Penalizing local correlations in the residual improves image denoising performance

Paul Riot, Andrés Almansa, Yann Gousseau, Florence Tupin
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
Email: paul.riot@telecom-paristech.fr

***Abstract*—In this work, we address the problem of denoising an image corrupted by an additive white Gaussian noise. This hypothesis on the noise, despite being very common and justified as the result of a variance normalization step, is hardly used by classical denoising methods. Indeed, very few methods directly constrain the whiteness of the residual (the removed noise). We propose a new variational approach defining generic fidelity terms to locally control the residual distribution using the statistical moments and the correlation on patches. Using different regularizations such as TV or a nonlocal regularization, our approach achieves better performances than the L2 fidelity, with better texture and contrast preservation.**

***Index Terms*—Image denoising, White noise, Cost function, Probability distribution.**

I. Introduction

Image denoising is a mandatory step in most image processing chains, potentially useful both for image interpretation and image enhancement. The most common assumption on the image noise is that is it additive, white and Gaussian. This assumption, although false on the raw images because of photonic noise, can be made realistic by a variance normalization step as shown in [1]. We note $g = f + n$ with $f \in \mathbb{R}^D$ the noise-free image (D the number of pixels in the image domain Ω), n a white Gaussian noise of variance σ^2 and g the noisy image. When an estimation u of the noise-free image is computed, we call residual $\epsilon = u - g$ the difference between the denoised image and the noisy image.

The maximum a posteriori method consists in finding a denoised image u that maximizes $P(u|g)$. This term can be decomposed using the Bayes law into a product of two terms: $P(g|u)$, the data fidelity term, and $P(u)$ the regularization term. This is usually expressed as

$$u = \underset{u \in \mathbb{R}^D}{\arg\min} \, ||g - u||^2 + \lambda \; J_{reg}(u) \tag{1}$$

with $J_{reg}(u)$ a regularization function adapted to the image. The data fidelity term is obtained from the negative log-likelihood of $P(g|u)$ which boils down to $||g - u||^2$ in the white Gaussian case. One can note that this is a pixelwise term which does not explicitely enforce the whiteness of the residual. Of course, the independance hypothesis is used to obtain the likelihood term and if the regularization model perfectly fits the image, this term would be optimal. In practice, the prior model is never globally optimal on the image. Then, $||g - u||^2$ gives very few guarantees about the distribution and the whiteness of the residual. In the litterature, most of the works try to improve the image model or locally adapt λ to obtain the best possible regularization term. In this paper, we introduce generic variational terms that directly constrain the mean, the variance and the whiteness of the residual. Those terms can be coupled with any regularization function $J_{reg}(u)$. In particular, we will study their performances using two different regularizations models presented in the following.

The Rudin, Osher and Fatemi model (ROF) [2] is a widely used method. It relies on a Total Variation (TV) regularization $J_{reg}(u) = TV(u) = \sum_{(i,j)\in\Omega} ||\nabla u(i,j)||$, where $\nabla u(i,j)$ is the gradient at pixel (i,j). This term smoothes the denoised image while preserving edges. However, it does not usually preserve well textures and contrast. This formulation has the advantage to be convex and a global minimum can be found using different algorithms such as [3], [4].

First introduced in [5] and extended in [6], the nonlocal methods rely on the redundancy of the image to achieve denoising. They find similar patches throughout the image and average them to obtain a denoised image. Interestingly, those methods, although being spatial filtering techniques, can be expressed as regularization terms suited for a variational approach [7], [8]. Weights w are computed on the noisy image in the same manner as in the Non Local Means method. Then, the regularization term is obtained as:

$$NL(u) = \sum_{i\in\Omega} \sum_{j\in W(i)} w(i,j)(u(i) - u(j))^2 \tag{2}$$

W being the search window. This term can be understood as a nonlocal gradient.

Here, we propose to study the data fidelity term and improve it to obtain some guarantees on the residual distribution and whiteness. Indeed, the residual obtained by the different methods mentionned earlier do not always respect the noise model, especially locally, as shown in [9]. To the best of our knowledge, only very few works proposed to take advantage of the noise statistics in this manner [10], [11], [12], [13], [14]. The first article [10] by Teuber et al. proposes to denoise 1D signals by cutting them into smaller parts and by controlling the mean and variance of the residual using L2 distances. It also proposes to control the correlation of the residual at one particular lag. Other methods [11], [12], [13] presented by Lanza et al. use an Alternating Direction Method on features that are computed globally. In the first two papers, the Fourier transform of the residual and its autocorrelation are respectively considered to constrain the whiteness of the noise. In [13], the residual cumulative histogram is used to control its distribution. Finally, the work presented by Fehrenbach et al. in [14] proposes to constrain a tiling of the Fourier transform of the residual. In this article, we propose to use features computed locally on patches such as the residual mean, variance and autocorrelation to constrain ϵ.

In the first section, we introduce the terms used to control the first and second order moments of the residual, respectively controlling its mean and variance. They make sure that the method removes the right amount of noise and preserves the contrast. The

second section is dedicated to the whiteness term. Its purpose is to constrain the whiteness of the residual, therefore ensuring that no information is removed from the image during the denoising process. Finally, we present some experimental results in the third section. In particular, experiments show that for both the TV and the NL regularizations, the proposed terms outperform the usual L2 data fidelity term.

II. Controlling the residual moments

Introduced in [10], the idea of replacing the likelihood term $||u-g||^2$ by terms controlling moments of various orders is motivated by several reasons. First, methods such as ROF can obtain residuals that do not match the noise model as it is shown in the experiments in section IV.

Second, the methods using $||u-g||^2$ as a data fidelity term are hard to parametrize globally. This problem is well recognized and many works try to tackle it [15], [16]. Indeed, $||u-g||^2$ is minimized when $u=g$, while the regularization is usually minimized on a totally different space. The estimation u is obtained with a trade-off of both functions. Thus, the choice of the parameter λ is critical to obtain the right amount of denoising on structures having different scales. The trade-off is often impossible to achieve globally and the flat areas are too noisy while the textures are blurred out.

We aim at designing generic fidelity terms which guarantee that the right amount of noise is removed with a relative stability regarding the regularization parameter λ. For that task, the statistical moments of the residual are particularly relevant. Indeed, controlling the first order moment, the mean, allows to obtain a centered residual, while the second order moment, the variance, is useful to remove the right amount of energy.

In order to preserve all the structures and textures in the image, we need to compute the features locally. Thus, we extract K overlapping square patches ϵ_k of size s^2 from the residual image ϵ.

The first order moments of the residual computed on each patch are defined as $\mu_{\epsilon_k} = \frac{1}{s^2}\sum_{i=1}^{s^2}\epsilon_k(i)$. As a sum of s^2 independent Gaussian variables, μ_{ϵ_k} follows a normal distribution with variance $\frac{\sigma^2}{s^2}$. Using the sum of the negative log-likelihoods of each μ_{ϵ_k}, we obtain the following term:

$$J_{mean}(\epsilon) = \frac{s^2}{2K\sigma^2}\sum_{k=1}^{K}\mu_{\epsilon_k}^2 \tag{3}$$

This term can be seen as a relaxed version of the term $||u-g||^2$ over each patch. Indeed, with $s^2=1$, $J_{mean}(\epsilon)$ is proportional to $||u-g||^2$.

The second order moment of the residual is defined as $\sigma^2_{\epsilon_k} = \frac{1}{s^2}\sum_{i=1}^{s^2}(\epsilon_k(i))^2$. This is a sum of s^2 squared centered Gaussian variables and, as such, it follows a χ^2 distribution with s^2 degrees of freedom. Once again, using the sum of the negative log-likelihoods of each $\sigma^2_{\epsilon_k}$, the following term is obtained:

$$J_{var}(\epsilon) = \frac{1}{K}\sum_{k=1}^{K}\left((\frac{s^2}{2}-1)\log(s^2\sigma^2_{\epsilon_k}) - \frac{s^2\sigma^2_{\epsilon_k}}{2\sigma^2}\right) \tag{4}$$

It guarantees that the right amount of energy is removed locally over each patch. The residuals that jointly minimize J_{mean} and J_{var} possess the right characteristics: zero mean and variance σ^2.

Let us stress the importance of computing the moments locally. Indeed, a global approach would result in spatially different behaviour of the method. Some textured part would be left untouched (not denoised), while smoother areas would be overly reguralized. The size of the patches is an important parameter since the quality of the estimators decreases when the number of samples is small.

Our experiments showed that J_{mean} and J_{var} present interesting performances in terms of SNR, but they introduce low frequency artifacts. Indeed, both are too constrained. They are minimized at the most probable realisation of noise, but are only estimated on a few samples. In order to fix this issue, we expand the space of the minimizers of both functions with the following expressions:

$$J_{mean+}(\epsilon) = \frac{s^2}{K\sigma^2}\sum_{k=1}^{K}(\mu^2_{\epsilon_k} - \frac{\sigma^2}{s^2})^+ \tag{5}$$

$$J_{var+}(\epsilon) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{s^2}{2\sigma^4}(\sigma^2_{\epsilon_k}-\sigma^2)^2 - 1\right)^+ \tag{6}$$

with $(x)^+ = max(0,x)$, the positive part function. Those relaxed versions of both terms are minimized on much wider spaces, as on Figure 1, and produce no artifacts.

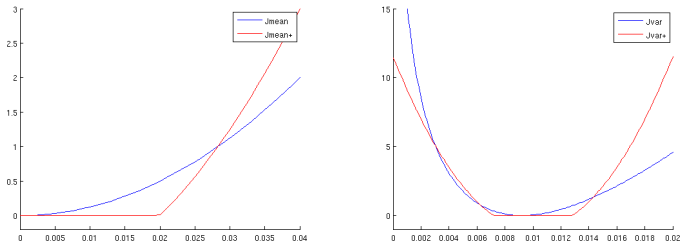

Fig. 1: Comparison of J_{mean}, J_{mean+} and J_{var}, J_{var+} as functions of respectively μ_{ϵ_k} and σ_{ϵ_k} for the parameters: $\sigma = 0.1$ and $s = 5$.

In this section, we designed new terms that control the mean and the variance of the residual to match the noise model. The next section explains how to further control the statistics of the residual and therefore ensure a better preservation of the information.

III. Imposing whiteness constraints

In this section, we explain how to constrain the whiteness of the residual, that is how to enforce its decorrelation. Although the decorrelation of the noise is often assumed, it is very rarely enforced. Some articles show that evaluating this hypothesis on the residual could be used as a way to estimate the quality of the denoising [17]. But, only a few [10], [11], [12], [13], proposed to use it directly to perform denoising. Lanza in [11], [12], [13] proposed different possible features to enforce the residual decorrelation such as the periodogram, the histogram or the autocorrelation matrix. The performances are not showing great differences between methods.

Here, we propose to use the most direct tool: the autocorrelation estimation matrices r_k computed on each patch. More precisely, in order to simplify the calculation, speed up the computation process and improve the higher lag estimation, we used the circular convolution to compute those matrices. This decision

should be harmless for our method given the noise model. The r_k are defined for all possible lags $(l, m) \in \{-\frac{s}{2}, ..., \frac{s}{2}\}$:

$$r_k(l,m) = \frac{1}{s^2} \sum_{(i,j)} \epsilon_k(i,j)\epsilon_k(i+l, j+m) \tag{7}$$

r_k is a classical estimator of the autocorrelation. As we have previously defined J_{var+} to control the variance, we discard $r_k(0,0)$ in the following. In practice, there is no known analytic expression for the distribution of r_k. Using the central limit theorem, it is possible to approximate it with a Gaussian and use its negative log-likelihood as our cost function. However, once again, the resulting expression is too constrained, since it penalizes too much the matrices that are not zero everywhere, and induce wave-like artifacts. Thus, the L2-norm of r_k is used to design our fidelity term: $||r_k||_2 = \sqrt{\sum_{(l,m)\neq(0,0)} r_k(l,m)^2}$.

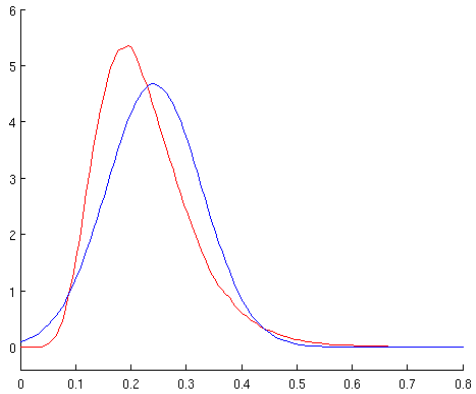

Fig. 2: Empirical distribution of $||r_k||_2$ (red) and its Gaussian approximation (blue) with the parameters: $\sigma = 0.1$ and $s = 5$.

The term $||r_k||_2$ follows the distribution shown in figure 2 obtained empirically on white Gaussian noise samples using the Kernel Density Estimation (KDE) algorithm from [18]. Supposing that all the $r(l,m)$ are independent, we get a variance equal to $(s^2-1)\sigma^4$ and a mean equal to $\sigma^2\sqrt{(s^2-1)}s$. However, the independence hypothesis is obviously not true and the variance is under-estimated. To correct this effect, we established empirically that a multiplicative factor of 3 should be applied. The same factor has been found for different noise levels and parameters. Finally, we approximate this distribution using a Gaussian with mean $\sigma^2\sqrt{(s^2-1)}s$ and variance $3(s^2-1)\sigma^4$. Then, we compute its log-likelihood to obtain:

$$J_{white}(\epsilon) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{6(s^2-1)\sigma^4}\left(||r_k||_2 - \sigma^2\sqrt{(s^2-1)}s\right)^2 \tag{8}$$

This term prevents the method from removing correlated content such as textures, contrast or edges. By constraining a weakly correlated residual, the method preserves this information. Once again, the choice of the patch size s is critical to obtain a good estimator. Also, the choice of the circular convolution allows for a better estimation of the lower frequencies (higher lags).

One could note that our expressions for J_{var+} and J_{white} are not centered (we did not remove the mean to estimate both the variance and the autocorrelation). This choice was motivated by our experiments which showed slightly better performances. However, the differences were not visually significant.

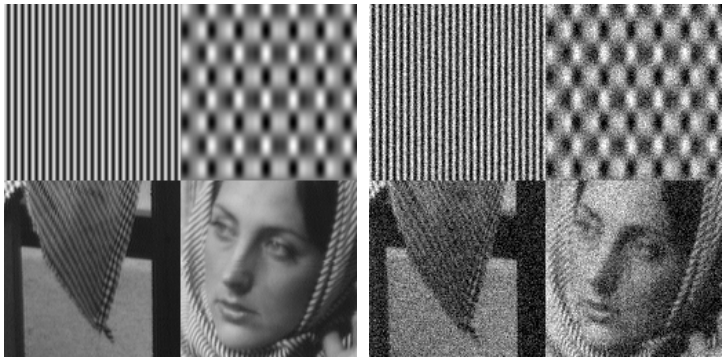

Fig. 3: Noise-free image and noisy image with $\sigma = 0.1$.

IV. Experimental results

A. Experimental settings

In this section, we will show experimental results obtained for a mosaic image using the L2 data fidelity term, and the proposed method. More complete experiments are shown on the web-page http://perso.telecom-paristech.fr/~riot/EUSIPCO2016. Both approaches are expressed with the following cost functions:

$$u = \arg\min_{u\in\mathbb{R}^D} J_{mean+}(u) + J_{var+}(u) + J_{white}(u) + \lambda\ J_{reg}(u) \tag{9}$$

$$u = \arg\min_{u\in\mathbb{R}^D} ||u-g||^2 + \lambda\ J_{reg}(u) \tag{10}$$

with $J_{reg}(u)$ being a regularizer. The two regularizers $TV(u)$ and $NL(u)$ introduced in section I will be studied in this work.

One of the main drawback of our method is that its cost function is not convex and not smooth. Thus, the optimization process is much more difficult than using the L2-norm which is convex and smooth. To obtain differentiable cost functions in all the considered cases, we used a smooth approximation of TV. In the same manner, a smooth approximation of the positive part was used, with a a large number:

$$(x)^+ = \frac{\log(1+\exp(-ax))}{a} + x \tag{11}$$

Still, since the cost function of the proposed method is not convex, the result depends on the initialization and can only be expected to be a local minimum. For our experiments, we used the L-BFGS method [19]. This approach performs a Quasi-Newton method using a low-rank Hessian. This allows the algorithm to run using a reasonnable amount of memory on large images. The method is initialized using the noisy image.

Our method requires one more parameter than the L2 fidelity: the patch size s. The experiments show that $s = 15$ gives the best performance on the considered images. It is large enough to obtain a good estimator of the moments and of the autocorrelation matrix, and small enough to constrain locally the residual to match the noise model. It is interesting to note that as we are working directly on the residual and not on the image itself, the choice of this parameter should not depend on the image. We also observed that extracting only a quarter of all the patches very marginally affects the performances and greatly reduces the computational time.

In order to compare performance on a fair basis, all the shown results use an oracle estimation of λ that maximizes the SNR:

$$SNR(\hat{u}) = 10\log_{10}\left(\frac{||\hat{u}||^2}{||f-\hat{u}||^2}\right) \tag{12}$$

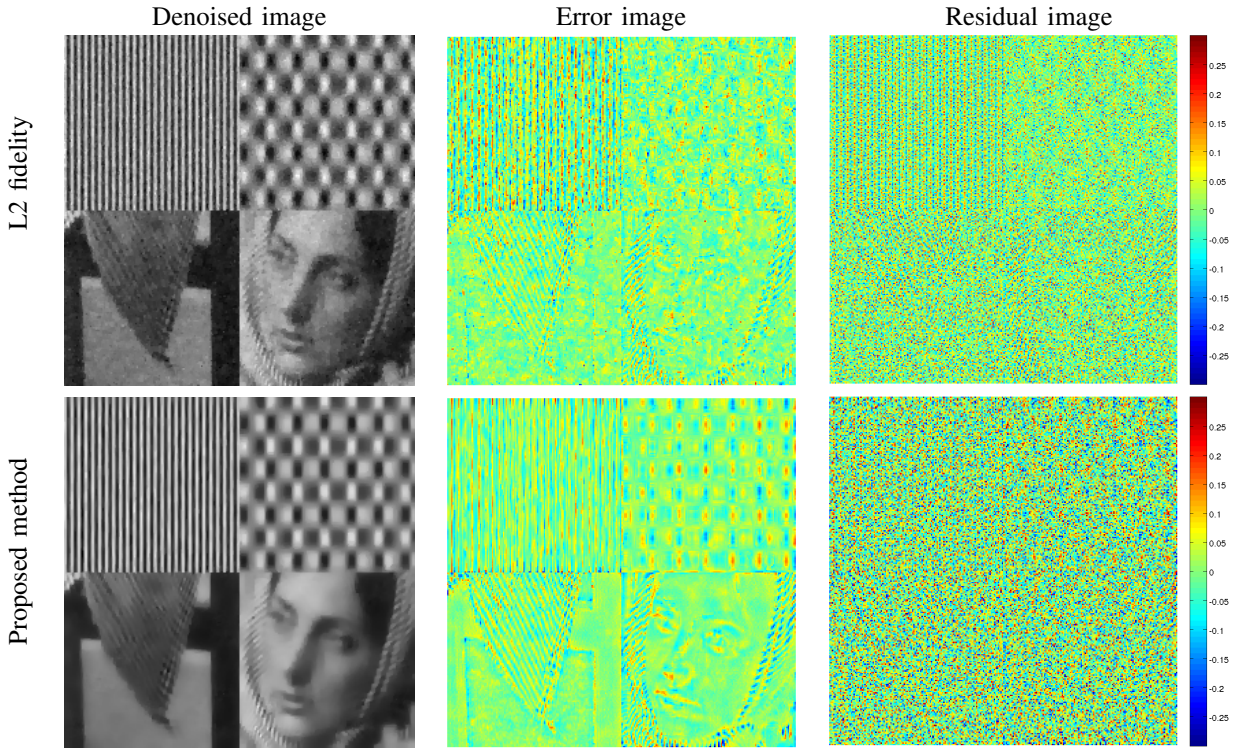

Fig. 4: Denoised, error and residual images for the L2 fidelity and the proposed method using *TV regularization*.

B. TV regularization

The results are shown on Figure 4. The overall impression from the result of the proposed method is more satisfying than with the L2 fidelity term. Indeed, the whole image cannot be well handled by the ROF model as it would require different values λ for each texture. Furthermore, the residual in the top left texture is highly correlated. A large amount of texture information was removed. Overall, the image does not seem sufficiently denoised although the parameter λ was chosen to obtain the best possible SNR. Increasing λ would remove more energy from the texture than from the noise.

On the other hand, the proposed algorithm presents a smoother denoised image. J_{mean+} also limits the losses of contrast which are still visible on the top right texture. Using the J_{mean} term instead, it is possible to completely solve this problem. But, as said earlier, it induces artifacts and was discarded. J_{var+} is a guarantee that the same amount of noise is removed from all parts of the image although λ was chosen globally. This can be verified on the residual image which is more homogeneous spatially and with the right energy when using the proposed method. Finally, J_{white} enforces the whiteness hypothesis and makes sure that a decorrelated residual is obtained. This is particularly visible for the top left textures. The residual appears much less correlated than using the ROF model where we can see the vertical stripes in the residual. Still, the napkin texture on the bottom left is damaged by both methods as it is particularly hard to retrieve. One could design a more constrained version of J_{white}, however our experiments showed that it leads to wave-like artifacts when the noise realisation is locally correlated.

C. NL regularization

Experiments shown on figure 5 were run using the NL regularizer proposed by [7], [8]. In the literature, several propositions were made to compute the weights $w(i,j)$. Here, we chose the most basic one presented in [5] with the following formula:

$$w(i,j) = \exp \frac{-||P_g(i) - P_g(j)||^2}{s_{NL}^2 h_{NL}^2} \tag{13}$$

with $P_g(i)$ the patches from the noisy image centered on i, s_{NL} the patch size, h_{NL} the bandwidth parameter. As stated in [5], and verified by our testing, it is better to only compute the weights and gradient over a search window W_{NL} rather than over the whole image domain to avoid the accumulation of wrong information. We chose $W_{NL} = 10$, $s_{NL} = 5$ and $h_{NL} = 0.8\sigma$.

Once again, the proposed method performs better than the classical one with respect to several criteria. First, it achieves a satisfying level of denoising throughout the whole image. This is allowed because each term is local. Using the L2 fidelity, the algorithm performs almost no denoising on the top left texture where the regularization finds very few similar patches. Second, the contrast losses are way less severe than using the L2 fidelity. Still visible on the top right texture, they are much lower in amplitude. Observe also that the residual contains less structure with our approach. In particular, the face is very visible on the L2 fidelity residual but hardly with the proposed method. This is also visible on the napkin on the bottom left image where less information is removed.

Several SNR results are available on table I. Overall, the proposed method improves the performances on heterogeneous images with different structure sizes and possibly different textures. On the other hand, on homogeneous and weakly textured images such as Pepper or Flinstones, where a single value for λ can be globally optimal using the L2 fidelity, it does not improve the SNR. However, one must remain careful with the SNR as it does not always carry all the needed information such as the presence of artifacts, or texture preservation.

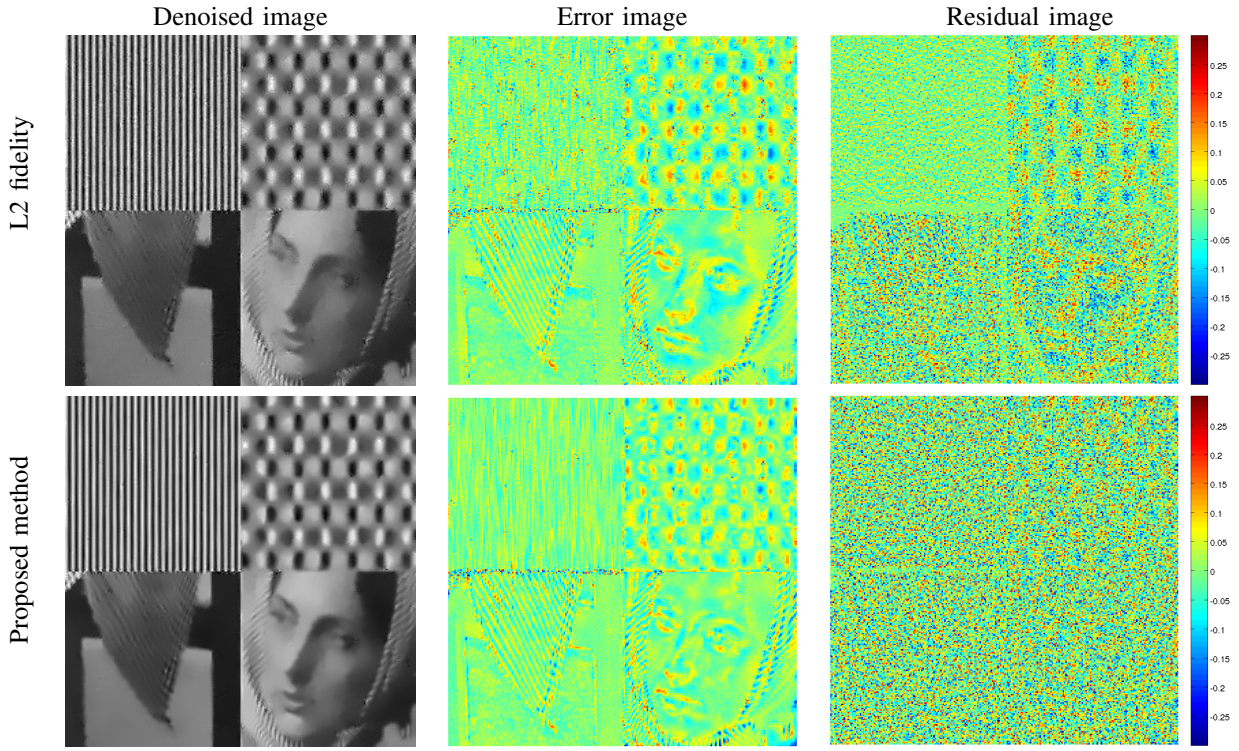

Fig. 5: Denoised, error and residual images for the L2 fidelity and the proposed method using *NL regularization*.

method	L2-TV	Proposed-TV	L2-NL	Proposed-NL
Composite image	19.12	**20.98**	19.85	**21.58**
Lena	22.01	**22.12**	21.36	**22.08**
Barbara	19.45	**21.05**	20.65	**21.59**
Pepper	**24.79**	24.69	23.60	**24.56**
Flinstones	**21.68**	21.31	21.78	**21.95**

TABLE I: SNR performances (in dB) of the different algorithms on 5 images. Composite image is the image shown in this paper, the other results are shown on the webpage.

Conclusion

In this work, we presented new terms to be used in a variational denoising framework. Those terms are shown to be more suited than the L2 fidelity term and to enforce the statistical hypothesis on the residual. The results appear smoother, with less contrast losses and better texture preservation. They also show significant improvements in terms of SNR. Multiple points could be improved such as the autocorrelation modelization or using a better initialization.

Acknowledgement: This work has been partially funded by the French Research Agency (ANR) under grant nro ANR-14-CE27-001 (MIRIAM).

References

[1] M. Colom, A. Buades, and J.-M. Morel, "Nonparametric noise estimation method for raw images," *JOSA A*, vol. 31, no. 4, pp. 863–871, 2014.

[2] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[3] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical imaging and vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[4] L. Condat, "A generic proximal algorithm for convex optimization—application to total variation minimization," *Signal Processing Letters, IEEE*, vol. 21, no. 8, pp. 985–989, 2014.

[5] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[6] M. Lebrun, A. Buades, and J. Morel, "A nonlocal bayesian image denoising algorithm," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1665–1688, 2013.

[7] A. Elmoataz, O. Lezoray, and S. Bougleux, "Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing," *Image Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1047–1060, 2008.

[8] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *Multiscale Modeling & Simulation*, vol. 7, no. 3, pp. 1005–1028, 2008.

[9] M. Nikolova, "Model distortions in bayesian map reconstruction," *Inverse Problems and Imaging*, vol. 1, no. 2, p. 399, 2007.

[10] T. Teuber, S. Remmele, J. Hesser, and G. Steidl, "Denoising by second order statistics," *Signal Processing*, vol. 92, no. 12, pp. 2837–2847, 2012.

[11] A. Lanza, S. Morigi, F. Sgallari, and A. J. Yezzi, "Variational image denoising based on autocorrelation whiteness," *SIAM Journal on Imaging Sciences*, vol. 6, no. 4, pp. 1931–1955, 2013.

[12] A. Lanza, S. Morigi, and F. Sgallari, "Variational image restoration with constraints on noise whiteness," *Journal of Mathematical Imaging and Vision*, pp. 1–17, 2014.

[13] A. Lanza, S. Morigi, F. Sgallari, and A. J. Yezzi, "Variational image denoising while constraining the distribution of the residual," *Electronic Transactions on Numerical Analysis*, vol. 42, pp. 64–84, 2014.

[14] J. Fehrenbach, M. Nikolova, G. Steidl, and P. Weiss, "Bilevel image denoising using gaussianity tests," in *Scale Space and Variational Methods in Computer Vision*. Springer, 2015, pp. 117–128.

[15] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Variational denoising of partly textured images by spatially varying constraints," *Image Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2281–2289, 2006.

[16] C. Sutour, C.-A. Deledalle, and J.-F. Aujol, "Adaptive regularization of the nl-means: Application to image and video denoising," *Image Processing, IEEE Transactions on*, vol. 23, no. 8, pp. 3506–3521, 2014.

[17] M. S. Almeida and M. A. Figueiredo, "Parameter estimation for blind and non-blind deblurring using residual whiteness measures," *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2751–2763, 2013.

[18] Z. Botev, J. Grotowski, D. Kroese *et al.*, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[19] J. Nocedal, "Updating quasi-newton matrices with limited storage," *Mathematics of computation*, vol. 35, no. 151, pp. 773–782, 1980.