A mathematical perspective of image denoising

Miguel Colom, Gabriele Facciolo, Marc Lebrun, Nicola Pierazzo,
Martin Rais, Yi-Qing Wang, and Jean-Michel Morel

Abstract. Digital images are matrices of regularly spaced samples, the pixels, each containing a photon count. Each pixel thus contains a random sample of a Poisson variable. Its mean would be the ideal image value at this pixel. It follows that all images are random discrete processes and therefore "noisy". Ever since digital images exist, numerical methods have been proposed to recover the ideal mean from its random observed value. This problem is obviously ill posed and makes sense only if there is an underlying image model. Inventing or learning from data a consistent mathematically image model is the core of the problem. Images being 2D projections of our complex surrounding visual world, this is a challenging problem, which is nevertheless beginning to find simple but mathematically innovative answers. We shall distinguish four classes of denoising principles, relying on functional or stochastic image models. We show that each of these principles can be summarized in a single formula. In addition these principles can be combined efficiently to cope with the full image complexity. This explains their immediate industrial impact. All current cameras and imaging devices rely directly on the simple formulas explained here. In the past ten years the image quality delivered to users has increased fast thanks to this exemplary mathematical modeling.

As an illustration of the universality and simplicity reached by the theory, most image denoising algorithms discussed in this paper can be tested directly on any digital image at *Image Processing On Line*, http://www.ipol.im/. In this web journal, each paper contains a complete algorithmic description, the corresponding source code, and can be run online on arbitrary images.

Mathematics Subject Classification (2010). Primary 62H35; secondary 68U10, 94A08.

Keywords. Image denoising, Fourier transform, Wiener estimate, wavelet threshold, discrete cosine transform, oracle estimate, Bayes formula, neighborhood filters, nonlocal methods, neural networks, blind denoising.

1. Introduction

Most digital images and movies are currently obtained by a matrix of sensors counting photons hitting the surface. We shall denote by $\mathbf{i}$ the indices of the matrix elements also called pixels. The value $\tilde{u}(\mathbf{i})$ observed by a sensor at a pixel $\mathbf{i}$ is a Poisson random variable whose mean $u(\mathbf{i})$ would be the ideal image. The difference between the observed image and the ideal image $\tilde{u}(\mathbf{i}) - u(\mathbf{i}) = n(\mathbf{i})$ is called "noise". By a well known property of Poisson random variables, the standard deviation of the noise $n(\mathbf{i})$ is equal to $\sqrt{u(\mathbf{i})}$. On a motionless scene with constant lighting, $u(\mathbf{i})$ can be approached by simply accumulating photons for a long exposure time, and by taking the temporal average of this photon count. Accumulating photon

Proceedings of the International Congress of Mathematicians, Seoul, 2014

impacts on a sensitive surface is therefore the essence of photography. The first Nicéphore Niépce photograph [16] was obtained after an eight hours exposure: it is very noisy, though! A digitization of it can be seen on the left hand side of Figure 1.1. The image in the middle is an attempt to denoise it. The image on the right is the "estimated noise", namely the difference between the noisy image and its denoised version. How was this done will be explained in section 7.

Figure 1.1. Left: A digitization of the first ever photograph by Nicéphore Niépce "View from the Window at Le Gras" ca. 1826 obtained after an eight hours exposure. Middle: an attempt to denoise it. Right: the "estimated noise", namely the difference between the noisy image and its denoised version.

Augmenting the exposure time of the camera amounts to augmenting the expectation $u(\mathbf{i})$ of the number of photons $\tilde{u}(\mathbf{i})$. The number of photons has mean $u(\mathbf{i})$ and variance $u(\mathbf{i})$. Since this variance measures the amount of noise, this implies that noise increases with the exposure. But the means increases faster than the noise. Indeed, the correctly scaled measurement of the noise is the Signal to Noise Ratio (SNR), which is defined by

$$SNR := \frac{\text{Mean}(u(\mathbf{i}))}{\sqrt{\text{Var}(\tilde{u}(\mathbf{i}))}} = \frac{u(\mathbf{i})}{\sqrt{u(\mathbf{i})}} = \sqrt{u(\mathbf{i})}. \tag{1.1}$$

The SNR increases like the square root of the exposure time. So the more photons we have, the better. The solution for getting a quality image, adopted from the beginning by Nicephore Niépce, was therefore to extend the exposure time as much as possible.

Yet, in a long exposure the photographed scene is exposed to variations due to changes in lighting, camera motion, and incidental motions of parts of the scene. For example in the town view of Figure 1.1, the walls on the right and left are bright because the Sun had moved during the eight hours exposure. Nowadays, digital cameras are much faster and capture fast moving objects. But even with a short exposure time, the photograph still risks motion blur on any animated scene. On the other hand, if the exposure time is too short, the image is noisy. Thus the main limitation to any imaging system is noise, regardless of its resolution.

At a first glance, the denoising problem is anyway hopeless: how to estimate the mean $u(\mathbf{i})$ of a random Poisson variable, given only one sample $\tilde{u}(\mathbf{i})$ of this variable? The best estimate of this mean knowing $\tilde{u}(\mathbf{i})$ is of course this unique sample $\tilde{u}(\mathbf{i})$. A glimpse of a solution comes from image formation theory. An optical image u is band-limited [63] and therefore smooth. Thus, one can restore the band-limited image u from its noisy version $\tilde{u}$, as was proposed in 1966 in [33], by imposing a decay to its Fourier spectrum. This classic Wiener-Fourier method multiplies the Fourier transform by optimal coefficients to attenuate the noise. It results in a convolution of the image with a "low-pass" kernel. As we shall see, this reduces the noise, but blurs the image. This is the functional perspective on the subject.

But the band-limitedness of u also implies that the random observed image values $\tilde{u}(\mathbf{j})$ at neighboring pixels $\mathbf{j}$ of a pixel $\mathbf{i}$ are positively correlated with $\tilde{u}(\mathbf{i})$. Thus, these values can be

taken into account to obtain a better estimate of $u(\mathbf{i})$. These values being nondeterministic, Bayesian approaches are relevant and have been proposed as early as 1972 [60]. This opens the stochastic perspective on the subject.

In short, there are two complementary early approaches to denoising, the Fourier-Wiener method, and the Bayesian estimation. A third hint is also given: the denoising of a given pixel value $\tilde{u}(\mathbf{i})$ must involve the values of neighboring pixels $\tilde{u}(\mathbf{j})$. This leads us to the question: where are the extra image samples that could be used to denoise the single sample $\tilde{u}(\mathbf{i})$? This question will lead us a long way. It turns out that, not only neighboring pixels in the same image can be used, but actually even pixels from other images! The mathematical innovation here comes from a non-local, or fully non-local approach to image processing, under the generic name of *neighborhood filters*, *nonlocal filters*, and even *global filters*, involving a whole set of images to denoise one.

These three main perspectives will permit us to review the main algorithmic principles which have been proposed for noise removal. All of them require a noise model, which in most of our study will be the Gaussian white noise (we will explain in the next section why this is not a limitation). The three rough denoising principles sketched above will be further combined into five algorithm classes, each one relying on a single formula.

- **The Fourier-Wiener transform thresholding principle**, section 2 : uses the regularity of the image (reflected by its sparsity in a well-chosen orthonormal transform). For the associated Fourier-Wiener image filters, the assumption is that the Fourier (or cosine transform, or wavelet transform) of the image decays quickly, and therefore faster than white noise, which is homoscedastic over all frequencies. An extreme view of this denoising principle is called "sparsity". According to this popular assumption used in *compressed sensing* [13], the ideal image has a few "sparse" coefficients in the right basis. If that is true, a simple threshold on the transform coefficients (on the right Hilbert basis) maintains the signal and kills most of the noise;
- **The self-similarity principle** and the patch based methods (section 3): The image is self-similar, and one can therefore use other "neighboring" pixels of the same image with the same expected colour to denoise a given pixel. The *neighborhood filters* propose to average the samples with similar colours, thus performing an artificial photon accumulation. This self-similarity principle is enhanced by deciding on the similarity of two pixels $\mathbf{i}$ and $\mathbf{j}$ by comparing two image patches surrounding them.
- **The Bayesian patch denoising principle,** section 4: The Bayesian principle extends the above considerations by giving them an optimal formulation, under the assumption that the patches similar to a given image patch follow a stochastic model.
- **The global denoising principle,** section 5. In this extension of the Bayesian model, not only image patches from the same image, but also image patches from *other images* can be used for image denoising. In its maximal extension, this principle can use literally all images of the world, thus giving an explicit point density function for the patch stochastic model.
- **Global neural denoising**, section 6 learns directly the denoising algorithm by a supervised learning algorithm, again learnt from a huge patch database.
- **Blind denoising**, section 7 is the ultimate achievement of the theory, as it considers denoising the image with a completely flexible noise model, learnt from the image

itself. This is the principle that must be used for old photographs and for degraded digital photographs, for which the noise model is unknown.

2. Fourier-Wiener transform thresholding

The white noise model. In this section and in the rest of the paper we shall adopt a convenient simplification of the noise model. We defined the noise as the difference between the observed image and the ideal image $\tilde{u}(\mathbf{i}) - u(\mathbf{i}) = n(\mathbf{i})$. For large enough values of $u(\mathbf{i})$ this random variable tends to be Gaussian. Furthermore, the Anscombe scalar transform $f(\tilde{u}(\mathbf{i}))$, where f is a special function proposed by Anscombe [2] transforms this Poisson noise with a variance depending on the signal $u(\mathbf{i})$ into a nearly Gaussian variable with fixed variance. By applying the Anscombe transform to the image its noise becomes *white, homoscedastic and Gaussian*. White means that the random value is independent at each pixel, which is true because the fluctuations of the photon numbers hitting each pixel are independent. Homoscedastic means that all pixels noises have the same variance which we will denote by σ^2. This noise model will simplify the discussion without loss of generality.

Classic transform thresholding algorithms use the observation that images are faithfully described by keeping only their large coefficients in a well-chosen basis. By keeping these large coefficients and setting to zero the small ones, noise should be removed and image geometry kept. By any orthogonal transform, the coefficients of an homoscedastic de-correlated noise remain de-correlated and homoscedastic. Here we refer to the classic Fourier, wavelet or cosine transforms, in their discrete version applied to the image matrix viewed as a vector in a large but finite dimension. Applied to digital images, each one of these transforms is an orthogonal transform in the finite dimensional image space. For the Fourier method this amounts to use the DFT (Discrete Fourier Transform). This Fourier method has been extended in the past thirty years to generalized linear space-frequency transforms such as the windowed cosine transform [70] or the many wavelet transforms [50].

The wavelet, or DCT, or Fourier coefficients of a Gaussian white noise with variance σ^2 remain a Gaussian diagonal vector with variance σ^2. The sparsity model assumes that the most "important" image coefficients are much larger than 3σ. Thus, cancelling the coefficients of the noisy image that are smaller (in absolute value) than, for example, 3σ will remove most of the coefficients that are only due to noise, while keeping the large image coefficients. This *sparsity* of image coefficients in certain bases is an empirical observation, used in most denoising and compression algorithms. For example the established image compression algorithms are based on the DCT (in the JPEG 1992 format) or, like the JPEG 2000 format [3], on biorthogonal wavelet transforms [17]. A bit more formally, let $\mathcal{B} = \{g_i\}_{i=1}^{M}$ be an orthonormal basis of $\mathbb{R}^M$, where M is the number of pixels of the noisy image $\tilde{u}$ (handled here as a vector). Then

$$\tilde{u} = \sum_{i=1}^{M} \langle \tilde{u}, g_i \rangle g_i, \quad \text{with} \quad \langle \tilde{u}, g_i \rangle = \langle u, g_i \rangle + \langle n, g_i \rangle \,, \tag{2.1}$$

where $\tilde{u}$, u and n denote respectively the noisy, ideal and noise images and $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product in $\mathbb{R}^M$. Being independent, the noise values $n(\mathbf{i})$ are uncorrelated. They have by assumption zero mean and variance σ^2. We can deduce that the noise coefficients in the new basis remain uncorrelated, with zero mean and variance σ^2. Indeed, denoting by $\mathbb{E}$

the expectation (with respect to the stochastic noise model) we have $\langle n, g_i \rangle = \sum_{\mathbf{r}=1}^{M} g_i(\mathbf{r}) n(\mathbf{r})$ and therefore

$$\begin{aligned}\mathbb{E}[\langle n, g_i \rangle \langle n, g_j \rangle] &= \sum_{\mathbf{r},\mathbf{s}=1}^{M} g_i(\mathbf{r}) g_j(\mathbf{s}) \mathbb{E}[n(\mathbf{r}) n(\mathbf{s})] \\ &= \langle g_i, g_j \rangle \sigma^2 = \sigma^2 \delta[j-i].\end{aligned}$$

In the Fourier-Wiener method, each noisy transform coefficient $\langle \tilde{u}, g_i \rangle$ is modified independently and then the denoised image is estimated by the inverse transform of the new coefficients. Denoting by $a(i)$ the attenuation factor $a(i)$ for the i-th coefficient, the inverse transform yields the denoised version

$$\mathbf{D}\tilde{u} = \sum_{i=1}^{M} a(i) \langle \tilde{u}, g_i \rangle g_i, \tag{2.2}$$

to be compared with (2.1). **D** is often called a *diagonal operator*. The following result, generally attributed to Norbert Wiener, gives the ideal values for $a(i)$:

Theorem 2.1. *The operator $\boldsymbol{D}_{inf}$ minimizing the mean squared error (MSE) $\boldsymbol{D}_{inf} = arg\min_{\boldsymbol{D}} \mathbb{E}\{\|u - \boldsymbol{D}\tilde{u}\|^2\}$ satisfies*

$$a(i) = \frac{|\langle u, g_i \rangle|^2}{|\langle u, g_i \rangle|^2 + \sigma^2}. \tag{2.3}$$

The previous optimal operator attenuates all noisy coefficients. In the methods assuming a "sparsity" for the ideal image u, one further restricts $a(i)$ to be 0 or 1. Then the diagonal operator becomes a projection operator. In that case, a subset of coefficients is kept, and the rest are set to zero. The projection operator that minimizes the MSE under that constraint is obtained with

$$a(i) = \begin{cases} 1 & \text{if } |\langle u, g_i \rangle|^2 \geq \sigma^2, \\ 0 & \text{otherwise.} \end{cases}$$

A *transform thresholding* algorithm therefore keeps the coefficients with a magnitude larger than the noise, while setting to zero the rest. Note that both above mentioned filters are "ideal", or "oracular" operators. Indeed, they use the coefficients $\langle u, g_i \rangle$ of the original image, which are not known. For this reason, such algorithms are called *oracle filters*. The classical *transform threshold filters* must approximate the oracle coefficients by using the observable noisy coefficients. The real denoising method is therefore called *empirical Wiener filter*, because it approximates the unknown original coefficients $\langle u, g_i \rangle$ by invoking the identity

$$\mathbb{E}|\langle \tilde{u}, g_i \rangle|^2 = |\langle u, g_i \rangle|^2 + \sigma^2$$

to replace the optimal attenuation coefficients $a(i)$ by the empirical attenuation coefficients

$$\alpha(i) = \max\left\{0, \frac{|\langle \tilde{u}, g_i \rangle|^2 - c\sigma^2}{|\langle \tilde{u}, g_i \rangle|^2}\right\} \tag{2.4}$$

where c is a parameter, usually larger than one.

The global Fourier basis is not used for denoising. Indeed, modifying Fourier coefficients by the diagonal operator often causes undue oscillation. To avoid this effect, the orthogonal bases are usually more local, of the wavelet or block DCT type. We give now two examples.

The sliding window DCT. The local adaptive filters were introduced by Yaroslavsky and Eden [70] and Yaroslavsky [72]. The noisy image is analyzed in a moving square block, typically with dimensions 8×8. At each position of the block center, its DCT spectrum is computed and modified by using the empirical coefficients (2.4). Finally, an inverse transform is used to estimate only the signal value in the central pixel of the block.

Wavelet thresholding. Let $\mathcal{B} = \{g_i\}_i$ be a wavelet orthonormal basis [49]. The so-called *hard wavelet thresholding method* [26] is a (nonlinear) projection operator setting to zero all wavelet coefficients smaller than a certain threshold. The performance of the method depends on the ability of the basis to approximate the image U by a small set of large coefficients. There has been a strenuous search for wavelet bases adapted to images [52].

Unfortunately, the brutal cancelation of DCT coefficients near the image edges[1] creates small oscillations by the Gibbs phenomenon. Similarly, the undue cancelation of some of the small wavelet coefficients may also cause the appearance of isolated wavelets in flat image regions. These annoying artifacts are sometimes called *wavelet outliers* [27]. They can be partially avoided with the use of a soft thresholding [25],

$$\alpha(i) = \begin{cases} \frac{\langle \tilde{u}, g_i \rangle - sgn(\langle \tilde{u}, g_i \rangle)\mu}{\langle \tilde{u}, g_i \rangle}, & \text{if } |\langle \tilde{u}, g_i \rangle| \geq \mu, \\ 0 \quad \text{otherwise}, \end{cases}$$

which reduces the Gibbs oscillation near image discontinuities.

Several carefully designed orthogonal bases adapt better to image local geometry and discontinuities than wavelets, particularly the "bandlets" [52] and "curvelets" [65]. This tendency to adapt the transform locally to the image is accentuated with the methods adapting a different basis to each pixel, or selecting a few elements or "atoms" from a huge patch dictionary to linearly decompose the local patch on these atoms. This point of view is developed in *sparse coding methods* and the K-SVD algorithm [1, 29, 47].

2.1. A case study: DCT denoising. We shall illustrate transform thresholding by at least one good detailed example. A basic DCT denoising can be drastically improved by several ingredients illustrated in Figure 2.1. This figure shows how the result improves by successively using a better colour space[2], by aggregating [18] the 64 denoised values obtained for each pixel, which is contained in 64 patches with 8×8 dimensions, by making a statistically more correct aggregation of these estimates, and finally by iterating the method, using the first denoised image as "oracle" for applying the Wiener filter a second time. The method is summarized in Algorithm 1. See [32] for an online implementation.

3. The self-similarity principle and the patch based methods

If m noisy independent pixels with the same expected colour are averaged, the noise (namely the variance of the average of these m values) is divided by m. The first application of this

[1] So are called the strong image discontinuities along apparent contours of visible objects.

[2] A colour image is a set of three images (R, G, B) giving scalar values to three chromatic components, Red, Green, Blue. The linear transform improving the denoising performance is simply $Y_0 = (R + G + B)/3, U_0 = \frac{1}{2}(R - B), V_0 = \frac{1}{4}(R + B) - \frac{1}{2}G$, where Y_0 is the luminance, and U_0 and V_0 contain the colour contrast between green and blue and green and red respectively.

Algorithm 1 DCT denoising algorithm. DCT coefficients lower than 3σ are canceled in the first step and a Wiener filter is applied in the "oracle" second step. In colour this strategy is applied to Y_0 . Its attenuation coefficients are also applied to U_o, V_o.

Input: noisy image $\tilde{u}$, σ noise standard deviation, (optional) prefiltered image $\hat{u}_1$ for "oracle" estimation, $h = 3\,\sigma$: threshold parameter.
Output: output denoised image u.
for each patch $\tilde{P}$ of size 8×8 (if $\hat{u}_1$, patch P_1 in $\hat{u}_1$) **do**
  Compute the DCT transform of $\tilde{P}$ (if $\hat{u}_1$, of P_1).
  if $\hat{u}_1$ **then**
    Modify DCT coefficients of $\tilde{P}$ as $\tilde{P}(i) = \tilde{P}(i)\frac{P_1(i)^2}{P_1(i)^2+\sigma^2}$.
  else
    Cancel coefficients of $\tilde{P}$ with magnitude lower than h.
  end if
  Compute the inverse DCT transform obtaining $\hat{P}$.
  Compute the aggregation weight $w_{\tilde{P}} = 1/\#\{\text{number of non-zero } DCT \text{ coefficients}\}$.
end for
for each pixel **i** **do**
  Aggregation: recover the denoised value at each pixel **i** by averaging all values at **i** of all denoised patches $\hat{Q}$ containing **i**, weighted by $w_{\tilde{Q}}$.
end for

very simple denoising principle is the use of accumulation: when the camera and the scene do not move, the larger the photon count, the larger the signal (mean) to noise (standard deviation) ratio. When we only dispose of a single image, some succedaneous of the above averaging principle must be found to compensate for the limited amount of observed photons. A rather trivial idea is to average the closest pixels to a given pixel. This amounts to convolve the image with a fixed radial positive kernel, for example a Gaussian kernel. This approach works only for pixels inside the homogeneous image regions, but not for those in contrasted image regions. A convolution with a Gaussian may reduce the noise, but it makes the image blurry.

Averaging pixels with similar colours. The sigma-filter [43] or neighborhood filter [71] is an elegant solution to avoid this blur risk. Neighborhood filters average nearby pixels of **i**, but under the condition that they have a colour value similar to that of **i**. These filters denoted by NF for neighborhood filter are defined by

$$NF_{h,\rho}\tilde{u}(\mathbf{i}) = \frac{1}{C(\mathbf{i})} \sum_{\mathbf{j}\in B_\rho(\mathbf{i})} \tilde{u}(\mathbf{j})\, e^{-\frac{|\tilde{u}(\mathbf{i})-\tilde{u}(\mathbf{j})|^2}{h^2}}, \tag{3.1}$$

where $B_\rho(\mathbf{i})$ is a ball of center **i** and radius $\rho > 0$, $h > 0$ is the filtering parameter and $C(\mathbf{i}) = \sum_{\mathbf{j}\in B_\rho(\mathbf{i})} e^{-\frac{|\tilde{u}(\mathbf{j})-\tilde{u}(\mathbf{i})|^2}{h^2}}$ is the normalization factor to make the above an averaging filter. The parameter h expresses the required degree of colour similarity between **i** and **j**. The filter (3.1) is so powerful that it has been been reinvented several times and received several names: *σ-filter* [43], *SUSAN filter* [64] and *Bilateral filter* [66].

3.1. Non-local means. The Non-local means filter extends the concept of a neighborhood filter by implicitly assuming a Markov field structure for the image. Its idea stems from the now famous algorithm to synthesize textures from examples [28]. Its Markovian assumption

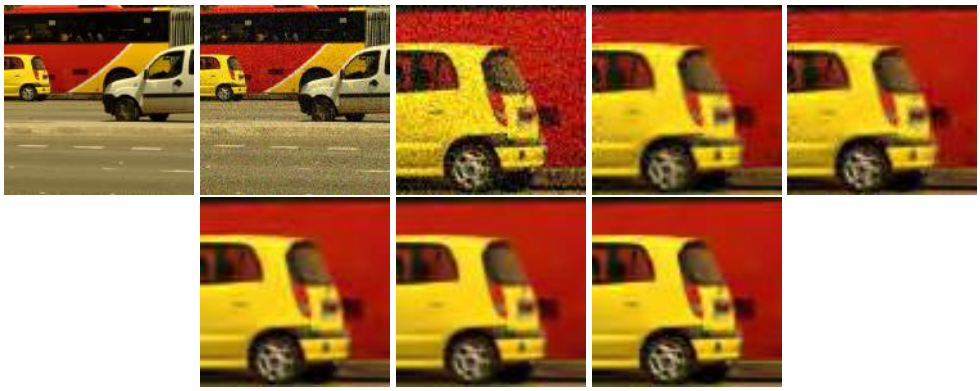


Figure 2.1. Original and noisy images with additive Gaussian white noise; crops of denoised images by Algorithm 1 when incrementally adding the use of a $Y_oU_oV_o$ colour system, uniform aggregation of the 64 estimated values at each pixel, statistically optimal aggregation of the same estimates, and iteration of the Wiener filter with the "oracle" given by the first step. Image quality and SNR increase significantly at each step.

is that, in a textured image, the stochastic model for a given pixel **i** can be predicted from a local image neighborhood P of **i**, which we shall call "patch".

The assumption for recreating new textures from samples is that there are enough pixels **j** similar to **i** in a texture image $\tilde{u}$ to recreate a new but similar texture u. This algorithm goes back to Shannon's theory of communication [63], where it was used for the first time to synthesize a probabilistically correct text from a sample.

An adaptation of the above synthesis principle yields an image denoising algorithm [7][3]. The observed image is the noisy image $\tilde{u}$. The reconstructed image is the denoised image $\hat{u}$. A noisy patch $\tilde{P}$ surrounding a pixel **i** is restored by looking for the patches $\tilde{Q}$ in $\tilde{u}$ with the same dimensions as $\tilde{P}$ and resembling P. Then the restored value $\hat{u}(\mathbf{i})$ is a weighted average of the central values $\tilde{u}(\mathbf{j})$ of the patches resembling P. This defines the"non-local means" algorithm, called "non-local" because it uses patches $\tilde{Q}$ that can be lie far away from $\tilde{P}$, and *even patches taken from other images*.

The underlying self-similarity hypothesis is that for every small patch in a natural image one can find several similar patches in the same image, as illustrated in figure 3.1. Let us now give the formula. NL-means denoises a square reference patch $\tilde{P}$ around **i** of dimension $\kappa \times \kappa$ by replacing it by an average of all similar patches $\tilde{Q}$ in a square neighborhood of **i** of size $\lambda \times \lambda$. To do this, a normalized Euclidean distance between $\tilde{P}$ and $\tilde{Q}$, $d(\tilde{P},\tilde{Q}) = \frac{1}{\kappa^2}\|\tilde{P}-\tilde{Q}\|^2$ is computed for all patches $\tilde{Q}$ is the search neighborhood. Then the weighted average is

$$\hat{P} = \frac{\sum_{\tilde{Q}} \tilde{Q} e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}}{\sum_{\tilde{Q}} e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}}. \tag{3.2}$$

The whole method is given in Algorithm 2 and can be tested in IPOL [10].

NL-means works better than the neighborhood filters because the distances of colours between pixels are computed on a patch surrounding the pixel instead of just the central pixel.

[3]See also the related attempts [4, 23, 51, 69].

Algorithm 2 NL-means algorithm.

Input: noisy image $\tilde{u}$, σ noise standard deviation. **Output**: denoised image $\hat{u}$.
Parameters: $\kappa = 3$: patch size, $\lambda = 31$: size of search zone for similar patches, $h = 0.6\,\sigma$: filtering parameter (these values may depend on the noise level)
for each pixel **i** **do**
  Select a square reference patch $\tilde{P}$ around **i** of dimension $\kappa \times \kappa$. Set $\hat{P} = 0$ and $\hat{C} = 0$.
  for each patch $\tilde{Q}$ in a square neighborhood of **i** of size $\lambda \times \lambda$ **do**
    Compute the normalized Euclidean distance $d(\tilde{P}, \tilde{Q}) = \frac{1}{\kappa^2}\|\tilde{P} - \tilde{Q}\|^2$.
    Accumulate $\tilde{Q}e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}$ to $\hat{P}$ and $e^{-\frac{d(\tilde{P},\tilde{Q})^2}{h^2}}$ to $\hat{C}$.
  end for
  Normalize the average patch $\hat{P}$ by dividing it by the sum of weights $\hat{C}$.
end for
for each pixel x **do**
  Aggregation: recover the denoised value at each pixel **i** by averaging all values at **i** of all denoised patches $\hat{Q}$ containing **i**.
end for

Thus only values of really similar pixels are averaged. This progress is illustrated in Figure 3.2 where the pixel "neighborhoods" have an increasing sophistication: the first result, on an original scanned image, is obtained by a Gaussian convolution. Efficient in flat regions, this filter blurs the edges. The second result is obtained by Yaroslavsky's neighborhood filter: each pixel is replaced by an average of the pixels which are close to it in both the image domain and colour range. The result is much sharper. The last result is obtained by NL-means. The choice of resembling pixels is still more selective. The image differences between original and denoised demonstrate the progress. This difference looks increasingly like noise when the pixel neighborhood becomes more sophisticated. The underlying self-similarity assumption can be formalized by an ergodic assumption, under which NL-means can be proved to converge asymptotically to the noiseless image [4]. The more samples the better, so the algorithm is immediately extendable to video [9]. Figure 3.1 illustrates how NL-means chooses the right weight configuration for each sort of image self-similarity.

4. The Bayesian patch denoising principle

Given u the noiseless ideal image and $\tilde{u}$ the noisy image corrupted with Gaussian noise of standard deviation σ so that

$$\tilde{u} = u + n, \tag{4.1}$$

the conditional distribution $\mathbb{P}(\tilde{u} \mid u)$ is

$$\mathbb{P}(\tilde{u} \mid u) = \frac{1}{(2\pi\sigma^2)^{\frac{M}{2}}} e^{-\frac{||u-\tilde{u}||^2}{2\sigma^2}}, \tag{4.2}$$

where M is the total number of pixels in the image. In order to compute the probability of

[4] It can be proved [7] that if the image is a fairly general stationary and mixing random process, for every pixel **i**, NL-means converges to the conditional expectation of **i** knowing its neighborhood, which is the best Bayesian estimate.

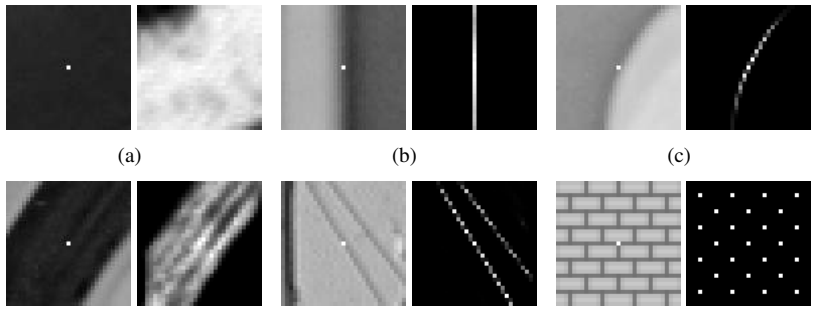


Figure 3.1. On the right-hand side of each pair, one can see the weights in the NL-means average used to estimate a 3×3 patch located in the center of the left image by NL-means.

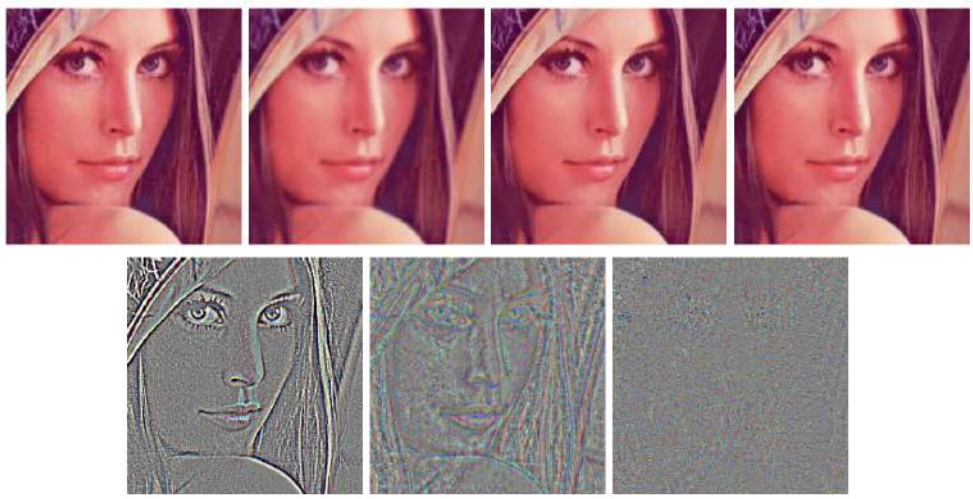


Figure 3.2. A comparison of the efficiency of neighborhood filters. The first row shows a piece of a famous test image (Lena) followed by its denoised version by a Gaussian convolution, a neighborhood filter, and NL-means. The second row shows the difference between the image and its denoised version, which increasingly resembles white noise.

the original image given the degraded one, $\mathbb{P}(u \mid \tilde{u})$, we need a prior on u. In the first models [30], this prior was a parametric Markov random field, specified by its Gibbs distribution. A Gibbs distribution for an image u takes the form

$$\mathbb{P}(u) = \frac{1}{Z} e^{-E(u)/T},$$

where Z and T are constants and E is called the energy function and writes

$$E(u) = \sum_{C \in \mathcal{C}} V_C(u),$$

where $\mathcal{C}$ denotes the set of cliques associated to the image and V_C is a potential function. The maximization of the *a posteriori* distribution writes by Bayes formula

$$\operatorname*{Arg\,max}_u \mathbb{P}(u \mid \tilde{u}) = \operatorname*{Arg\,max}_u \mathbb{P}(\tilde{u} \mid u)\mathbb{P}(u),$$

which is equivalent to the minimization of $-\log \mathbb{P}(u \mid \tilde{u})$,

$$\operatorname*{Arg\,min}_u \|u - \tilde{u}\|^2 + \frac{2\sigma^2}{T} E(u).$$

This energy writes as a sum of local derivatives of pixels in the image, thus being equivalent to a classical Tikhonoff regularization, [30], [6].

Recent Bayesian methods have abandoned as too simplistic the global patch models formulated by a parametric Gibbs energy. Instead, the methods build local non parametric patch models learnt from the image itself, usually as a local Gaussian model around each given patch, or as a Gaussian mixture. The term "patch model" is now preferred to the notion of "clique" previously used for the Markov field methods. But the underlying notion is the same: a "patch" is nothing but a clique. The difference is that the patch model is local and empirical while the clique probability model was usually global and parametric. In the nonparametric local patch models, the patches can become larger, up to an 8×8 size, while the cliques were often confined to very small neighborhoods. Given a noiseless patch P of u with dimension $\kappa \times \kappa$, and $\tilde{P}$ an observed noisy version of P, the same model gives by the independence of noise pixel values

$$\mathbb{P}(\tilde{P}|P) = c \cdot e^{-\frac{\|\tilde{P}-P\|^2}{2\sigma^2}} \tag{4.3}$$

where P and $\tilde{P}$ are considered as vectors with κ^2 components $||P||$ denotes the Euclidean norm of P, and c is the normalizing constant. Knowing $\tilde{P}$, our goal is to deduce P by maximizing $\mathbb{P}(P|\tilde{P})$. Using Bayes' rule, we can compute this last conditional probability as

$$\mathbb{P}(P|\tilde{P}) = \frac{\mathbb{P}(\tilde{P}|P)\mathbb{P}(P)}{\mathbb{P}(\tilde{P})}. \tag{4.4}$$

$\tilde{P}$ being observed, this formula can in principle be used to deduce the patch P maximizing the right term, viewed as a function of P. This is only possible if we know the probability model $\mathbb{P}(P)$. This model will be learnt from the image itself, or from a set of images[5]. For example, once we have obtained (like with NL-means) a group of similar patches Q similar to a given noisy patch P, these patches can be treated as a set of samples of a Gaussian vector. This permits to denoise each observed patch by a Bayesian estimation under this Gaussian model [38]. Let us assume that the patches Q similar to P follow a Gaussian model with (observable, empirical) covariance matrix $\mathbf{C}_P$ and (observable, empirical) mean $\overline{P}$. This means that

$$\mathbb{P}(Q) = c.e^{-\frac{(Q-\overline{P})^t \mathbf{C}_P^{-1}(Q-\overline{P})}{2}} \tag{4.5}$$

From (4.2) and (4.4) we obtain for each observed $\tilde{P}$ the following equivalence of problems:

$$\max_P \mathbb{P}(P|\tilde{P}) \quad \Leftrightarrow \quad \max_P \mathbb{P}(\tilde{P}|P)\mathbb{P}(P)$$

[5]For example [15], [68] or [75] apply a clustering method to the set of patches of a given image before restoration, and [77] applies it to a huge set of patches extracted from many images.

$$\Leftrightarrow \quad \max_P e^{-\frac{\|P-\tilde{P}\|^2}{2\sigma^2}} e^{-\frac{(P-\overline{P})^t \mathbf{C}_{\overline{P}}^{-1}(P-\overline{P})}{2}}$$

$$\Leftrightarrow \quad \min_P \frac{\|P-\tilde{P}\|^2}{\sigma^2} + (P-\overline{P})^t \mathbf{C}_P^{-1}(P-\overline{P}).$$

This expression does not yield an algorithm. Indeed, the noiseless patch P and the patches similar to P are not observable. So we face the same problem as with the oracular Fourier-Wiener filter. Nevertheless, we dispose of the noisy version $\tilde{P}$ and can compute the patches $\tilde{Q}$ similar to $\tilde{P}$. An empirical covariance matrix can therefore be obtained for the patches $\tilde{Q}$ similar to $\tilde{P}$. Furthermore, using (4.1) and the fact that P and the noise n are independent, it is easily checked that

$$\mathbf{C}_{\tilde{P}} = \mathbf{C}_P + \sigma^2 \mathbf{I}; \qquad E\tilde{Q} = \overline{P}. \tag{4.6}$$

If the above empirical estimates are reliable, the maximum *a posteriori* estimation problem finally boils down by (4.6) to the minimization problem:

$$\max_P \mathbb{P}(P|\tilde{P}) \Leftrightarrow \min_P \frac{\|P-\tilde{P}\|^2}{\sigma^2} + (P-\overline{\tilde{P}})^t (\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I})^{-1}(P-\overline{\tilde{P}}).$$

Differentiating this quadratic function with respect to P and equating to zero yields the amazingly simple denoising formula

$$\hat{P}_1 = \overline{\tilde{P}} + \left[\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I}\right] \mathbf{C}_{\tilde{P}}^{-1}(\tilde{P} - \overline{\tilde{P}}). \tag{4.7}$$

The formula (4.7) gives a direct denoising algorithm, provided we can compute the patch expectations and patch covariance matrices. This is done in [38] by computing empirical means and covariances from the patches similar to a given noisy patch. Since the first such estimate is not accurate, it is natural to iterate the algorithm, so that means and covariances are computed again from denoised patches at the first step. Thus, Algorithm 3 is a self-explanatory application of the single formula (4.7).

As pointed out in [41], the above Nonlocal Bayes algorithm is a Bayesian interpretation (with some generic improvements like the aggregation) of the PCA based algorithm proposed in [76][6].

5. The global patch denoising principle

The most recent denoising methods tend to give up any image model. Indeed, they directly use the observed set of images to denoise a new one. More specifically they denoise image patches by a fully non-local algorithm, in which the patch is compared to a patch model obtained from a large or *very large* patch set, of up to 10^{10} patches. Each patch is denoised by deducing its likeliest estimate from the set of all patches. In the method proposed in [77], this patch space is organized as a Gaussian mixture with about 200 components[7].

[6] See also [24] for a comparison of several local and more global strategies. Non Gaussian, Bayesian models are possible, depending on the patch and noise models. For example [59] treats the case of a local exponential density model for the noisy data.

[7] A similar idea was used in [34] who claim performing a "Scene completion using millions of photographs" to fill in missing parts of a given image.

Algorithm 3 Non local Bayes image denoising

Input: noisy image $\tilde{u}$, σ noise standard deviation. **Output**: denoised image $\hat{u}$.

for all patches $\tilde{P}$ of the noisy image **do**

Find a set $\mathcal{P}(\tilde{P})$ of patches $\tilde{Q}$ similar to $\tilde{P}$.

Compute the expectation $\overline{\tilde{P}}$ and covariance matrix $\mathbf{C}_{\tilde{P}}$ of these patches by

$$\mathbf{C}_{\tilde{P}} \simeq \frac{1}{\#\mathcal{P}(\tilde{P})-1} \sum_{\tilde{Q}\in\mathcal{P}(\tilde{P})} \Big(\tilde{Q}-\overline{\tilde{P}}\Big)\Big(\tilde{Q}-\overline{\tilde{P}}\Big)^t, \qquad \overline{\tilde{P}} \simeq \frac{1}{\#\mathcal{P}(\tilde{P})} \sum_{\tilde{Q}\in\mathcal{P}(\tilde{P})} \tilde{Q}.$$

Obtain the first step estimation $\hat{P}_1 = \overline{\tilde{P}} + \left[\mathbf{C}_{\tilde{P}} - \sigma^2\mathbf{I}\right]\mathbf{C}_{\tilde{P}}^{-1}(\tilde{P}-\overline{\tilde{P}})$.

end for

Obtain the pixel value of the basic estimate image $\hat{u}_1$ as an average of all values of all denoised patches $\hat{Q}_1$ which contain $\mathbf{i}$.

for all patches $\tilde{P}$ of the noisy image **do**

Find a new set $\mathcal{P}_1(\tilde{P})$ of noisy patches $\tilde{Q}$ similar to $\tilde{P}$ by comparing their denoised "oracular" versions Q_1 to P_1.

Compute the new expectation $\overline{\tilde{P}}^1$ and covariance matrix $\mathbf{C}_{\hat{P}_1}$ of these patches:

$$\mathbf{C}_{\hat{P}_1} \simeq \frac{1}{\#\mathcal{P}(\hat{P}_1)-1} \sum_{\hat{Q}_1\in\mathcal{P}(\hat{P}_1)} \Big(\hat{Q}_1-\overline{\tilde{P}}^1\Big)\Big(\hat{Q}_1-\overline{\tilde{P}}^1\Big)^t, \qquad \overline{\tilde{P}}^1 \simeq \frac{1}{\#\mathcal{P}(\hat{P}_1)} \sum_{\hat{Q}_1\in\mathcal{P}(\hat{P}_1)} \tilde{Q}.$$

Obtain the second step patch estimate $\hat{P}_2 = \overline{\tilde{P}}^1 + \mathbf{C}_{\hat{P}_1}\left[\mathbf{C}_{\hat{P}_1} + \sigma^2\mathbf{I}\right]^{-1}(\tilde{P}-\overline{\tilde{P}}^1)$.

end for

Obtain the pixel value of the denoised image $\hat{u}(\mathbf{i})$ as an average of all values of all denoised patches $\hat{Q}_2$ which contain $\mathbf{i}$.

In image denoising, the same idea [45] leads to define the simplest universal method, where a huge set of patches is used to estimate the upper limits a patch-based denoising method will ever reach[8]. The preliminary experiments of this paper involved a set of 20, 000 images [62]. The method, even if certainly not practical, is of exquisite simplicity. Given a clean patch P the noisy patch $\tilde{P}$ with Gaussian noise of standard deviation σ has probability distribution

$$\mathbb{P}(\tilde{P} \mid P) = \frac{1}{(2\pi\sigma^2)^{\frac{\kappa^2}{2}}} e^{-\frac{||P-\tilde{P}||^2}{2\sigma^2}}, \tag{5.1}$$

where κ^2 is the number of pixels in the patch. Then given a noisy patch $\tilde{P}$ its optimal estimator for the Bayesian minimum squared error (MMSE) is by Bayes' formula

$$\hat{P} = \mathbb{E}[P \mid \tilde{P}] = \int \mathbb{P}(P \mid \tilde{P}) P dP = \int \frac{\mathbb{P}(\tilde{P} \mid P)}{\mathbb{P}(\tilde{P})} \mathbb{P}(P) P dP. \tag{5.2}$$

Using a huge set of M natural patches (with a distribution supposedly approximating the real natural patch density), we can approximate the terms in (5.2) by $\mathbb{P}(P)dP \simeq \frac{1}{M}$ and

[8]The results of this paper support the "near optimality of state of the art denoising results", the results obtained by the classic state of the art BM3D algorithm being only 0.1 decibel away from optimality for methods using small patches (typically 8×8.) See also [14].

$\mathbb{P}(\tilde{P}) \simeq \frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i)$, which in view of (5.1) yields

$$\hat{P} \simeq \frac{\frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i) P_i}{\frac{1}{M}\sum_i \mathbb{P}(\tilde{P} \mid P_i)}. \tag{5.3}$$

Thus the final MMSE estimator is nothing but the exact application of NL-means, denoising each patch by matching it to the huge patch database. The final algorithm is summarized in Algorithm 4. Although this algorithm is optimal, it is not yet fully realizable in our current technology[9].

Algorithm 4 Global Bayesian denoising

Inputs: Noisy image $\tilde{u}$ in vectorial form; very large set of M patches P_i extracted from a large set of noiseless natural images. **Output**: Denoised image $\hat{u}$.
for all patches $\tilde{P}$ extracted from $\tilde{u}$ **do**
Compute the MMSE denoised estimate of $\tilde{P}$

$$\hat{P} \simeq \frac{\sum_{i=1}^{M} \mathbb{P}(\tilde{P} \mid P_i) P_i}{\sum_{i=1}^{M} \mathbb{P}(\tilde{P} \mid P_i)}$$

where $\mathbb{P}(\tilde{P} \mid P_i)$ is known from (5.1).
end for
At each pixel **i** get $\hat{u}(\mathbf{i})$ as $\hat{P}(\mathbf{i})$, where the patch P is centered at **i**.
(optional Aggregation) : for each pixel **j** of u, compute the denoised version $\hat{u}_{\mathbf{j}}$ as the average of all values $\hat{P}(\mathbf{j})$ for all patches containing **j**. (This step in not considered in [45].)

5.1. Comparing visual quality. The visual quality of the restored image is obviously a necessary, if not sufficient, criterion to judge the performance of a denoising algorithm. It permits to control the absence of artifacts and the correct reconstruction of edges, texture and fine structure. Figure 5.1 displays the noisy and denoised images for several classic algorithms for noise standard deviations of 30 (where each colour image is on a scale from 0 to 255). The experiment illustrates that algorithms based on wavelets or DCT, like DCT and BLS-GSM, suffer of a strong Gibbs effect near all image edges. This Gibbs effect is nearly not noticeable in the denoised image by K-SVD which uses a transform method in a learned redundant patch basis, or patch dictionary. The NL-means denoised image has no visual artifacts but is more blurred than those given by BM3D and Non-Local Bayes, that have a clearly superior performance to the rest of the algorithms. The BM3D denoised image has some Gibbs effect near edges, which sometimes degrades the visual quality of the solution. Indeed, the BM3D method is a syncretic method combining the grouping of similar patches with a DCT transform thresholding.

In short, the visual quality of DCT, BLS-GSM and K-SVD is inferior to that of NL-means, BM3D and NL-Bayes, because of strong colour noise low frequencies in flat zones, and of a Gibbs effect.

[9] A clever change of variables in the integral (5.2) found in [53] permits to accelerate the calculation in (5.3) by a 1000 factor, but this is still insufficient!

Figure 5.1. Comparison of visual quality. The noisy image was obtained adding a Gaussian white noise of standard deviation 30. From top to bottom and left to right: original, noisy, DCT sliding window, BLS-GSM, NL-means, K-SVD, BM3D, and Non-local Bayes.

6. Global neural denoising

Though optimal in theory, the global Bayesian denoising formula (5.2) has been recently very well approximated by a neural network learning from an equally huge set of image patches. A feed-forward neural network is a succession of non-linear hidden layers followed by an application-dependent decoder

$$f(\cdot,\theta) = \mathfrak{d} \circ h_n \circ \cdots \circ h_1(\cdot), \quad n \geq 1$$

with

$$\forall 1 \leq l \leq n,\ h_l(z_l) = \mathfrak{a}(W_l z_l + b_l)$$

and

$$\mathfrak{d}(z_n) = W_{n+1} z_{n+1} + b_{n+1}$$

in case of a linear decoder. The parameters θ comprise the connection weights W_l and biases b_l. The activation function $\mathfrak{a}(\cdot)$, typically implemented with the hyperbolic tangent or the logistic function, is applied to its input vector element-wise.

Besides being infinitely differentiable, neural networks can approximate arbitrarily well any continuous function on a compact set [35, 44], thereby making them a candidate for regression tasks

$$\begin{aligned} \theta^* &= \operatorname*{Arg\,min}_{\theta} \mathbb{E}\|f(\tilde{x},\theta) - x\|_2^2 \\ &= \operatorname*{Arg\,min}_{\theta} \mathbb{E}\|f(\tilde{x},\theta) - \mathbb{E}[x|\tilde{x}]\|_2^2 \end{aligned} \tag{6.1}$$

where $(\tilde{x}, x)$ denotes a random pair of observation and its ideal prediction, whose joint behavior is governed by some probability law used to define the expectation in (6.1). Note

	BM3D	NL-Bayes	DNN
$\sigma = 25$	32.53	32.61	**32.88**
$\sigma = 50$	29.20	29.34	**29.72**
$\sigma = 75$	27.28	27.22	**27.95**
$\sigma = 170$	13.84	22.99	**24.56**

Table 6.1. Table comparing two state of the art denoising methods with DNN: the PSNR, qui is a logarithm of the SNR defined in (1.1) measures the image quality (the higher the better).

that although we can sample from it, the underlying probability does not have a closed form in general. Moreover, the function $\theta \mapsto f(\tilde{x}, \theta)$ is not convex, leaving us with little choice but to substitute the expectation with an empirical surrogate and rely on the method of steepest descent [5, 42] to conduct the minimization.

Recently, a set of image denoising neural networks [11] has been shown to outperform BM3D [22] and non-local Bayes [38] at several rather high levels of Gaussian noise for which they were trained. Note that these spin-offs of the original non-local means [7] seek information exclusively inside the noisy image while the neural networks learned to estimate the 17-by-17 patch lying at the center of a noisy 39-by-39 noisy observation by looking at noisy and clean patch pairs gathered from other many images. Table 6.1 is a comparison of these algorithms on a benchmark set and the deep neural networks (DNN) consistently dominate the other two for all the four noise levels.

A look at the output layer of the neural network trained at $\sigma = 25$ (Figure 6.2) reveals a locally oscillating behaviour akin to that of wavelets for those visually meaningful synthesis features. This suggests that a sort of optimal Fourier-Wiener filter is being performed.

This impressive performance is reached with neural networks of four hidden layers, each carrying up to 3000 nodes, thereby requiring a computational cost of several 10^6 operations per pixel. Moreover, their enormous sizes also mean long training time: it could take weeks on a modern GPU platform to train just one neural network [12] under a specific level of noise with tens of millions of example pairs. Although through an investigation of the natural patch distribution, it can be shown [67] that a simple linear transform is readily available to make a single neural network work well across all levels of Gaussian noise, the challenge lying ahead is to scale down such a neural network while preserving its performance.

7. Blind denoising

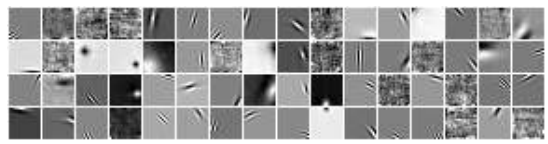
We have shown that all efficient denoising methods boil down to a single formula and to very simple image models. But we assumed a simple noise model, the Gaussian white noise. In this section the focus will be on performing "blind denoising", namely a fully automatic denoising on any digital image.

In most images handled by the public and even by scientists, the noise model is indeed imperfectly known or unknown. Recent progress in noise estimation permits to estimate from a single image a noise model which is simultaneously signal and frequency dependent. We describe here a multiscale denoising algorithm [39] adapted to this broad noise model. This leads to a blind denoising algorithm which can be tested for example on scans of old

(a) Original (b) BM3D

(c) Non local Bayes (d) Neural Network

Figure 6.1. (b), (c) and (d) are the denoised versions of the original image corrupted by some Gaussian noise with standard deviation at 25. The figure shows that a blindly learned algorithm by neural network can outperform all carefully hand-crafted algorithms. Nevertheless the resulting neural network is still unpractical, necessitating tens of millions of connections to denoise a single patch.



(a) output features

Figure 6.2. A random selection of output features. Most of the output features resemble classic wavelets.

photographs, for which the noise model is unknown.

Blind denoising is the conjunction of a noise estimation method followed by the application of an adapted denoising method. Yet, to cope with the broad variety of observed noises, the noise model must be far more comprehensive than the usual white Gaussian noise. Because images have undergone nonlinear operations and filters, a flexible denoising method must cope with a noise model that depends on the signal, but also on the spatial frequency (in technical terms, a coloured noise). The archives of the online executions at the IPOL journal of seven classic denoising methods, namely DCT denoising [72–74], TV denoising [31, 61], K-SVD [40, 48], NL-means [8, 10], BM3D [21, 37], BLS-GSM [58] and NL-Bayes [41] are full with puzzling noisy images.

There are only a few references on blind denoising approaches: Portilla [56], [55], Rabie [57] and Liu, Freeman, Szeliski and Kang [46]. Portilla's method is an adaptation of the BLS-GSM algorithm, which models wavelet patches at each scale by a Gaussian scale mixture (GSM), followed by a Bayesian least square (BLS) estimation for wavelet patches. The "noise clinic" described in this section is based on a noise signal and frequency noise estimator proposed by Colom et al. [19, 20], relying on a Ponomarenko et al. general principle [54] to build a noise patch model.

As evident in its formula (4.7), the NL-Bayes method described in section 4 only requires the knowledge of a local Gaussian patch model and of a Gaussian noise model. We already saw in Algorithm 3 how to estimate the local patch Gaussian model, described by an empirical mean and an empirical covariance. So we only need to hint at how to estimate the covariance matrix of the noise. The noise model being signal dependent, for each intensity $\mathbf{i}$ in the range intensity $[0, 255]$ of the image a noise covariance matrix $\mathbf{C}_{n\mathbf{i}}$ must be estimated. The noise model for each group of patches similar to $\tilde{P}$ will depend on $\tilde{P}$ through their mean $\mathbf{i}$. The reference intensity for the current 3D group $\mathcal{P}(\tilde{P})$ must therefore be estimated to apply (4.7) with the appropriate noise covariance matrix. This intensity is simply estimated as the average of all pixels contained in $\mathcal{P}(\tilde{P})$. So we need to estimate the noise covariance matrices $\{\mathbf{C}_{n\mathbf{i}}\}_{\mathbf{i}\in[0,255]}$. Colom et al., [20], proposed an adaptation of the Ponomarenko et al. [54] method estimating a frequency dependent noise to estimate noise in JPEG images. Given a patch size $\kappa \times \kappa$, the method extracts from the image a set with fixed cardinality of sample blocks with lowest variance, and with mean approximately equal to $\mathbf{i}$. These blocks are therefore likely to contain only noise. They are transformed by a DCT, and an empirical standard deviation of their DCT coefficients is computed. This algorithm computes for every intensity $\mathbf{i}$ with a multi-frequency noise estimate given by a $\kappa^2 \times \kappa^2$ matrix

$$\mathbf{M}_{\mathbf{i}} := \mathbb{E}\left(\mathcal{D}N_{\mathbf{i}}\left(\mathcal{D}N_{\mathbf{i}}\right)^t\right) \tag{7.1}$$

where $\mathcal{D}$ is the $\kappa^2 \times \kappa^2$ matrix of the discrete cosine transform (DCT) and $N_{\mathbf{i}}$ denotes the $\kappa \times \kappa$ stochastic noise patch model at intensity $\mathbf{i}$. This method estimates the variances of the DCT coefficients of noise blocks and not their covariances. The covariance matrices are assumed to be diagonal, since generally the DCT decorrelates the noise.

For a given intensity $\mathbf{i}$, the covariance matrix of the noise is $\mathrm{Cov}(N_{\mathbf{i}}) = \mathbb{E}\left(N_{\mathbf{i}}N_{\mathbf{i}}^t\right)$ which leads to

$$\mathcal{D}\mathrm{Cov}(N_{\mathbf{i}})\mathcal{D}^t = \mathcal{D}\mathbb{E}\left(N_{\mathbf{i}}N_{\mathbf{i}}^t\right)\mathcal{D}^t = \mathbb{E}\left(\mathcal{D}N_{\mathbf{i}}\left(\mathcal{D}N_{\mathbf{i}}\right)^t\right) = \mathbf{M}_{\mathbf{i}} \tag{7.2}$$

thanks to (7.1). The DCT being an orthogonal transform , from (7.2) we get $\mathrm{Cov}(N_{\mathbf{i}}) = \mathcal{D}^t\mathbf{M}_{\mathbf{i}}\mathcal{D}$.

We shall apply the blind denoising to a real noisy image for which no noise model was available. To illustrate the algorithm structure and its action, we present the noisy input image, the denoised image, the difference image = noisy - denoised, the average noise curve over high frequencies, and the average noise curve over low frequencies. The results are shown in Figure 7.1. As the noise curves illustrate, the noise is frequency and signal dependent.

Results on old photographs. Scanned old photographs form a vast image corpus for which the noise model can't be anticipated. The noise is chemical, generally with big grain and further altered by the scanning and JPEG encoding. Figure 7.2 shows results obtained by the Noise Clinic over this kind of noisy images. The results compare well with those obtained with *blind BLS-GSM* [55, 56], another state-of-the-art blind denoising algorithm.

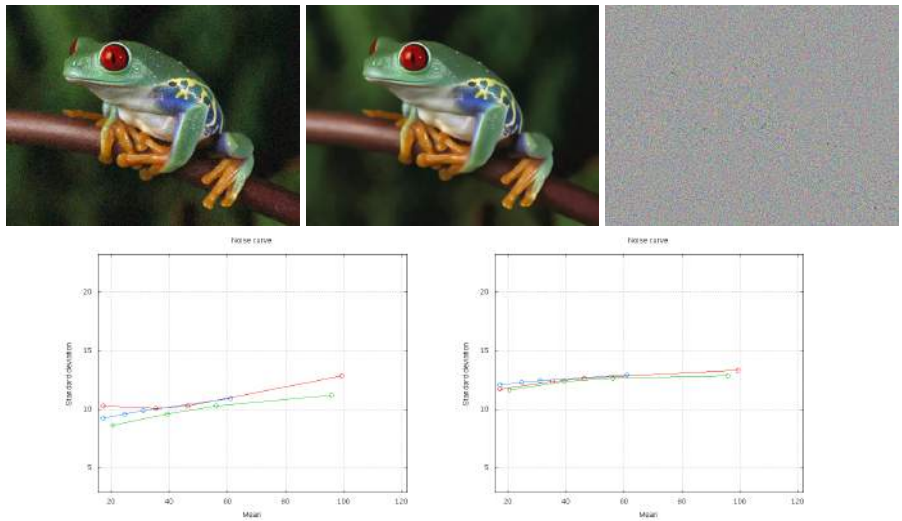


Figure 7.1. Top: Illustration of blind denoising of a JPEG image, the "Frog" image. It is advised to zoom in the high quality pdf file to see detail. Left, noisy image, middle denoised image and right, difference image. Bottom: noise variance estimation of the "Frog" image, as a function of the image value and of the local spatial DCT frequency . Left: average of the low frequency curves in the DCT. Right: average of high frequencies.

Figure 7.2. Blind denoising results on two old photographs. The first is a portrait of young Marylin Monroe. The second is a detail of a group photograph at the Solvay conference, 1927. For both, a crop of a scan of the original image is followed result of the Noise Clinic. It is advised to zoom in the pdf to see image details.

8. Conclusion

Fifty years effort have ended up with denoising methods that can be fully described with six short formulas that guarantee optimality for a definite image model: these formulas are : (2.2) and (2.4) for the Wiener-Fourier transform thresholding assuming an image sparsity model; (3.1) for the neighborhood filter and (3.2) for NL-means, both assuming a self-similarity model; (4.7) for nonlocal Bayes, which assumes again an image self-similarity and local Gaussian behavior for patches. Finally the single formula (5.3) for global Bayesian denoising, which is asymptotically optimal given a (virtually infinite) sample set of image patches. The

global (Bayesian or neural) methods bypass the question of a mathematical image model by using an *in extenso* model, namely *all* image patches of the world. For this precise reason, they are still impractical. On the other hand, the best simple image models obtain a denoising performance equivalent to global methods. This is encouraging for mathematical modeling! But are the three main image mathematical models compatible? The answer is yes: the Bayesian self-similarity image model (Nonlocal Bayes) combines the three main principles. Indeed, the Bayesian local estimate of a patch is a diagonal operator on the patch basis given by the local Gaussian model. Similarly, a recent method, *dual domain denoising* [36], also shows excellent performance by alternating and iterating a neighborhood filter with a DCT transform thresholding.

Acknowledgements. Research partially financed by the Office of Naval research under grant N00014-97-1-0839, DxO-Labs, Centre National d'Etudes Spatiales (CNES, MISS project), the European Research Council, advanced grant "Twelve labours", and the Spanish Ministerio de Ciencia e Innovación under grant TIN2011-27539.

References

[1] M. Aharon, Michael Elad, and A. Bruckstein, *K-SVD: Design of dictionaries for sparse representation*, IEEE Transactions on Image Processing (2005), 9–12.

[2] F. J. Anscombe. *The transformation of Poisson, binomial and negative-binomial data* , Biometrika, **35**(3) (1948), 246–254.

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

[4] S.P. Awate and R.T. Whitaker, *Unsupervised, information-theoretic, adaptive image filtering for image restoration*, IEEE Trans. PAMI **28**(3) (2006), 364–376.

[5] L. Bottou. *Large-scale machine learning with stochastic gradient descent*, In Proc. Int. Conf. Computational Statistics, pp. 177–186, Springer, 2010.

[6] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31, Springer Verlag, 1999.

[7] A. Buades, B. Coll, and J. M. Morel, *A review of image denoising algorithms, with a new one*, Multiscale Modeling & Simulation **4**(2) (2005), 490–530.

[8] ———, *A non local algorithm for image denoising*, IEEE Computer Vision and Pattern Recognition **2** (2005), 60–65.

[9] ———, *Nonlocal image and movie denoising*, International Journal of Computer Vision **76**(2) (2008), 123–139.

[10] ———, *Non-local means denoising*, Image Processing On Line **1** (2011).

[11] H. Burger, C. Schuler, and S. Harmeling, *Image denoising: Can plain neural networks compete with BM3D?*, In IEEE Conf. Computer Vision and Pattern Recognition, pp. 2392–2399, 2012.

[12] H.C. Burger, *Modelling and Learning Approaches to Image Denoising*. PhD thesis, Eberhard Karls Universität Tübingen, Wilhelmstr. 32, 72074 Tübingen, 2013.

[13] E.J. Candès and M.B. Wakin, *An introduction to compressive sampling*, Signal Processing Magazine, IEEE **25**(2) (2008), 21–30.

[14] P. Chatterjee and P. Milanfar, *Is denoising dead?*, IEEE Transactions on Image Processing **19**(4) (2010), 895–911.

[15] _____, *Patch-based near-optimal image denoising*, IEEE Transactions on Image Processing, 2011.

[16] C. Chevalier, G. Roman, and J.N. Niepce, *Guide du photographe*, C. Chevalier, 1854.

[17] A. Cohen, I. Daubechies, and J.C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Communications on pure and applied mathematics **45**(5) (1992), 485–560.

[18] R.R. Coifman and D.L. Donoho, *Translation-invariant de-noising*, Lecture Notes In Statistics, pp. 125–125, 1995.

[19] A. Colom, M. Buades, *Analysis and extension of the Ponomarenko et al. method, estimating a noise curve from a single image*, Image Processing On Line **3** (2013), 173–197.

[20] M. Colom, M. Lebrun, A. Buades, and J.M. Morel, *A non-parametric approach for the estimation of intensity-frequency dependent noise*, submitted, 2014.

[21] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising by sparse 3D transform-domain collaborative filtering*, IEEE Transactions on Image Processing **16**(8) (2007).

[22] _____, *Image restoration by sparse 3D transform-domain collaborative filtering* , In Electronic Imaging, 2008.

[23] J.S. De Bonet, *Noise reduction through detection of signal redundancy*, Rethinking artificial intelligence, 1997.

[24] J. Dalalyan A. Deledalle, C.A. Salmon, *Image denoising with patch based PCA: local versus global*, In Proceedings of the British Machine Vision Conference, pp. 25.1–25.10, 2011.

[25] D.L. Donoho, *De-noising by soft-thresholding*, IEEE Transactions on Information Theory **41**(3) (1995), 613–627.

[26] D.L. Donoho and J.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81**(3) (1994), 425–455.

[27] S. Durand and M. Nikolova, *Restoration of wavelet coefficients by minimizing a specially designed objective function*, In Proc. IEEE Workshop on Variational, Geometric and Level Set Methods in Computer Vision, pp. 145–152, 2003.

[28] A.A. Efros and T.K. Leung, *Texture synthesis by non-parametric sampling*. In International Conference on Computer Vision, volume 2, pp. 1033–1038, 1999.

[29] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Transactions on Image Processing **15**(12) (2006), 3736–3745.

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

[31] P. Getreuer, *Rudin-Osher-Fatemi total variation denoising using split Bregman*, Image Processing On Line **2** (2012).

[32] Guillermo Sapiro Guoshen Yu, *DCT image denoising: a simple and effective image denoising algorithm*, Image Processing On Line **1** (2011).

[33] S.R.J.L. Harris, *Image evaluation and restoration*, Journal of the Optical Society of America **56**(5) (1966), 569–570.

[34] J. Hays and A.A. Efros, *Scene completion using millions of photographs*, In ACM Transactions on Graphics, volume 26, p. 4, 2007.

[35] K. Hornik, M. Stinchcombe, and H. White. *Multilayer feedforward networks are universal approximators*, Neural Networks **2**(5) (1989), 359–366.

[36] C. Knaus and M. Zwicker, *Dual-domain image denoising*, In IEEE International Conference on Image Processing, pp. 440–444, 2013.

[37] M. Lebrun, *An analysis and implementation of the BM3D image denoising method*, Image Processing On Line **2** (2012).

[38] M. Lebrun, A. Buades, and J. M. Morel, *A nonlocal Bayesian image denoising algorithm*, SIAM Journal on Imaging Sciences **6**(3) (2013), 1665–1688.

[39] M. Lebrun, M. Colom, and J.M. Morel, *The noise clinic: a universal blind denoising algorithm*, submitted, 2014.

[40] M. Lebrun and A. Leclaire, *An implementation and detailed analysis of the K-SVD image denoising algorithm*, Image Processing On Line **2** (2012).

[41] Marc Lebrun, Antoni Buades, and Jean-Michel Morel, *Implementation of the "Non-Local Bayes" (NL-Bayes) image denoising algorithm*, Image Processing On Line, **3** (213), 1–42. http://dx.doi.org/10.5201/ipol.2013.16.

[42] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient backprop*, In Neural networks: Tricks of the trade, pp. 9–50. Springer, 1998.

[43] J.S. Lee, *Digital image smoothing and the sigma filter*, Computer Vision, Graphics, and Image Processing **24**(2) (1983), 255–269.

[44] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks **6**(6) (1993), 861–867.

[45] A. Levin and B. Nadler, *Natural image denoising: Optimality and inherent bounds*, In IEEE Conference on Computer Vision and Pattern Recognition (2011), 2833–2840.

[46] Liu, W. Freeman, R. Szeliski, and S. Kang, *Automatic estimation and removal of noise from a single image*, IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2) (February 2008), 299–314.

[47] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, *Non-local sparse models for image restoration*, In International Conference on Computer Vision, (2009), 2272–2279.

[48] J. Mairal, G. Sapiro, and M. Elad, *Learning multiscale sparse representations for image and video restoration*, SIAM Multiscale Modeling and Simulation **7**(1) (2008), 214–241.

[49] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic press, 1999.

[50] Y. Meyer, *Wavelets-algorithms and applications*, Wavelets-Algorithms and applications Society for Industrial and Applied Mathematics Translation., p. 142, 1, 1993.

[51] E. Ordentlich, G. Seroussi, S. Verdu, M. Weinberger, and T. Weissman, *A discrete universal denoiser and its application to binary images*, In International Conference on Image Processing, volume 1, 2003.

[52] E. Le Pennec and S. Mallat, *Geometrical image compression with bandelets*, In Proceedings of the SPIE 2003, volume 5150, pp. 1273–1286.

[53] N. Pierazzo and M. Rais, *Boosting shotgun denoising by patch normalization*, In IEEE International Conference on Image Processing (2013).

[54] N. Ponomarenko, V. Lukin, K. Egiazarian, and J. Astola, *A method for blind estimation of spatially correlated noise characteristics*, In IS&T/SPIE Electronic Imaging, pp. 753208–753208, International Society for Optics and Photonics, 2010.

[55] J. Portilla, *Blind non-white noise removal in images using Gaussian scale mixtures in the wavelet domain*, Benelux Signal Processing Symposium, 2004.

[56] ———, *Full blind denoising through noise covariance estimation using Gaussian scale mixtures in the wavelet domain*, IEEE International Conference on Image Processing **2** (2004), 1217–1220.

[57] T. Rabie, *Robust estimation approach for blind denoising*, IEEE Transactions on Image Processing **14**(11) (November 2005), 1755–1765.

[58] Boshra Rajaei, *An Analysis and Improvement of the BLS-GSM Denoising Method*, Image Processing On Line **4** (2014), 44–70.

[59] M. Raphan and E.P. Simoncelli, *An empirical Bayesian interpretation and generalization of NL-means*, Technical report, TR2010-934, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University, 2010.

[60] W.H. Richardson, *Bayesian-based iterative method of image restoration*, Journal of the Optical Society of America **62**(1) (1972), 55–59.

[61] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D **60** (1992), 259–268.

[62] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, *Labelme: a database and web-based tool for image annotation*, International journal of computer vision **77**(1) (2008), 157–173.

[63] C.E. Shannon, *A mathematical theory of communication*, ACM SIGMOBILE Mobile Computing and Communications Review **5**(1) (2001), 3–55.

[64] S.M. Smith and J.M. Brady, *SUSANA new approach to low level image processing*, International Journal of Computer Vision **23**(1) (1997), 45–78.

[65] J.L. Starck, E.J. Candes, and D.L. Donoho, *The curvelet transform for image denoising*, IEEE Transactions on Image Processing **11**(6) (2002), 670–684.

[66] C. Tomasi and R. Manduchi, *Bilateral filtering for gray and color images*, In Computer Vision, 1998. Sixth International Conference on, pp. 839–846, 1998.

[67] Y. Q. Wang and J. M. Morel, *Can a single image denoising neural network handle all levels of Gaussian noise?*, IEEE Signal Processing Letters, 2014. to appear.

[68] ———, *SURE guided Gaussian mixture image denoising*, SIAM Journal on Imaging Sciences **6**(2) (2013), 999–1034.

[69] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and MJ Weinberger, *Universal discrete denoising: Known channel*, IEEE Transactions on Information Theory **51**(1) (2005), 5–28.

[70] L. Yaroslavsky and M. Eden, Fundamentals of Digital Optics, 2003.

[71] L. P. Yaroslavsky, *Digital Picture Processing*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1985.

[72] ———, *Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window*, In Proceedings of SPIE, volume 2825, pp. 2–13, 1996.

[73] L.P. Yaroslavsky, K.O. Egiazarian, and J.T. Astola, *Transform domain image restoration methods: review, comparison, and interpretation*, In Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, volume 4304, pp. 155–169, May 2001.

[74] G. Yu and G. Sapiro, *DCT image denoising: a simple and effective image denoising algorithm*, Image Processing On Line, 1, 2011.

[75] G. Yu, G. Sapiro, and S. Mallat, *Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity*, IEEE Transactions on Image Processing **21**(5) (2012), 2481–2499.

[76] L. Zhang, W. Dong, D. Zhang, and G. Shi, *Two-stage image denoising by principal component analysis with local pixel grouping*, Pattern Recognition **43**(4) (2010), 1531–1549.

[77] D. Zoran and Y. Weiss, *From learning models of natural image patches to whole image restoration*, International Conference on Computer Vision, 2011.

CMLA, ENS Cachan, 61 av. du Psdt Wilson, 94235 Cachan Cedex France
E-mail: morel@cmla.ens-cachan.fr