Process.*, 11(10): 1141-1151, 2002.
29. Elad, M. Aharon, M.: Image denoising via learned dictionaries and sparse representation. In *Proc. Conf. Vis. Patt. Recog. (CVPR'06)*, Vol. 1, pp. 895-900, New-York, 2006.
30. B. Fischl, E.L. Schwartz. Adaptive nonlocal filtering: a fast alternative to anisotropic diffusion for image enhancement. *IEEE Trans. Patt. Anal. Mach. Intell.*, 21(1): 42-48, 1999.
31. A. Fitzgibbon, Y. Wexler, A. Zisserman. Image-based rendering using image-based priors. In *Proc. Int. Conf. Comp. Vis. (ICCV'03)*, Nice, France, 2003.
32. W. T. Freeman, E. C. Pasztor, O. T. Carmichael. Learning low-level vision. *Int. J. Comp. Vis.*, 40(1): 25-47, 2000.
33. T. Gasser, L. Sroka, C. Jennen Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73: 625-633, 1986.
34. D. Geman, D. Geman. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 6(6): 721-741, 1984.
35. D. Geman, S. Geman, C. Graffigne, P. Dong. Boundary detection by constrained optimization. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(7): 609-628, 1990.

36. M. Ghazel, G.H. Freeman, E.R. Vrscay. Fractal image denoising. *IEEE Trans. Image Process.*, 12(12): 1560-1578, 2003.

37. I. Gijbels, A. Lambert, P. Qiu. Edge-preserving image denoising and estimation of discontinuous surfaces. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(7): 1075- 1087, 2006.

38. G. Gilboa, Y.Y. Zeevi, N. Sochen. Texture preserving variational denoising using an adaptive fidelity term. In *Proc. VLSM'03*, Nice, France, 2003.

39. G. Gilboa, S. Osher. Nonlocal linear image regularization and supervised segmentation. *SIAM J. Multiscale Model. Simul.* **6** (2007) 595-630.

40. F. Godtliebsen, E. Spjotvoll, J.S. Marron. A nonlinear Gaussian filter applied to images with discontinuities. *J. Nonparam. Statist.*, 8: 21-43, 1997.

41. A. Goldenshluger, A. Nemirovsky. On spatial adaptive estimation of nonparametric regression. *Math. Meth. Statist.*, 6(2): 135-170, 1997.

42. G. Gomez, J.L. Marroquin, L.E. Sucar. Probabilistic estimation of local scale. In *Proc. Int. Conf. Patt. Recog. (ICPR'00)*, Vol. 3, pp. 798-801, Barcelona, Spain, 2000.

43. W. Hardle, O. Linton. Applied nonparametric methods. *In R.F. Engle, D.L. McFadden ed., Handbook of Econometrics*, Vol. IV, Amsterdam, North Holland, pp. 2295-2381, 1994.

44. N. Jojic, B. Frey, A. Kannan. Epitomic analysis of appearance and shape. In *Proc. Int Conf. Comp. Vis. (ICCV'03)*, Vol. 1, pp. 34-41, Nice, France, 2003.

45. A. Juditsky. Wavelet estimators: adapting to unknown smoothness. *Math. Meth. Statist.*, 1:1-20, 1997.

46. V. Katkovnik, K. Egiazarian, J Astola. Adaptive window size image denoising based on intersection of confidence intervals (ICI) rule. *J. Math. Imag. Vis.*, 16(3): 223-235, 2002.

47. C. Kervrann, F. Heitz. A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics. *IEEE Trans. Image Process.*, 4(6): 856-862, 1995.

48. C. Kervrann. An adaptive window approach for image smoothing and structures preserving. In *Proc. Eur. Conf. Comp. Vis. (ECCV'04)*, Vol. 3, pp. 132-144, Prague, Czech Republic, 2004.

49. C. Kervrann, J. Boulanger. Local adaptivity to variable smoothness for exemplar-based image denoising and representation. INRIA Research Report, RR-5624 , July 2005.

50. C. Kervrann, J. Boulanger. Unsupervised patch-based image regularization and representation. In *Proc. Eur. Conf. Comp. Vis. (ECCV'06)*, Vol. 4, pp. 555-567, Graz, Austria, 2006.

51. C. Kervrann, J. Boulanger, P. Coupé. Bayesian non-local means filter, image redundancy and adaptive dictionaries for noise removal. In *Proc. Conf. Scale-Space and Variational Meth. (SSVM' 07)*, Ischia, Italy, 2007.

52. S. Kinderman, S. Osher, P.W. Jones. Deblurring and denoising of images by nonlocal functionals. *SIAM J. Multiscale Model. Simul.* **4** (2005) 1091-1115

53. J.S. Lee. Digital image smoothing and the sigma filter. *Comp. Vis. Graph. Image Process.*, 24: 255-269, 1983.

54. E. Le Pennec, S. Mallat. Sparse geometric image representation with Bandelets. *IEEE Trans. Image Process.*, 14(4):423-438, 2005.

55. O. Lepskii. On a problem of adaptive estimation on white Gaussian noise. *Th. Prob. Appl.*, 35: 454-466, 1990.

56. O. Lepskii. Asymptotically minimax adaptive estimation 1: uppers bounds. *SIAM J. Th. Prob. Appl.*, 36(4): 654-659, 1991.

57. O. V. Lepski, E. Mammen, V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3): 929-947, 1997.

58. T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *Int. J. Comp. Vis.*, 30(2): 117-154, 1998.

59. M. Maurizot, P. Bouthemy, B. Delyon, A. Juditski, J.-M. Odobez. Determination of singular points in 2D deformable flow fields. In *IEEE Proc. Int. Conf. Image Processing (ICIP'95)*, Vol. 3, pp. 488-491, Washington DC, 1995.

60. Y. Meyer. Oscillating patterns in image processing and nonlinear evolution equations. *University Lecture Series*, 22, AMS 2002.

61. J. Mairal, G. Sapiro, M. Elad. Multiscale sparse image representation with learned dictionaries. In *Proc. Int. Conf. Image Processing (ICIP'07)*, San Antonio, Texas, USA, 2007.

62. P. Mrazek. Selection of optimal stopping time for nonlinear diffusion filtering. *Int. J. Comp. Vis.*, 52(2/3): 189-203, 2003.

63. P. Mrazek, J. Weickert, A. Bruhn. On robust estimation and smoothing with spatial and tonal kernels. *Preprint no 51*, University of Bremen, Germany, 2004.

64. D. Mumford, J.Shah. Optimal approximations by piecewise smooth functions and variational problems. *Comm. Pure Appl. Math.*, 42(5): 577-685, 1989.

65. M. Nitzberg, T. Shiota. Nonlinear image filtering with edge and corner enhancement. *IEEE Trans. Patt. Anal. Mach. Intel.*, 14(8): 826-833, 1992.

66. S. Osher, A. Solé, L. Vese. Image decomposition and restoration using total variation minimization and the H^{-1} norm. *SIAM J. Multiscale Model. Simul.*, 1(3): 349-370, 2003.

67. P. Perona, J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(7): 629-239, 1990.

68. A. Pizurica, W. Philips. Estimating probability of presence of a signal of interest in multiresolution single and multiband image denoising. *IEEE Trans. Image Process. *, 15(3):654-665, 2006.

69. J. Polzehl, V. Spokoiny. Adaptive weights smoothing with application to image restoration. *J. R. Stat. Soc. B.*, 62(2): 335-354, 2000.

70. J. Portilla, V. Strela, M. Wainwright, E. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.*, 12(11): 1338-1351, 2003.

71. S. Roth, M.J. Black. Fields of experts: a framework for learning image priors with applications. In *Proc. Comp. Vis. Patt. Recog. (CVPR'05)*, Vol. 2, pp. 860-867, San Diego, CA, 2005.

72. L. Rudin, S. Osher, E. Fatemi. Nonlinear Total Variation based noise removal algorithms. *Physica D*, 60: 259-268, 1992.

73. P. Saint-Marc, J.S. Chen, G. Médioni. Adaptive smoothing: a general tool for early vision. *IEEE Trans. Patt. Anal. Mach. Intel.*, 13(6): 514-529, 1991.

74. D. W. Scott. *Multivariate Density Estimation*. Wiley, 1992.

75. M. Singh, N. Ahuja. Regression based bandwidth selection for segmentation using Parzen windows. In *Proc. Int Conf. Comp. Vis. (ICCV'03)*, Vol. 1, pp. 2-9, Nice, France, 2003.

76. S.M. Smith, M. Brady. SUSAN - a new approach to low-level image processing. *Int. J. Comp. Vis.*, 23(1): 45-78, 1997.

77. V. G. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, 26(4): 141-170, 1998.

78. L. Stankovic. Performance analysis of the adaptive algorithm for bias-to-variance trade-off. *IEEE Trans. Signal Process.*, 52(5): 1228-1234, 2004.

79. N. Sochen, R. Kimmel, A. M. Bruckstein. Diffusions and Confusions in Signal and Image Processing. *J. Math. Imag. Vis.*, 14(3): 237-244, 2001.

80. A. Spira, R. Kimmel, N. Sochen. Efficient Beltrami flow using a short-time kernel. In *Proc. Int. Conf. Scale-Space Meth. Comp. Vis. (Scale-Space'03)*, pp. 551-522, Isle of Skye, Scotland, 2003.

81. J.L. Starck, E. Candes, D.L. Donoho. The Curvelet transform for image denoising. *IEEE Trans. on Image Process. *, 11(6): pp. 670-684, 2002.

82. C.V. Stewart, C.-L. Tsai, B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration *IEEE Trans. Med. Imag. *, 22(11): 1379-1394, 2003.

83. C. Tomasi, R. Manduchi. Bilateral filtering for gray and color images. In *Proc. Int Conf. Comp. Vis. (ICCV'98)*, pp. 839-846, Bombay, India, 1998.

84. D. Tschumperlé, D. Curvature-preserving regularization of multi-valued images using PDE's. In *Proc. Eur Conf. Comp. Vis. (ECCV'06)*, Vol. 2, pp. 295-307, Graz, Austria, 2006.

85. R. van den Boomgaard, J. van de Weijer. On the equivalence of local-mode finding, robust estimation and mean-shift analysis as used in early vision tasks. In *Proc. Int. Conf. Patt. Recogn. (ICPR'02)*, Vol. III, pp. 927-930, Quebec City, Canada, 2002.

86. Wang, Z., Zhang, D.: Restoration of impulse noise corrupted images using long-range correlation. IEEE Signal Process. Lett., 5: 4-6, 1998.

87. J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart, 1998.

88. J. Weickert. Coherence-enhancing diffusion filtering. *Int. J. Comp. Vis.*, 31(2/3): 111-127, 1999.

89. J. van de Weijer, R. van den Boomgaard. Local mode filtering. In *Proc. Comp. Vis. Patt. Recogn. (CVPR'01)*, vol. II, pp. 428-433, Kauai, Hawaii, 2001.

90. G.Z. Yang, P. Burger, D.N. Firmin, S.R. Underwood. Structure adaptive anisotropic image filtering. *Image Vis. Comput.*, 14: 135-145, 1996.

91. L.P. Yaroslavsky, M. Eden *Fundamentals of digital optics*, Birkhaüser, Boston, 1996.

92. D. Zhang, Z. Wang: Image information restoration based on long-range correlation. IEEE Trans. Circ. Syst. Video Technol. 12: 331-341, 2002.

93. S.C. Zhu, Y. Wu, D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *Int. J. Comp. Vis.* 27(2): 107-126, 1998.

Appendix

A.1. Proof of the inequality: $v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'}) \leq v^2(\widehat{u}_{i,n'})$, $\forall n' < n$.

We have $\widehat{u}_{i,n} - \widehat{u}_{i,n'} \sim \mathcal{N}(0, v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'}))$ since both biases are negligible ($\mathbb{E}[\widehat{u}_{i,n} - \widehat{u}_{i,n'}] \approx 0$) and then $v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'})) = \mathbb{E}\left[|\widehat{u}_{i,n} - \widehat{u}_{i,n'}|^2\right]$. We recall that $Y_i = u_{\text{true}}(\mathbf{x}_i) + \epsilon_i$ and

$$\widehat{u}_{i,n} = \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n}\, Y_j \quad \text{and} \quad \widehat{u}_{i,n'} = \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\, Y_j.$$

Since $\Delta_{i,n'} \subset \Delta_{i,n}$, we write

$$v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'}))$$

$$= \mathbb{E}\left[\left(\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n}\, \epsilon_j + u_{\text{true}}(\mathbf{x}_i) \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n} - \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\, \epsilon_j - u_{\text{true}}(\mathbf{x}_i) \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n}\, \epsilon_j - \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\, \epsilon_j\right)\left(\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n}\, \epsilon_j - \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\, \epsilon_j\right)\right]$$

$$= \sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\pi_{i\sim j,n})^2 + \sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n'}} (\pi_{i\sim j,n'})^2 - 2\, \mathbb{E}\left[\sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n}\, \epsilon_j \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\, \epsilon_j\right]$$

$$= \sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\pi_{i\sim j,n})^2 + \sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n'}} (\pi_{i\sim j,n'})^2 - 2\sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'} \pi_{i\sim j,n}$$

$$= v^2(\widehat{u}_{i,n}) + v^2(\widehat{u}_{i,n'}) - 2\sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\pi_{i\sim j,n}.$$

In addition, as $\pi_{i\sim j,n'} := 0$ when $\mathbf{x}_j \notin \Delta_{i,n'}$, it follows that

$$\sum_{\mathbf{x}_j \in \Delta_{i,n'}} \pi_{i\sim j,n'}\pi_{i\sim j,n} \geq \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n'}\pi_{i\sim j,n}.$$

By definition, $\pi_{i\sim j,n'} \geq \pi_{i\sim j,n}$ for $\mathbf{x}_j \in (\Delta_{i,n} \cap \Delta_{i,n'})$, hence

$$\sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} \pi_{i\sim j,n'}\pi_{i\sim j,n} \geq \sigma^2 \sum_{\mathbf{x}_j \in \Delta_{i,n}} (\pi_{i\sim j,n})^2 = v^2(\widehat{u}_{i,n}).$$

Finally, we obtain the inequality:

$$v^2(\widehat{u}_{i,n} - \widehat{u}_{i,n'})) \leq v^2(\widehat{u}_{i,n'}) + v^2(\widehat{u}_{i,n}) - 2v^2(\widehat{u}_{i,n}) \leq v^2(\widehat{u}_{i,n'}) - v^2(\widehat{u}_{i,n}) \leq v^2(\widehat{u}_{i,n'}).$$

$$\square$$

A.2. Proof of Proposition 1. If $X = \dfrac{\widehat{u}_{i,n} - \widehat{u}_{i,n'}}{v(\widehat{u}_{i,n'})}$ is $\mathcal{N}(0,1)$, then

$$\mathbb{E}\exp(\varrho X) = \exp\left(\frac{\varrho^2}{2}\right).$$

From this inequality and the exponential Chebychev's inequality stated as follows:

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[g(X)]}{g(a)}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a positive monotonous increasing function and $a \in \mathbb{R}$, we have

$$\mathbb{P}((\widehat{u}_{i,n} - \widehat{u}_{i,n'}) > \varrho v(\widehat{u}_{i,n'})) \leq \frac{\mathbb{E}\exp\left(\frac{\varrho(\widehat{u}_{i,n}-\widehat{u}_{i,n'})}{v(\widehat{u}_{i,n'})}\right)}{\exp \varrho^2}.$$

by taking $g(x) = \exp(\varrho x)$ and $a = \varrho$. Hence

$$\mathbb{P}((\widehat{u}_{i,n} - \widehat{u}_{i,n'}) > \varrho v(\widehat{u}_{i,n'})) \leq \exp\left(-\frac{\varrho^2}{2}\right)$$

and the following result comes from the symmetry of the normal distribution.

$$\mathbb{P}(|\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})) \leq 2\exp\left(-\frac{\varrho^2}{2}\right).$$

To prove Proposition 1., we note also that

$$\{\widehat{n}(\mathbf{x}_i) = n\} = \{\exists n' \in \{1,\ldots,n-1\} : |\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})\}$$
$$\subseteq \bigcup_{n'<n} \{|\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})\}.$$

Using this definition, we get

$$\mathbb{P}(\widehat{n}(\mathbf{x}_i) = n) \leq \sum_{n' < n} \mathbb{P}(|\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})) \leq \sum_{n' < n} 2 \exp\left(-\frac{\varrho^2}{2}\right)$$

$\square$

A.3. Proof of Proposition 2. Let the condition of the theorem be satisfied, that is $n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)$, then, from the inequality (18), we have

$$|u_{\text{oracle}}(\mathbf{x}_i) - \widehat{u}(\mathbf{x}_i)| \leq (2\gamma(\mathbf{x}_i) + \varkappa)\, v(u_{\text{oracle}}(\mathbf{x}_i)).$$

By taking the expectation of this expression and using (22), we obtain:

$$\begin{aligned}
\mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - \widehat{u}(\mathbf{x}_i)|^2]^{1/2}\mathbf{1}(n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)) &\leq (2\gamma(\mathbf{x}_i) + \varkappa)\, v(u_{\text{oracle}}(\mathbf{x}_i)) \\
&\leq \frac{2\gamma(\mathbf{x}_i) + \varkappa}{\sqrt{1 + \gamma^2(\mathbf{x}_i)}}\mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2}.
\end{aligned}$$

Finally, by applying the triangular inequality, we get

$$\begin{aligned}
\mathbb{E}[|\widehat{u}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2}\mathbf{1}(n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)) &\leq \mathbb{E}[|\widehat{u}(\mathbf{x}_i) - u_{\text{oracle}}(\mathbf{x}_i)|^2]^{1/2}\mathbf{1}(n_{\text{oracle}}(\mathbf{x}_i) \leq \widehat{n}(\mathbf{x}_i)) \\
&\quad + \mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2} \\
&\leq \frac{2\gamma(\mathbf{x}_i) + \varkappa}{\sqrt{1 + \gamma^2(\mathbf{x}_i)}}\mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2} \\
&\quad + \mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2} \\
&\leq \left[\frac{2\gamma(\mathbf{x}_i) + \varkappa}{\sqrt{1 + \gamma^2(\mathbf{x}_i)}} + 1\right]\mathbb{E}[|u_{\text{oracle}}(\mathbf{x}_i) - u_{\text{true}}(\mathbf{x}_i)|^2]^{1/2}
\end{aligned}$$

$\square$

A.4. Proof of Proposition 3. We have

$$\{\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)\} \subseteq \bigcup_{n < n_{\text{oracle}}(\mathbf{x}_i)} \bigcup_{n' < n} \{|\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})\}.$$

and the probability of the event $\{\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)\}$ occurs can bounded

$$\mathbb{P}(\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)) \leq \sum_{n=1}^{n_{\text{oracle}}(\mathbf{x}_i)-1} \sum_{n'=1}^{n} \mathbb{P}(|u_{\text{oracle}}(\mathbf{x}_i) - \widehat{u}_{i,n}| > \varrho v(\widehat{u}_{i,n})).$$

In Appendix A.2, we proved $\mathbb{P}(|\widehat{u}_{i,n} - \widehat{u}_{i,n'}| > \varrho v(\widehat{u}_{i,n'})) \leq 2 \exp\left(-\varrho^2/2\right)$. Hence, we have

$$\begin{aligned}
\mathbb{P}(\widehat{n}(\mathbf{x}_i) < n_{\text{oracle}}(\mathbf{x}_i)) &\leq \sum_{n=1}^{n_{\text{oracle}}(\mathbf{x}_i)-1} \sum_{n'=1}^{n} 2 \exp\left(-\frac{\varrho^2}{2}\right) \\
&= n_{\text{oracle}}(\mathbf{x}_i)(n_{\text{oracle}}(\mathbf{x}_i) - 1) \exp\left(-\frac{\varrho^2}{2}\right)
\end{aligned}$$