A hybrid approach combining cnns and variational modelling for blind image denoising

Rim Rekik Dit Nekhili, Xavier Descombes, Luca Calatroni

A HYBRID APPROACH COMBINING CNNS AND VARIATIONAL MODELLING FOR BLIND IMAGE DENOISING

Rim Rekik Dit Nekhili[1]*, Luca Calatroni*[2]*, Xavier Descombes*[3]

[1]Université Grenoble Alpes, Inria, Grenoble, France
[2]Université Côte d'Azur, CNRS, Inria, Laboratoire I3S, Sophia Antipolis, France
[3]Université Côte d'Azur, Inria, CNRS, Laboratoire I3S, Sophia Antipolis, France

ABSTRACT

We consider the problem of image denoising with unknown noise distribution. We propose a hybrid approach where model-based space-variant total variation (TV) regularization is used for denoising with hyperparameters estimated locally using a Convolutional Neural Network (CNN) with a simple and light architecture. The special choice of the weighted TV prior allows for the use of a limited learning set, while the use of the proposed CNN approach allows for local parameter estimation independently of the type of noise in the data. The obtained results show that the proposed hybrid approach takes benefit from both the prior information encoded in the choice of the regularization model and the versatility of the CNN-based parameter estimation approach.

Index Terms— Image denoising, Total Variation, Convolutional Neural Networks, Local hyperparameter estimation.

1. INTRODUCTION

Image restoration is a standard imaging inverse problem that includes image denoising and image deconvolution, which is classically addressed by optimizing a composite model defined in terms of a prior or regularization term to smooth the solution and a data fidelity or likelihood term taking into account the attachment with the given image data. A reference approach in this field consists in combining the Total Variation (TV) prior promoting sparse image gradients with an ℓ_2 data term, see [2]. These two terms are typically weighted by a scalar hyperparameter to be estimated in order to obtain good reconstruction and data consistency. Recently, some approaches to allow this parameter to vary locally (i.e. at each pixel location) in order to take into account specific image structures such as edges or texture, see, e.g. [1, 3]. To estimate such space-variant parameters an iterative algorithm alternating image optimization and parameter estimation is used. These approaches necessitate the definition of a proper noise model, which is usually considered as Gaussian.

Following their success in classification or object detection problems, convolutional neural networks (CNN) have been also employed for image restoration. They offer an optimization framework to solve inverse problems without the use of a pre-defined model, thus avoiding model parameter estimation. In this framework, the model information is learned from a huge collection of examples constructed with corrupted images and associated denoised or deconvolved images. Transfer learning can be applied to reduce its size, but collecting a suitably large and representative learning set is still challenging.

In this paper, we propose to combine a model-based approach with CNNs, so as to benefit from the prior information of the model considered, while limiting the size of the learning set at the same time. As model parameters are locally estimated by the CNN, the proposed approach is flexible as it adapts also to possibly non-Gaussian noise distributions. Furthermore, as only one scalar hyperparameter is learned at each pixel, a light architecture for the CNN and a reduced learning set can be considered.

2. SPACE-VARIANT TV DENOISING

Given a noisy vectorized image $\mathbf{y} \in \mathbb{R}^N$, the standard formulation of the image denoising problem reads:

$$\text{find} \quad \mathbf{x} \quad \text{s.t.} \quad \mathbf{y} = \mathcal{T}(\mathbf{x}), \tag{1}$$

where $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$ is a possibly non-linear noise degradation operator whose definition depends on the noise statistics assumed. In the case of additive noise degradation, problem (1) takes the form $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{n} \in \mathbb{R}^n$ denotes the random vector whose components are drawn by, for instance, a Gaussian noise distribution of zero mean and given variance σ^2, so that $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{Id})$. More complicated, possibly non-linear noise degradation models can be considered to model, e.g., signal-dependent Poisson and/or multiplicative speckle noise.

RRDN performed her work while visiting the team Morpheme, Inria, Sophia-Antipolis, France. She is grateful to I. Boudali (University of Tunis MANAR). LC acknowledges the support of the H2020 RISE grant NoMADS, GA 777826. The authors are warmly thankful to A. Lanza (University of Bologna) and M. Pragliola (University of Naples) for useful discussions on the proposed model and the one in [1].

Bayesian formulation. A standard approach for solving (1) consists in considering the quantities into play in terms of their probability density functions (p.d.f.) and looking for solutions $\mathbf{x}^*$ of the following maximum-likelihood problem:

$$\mathbf{x}^* \in \arg\max_{\mathbf{x}} \ \mathrm{p}(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} \ \frac{\mathrm{p}(\mathbf{y}|\mathbf{x})\mathrm{p}(\mathbf{x})}{\mathrm{p}(\mathbf{y})}, \quad (2)$$

by Bayes formula, where $\mathrm{p}(\mathbf{y}|\mathbf{x})$ is the noise-likelihood function associated to the distribution of noise in the data, $\mathrm{p}(\mathbf{x})$ is the so-called *prior* p.d.f. whose form depends on prior knowledge on the solution and $\mathrm{p}(\mathbf{y})$ is a normalization constant.

A popular choice for $\mathrm{p}(\mathbf{x})$ which has become increasingly popular over the last decades, consists in assuming that $\mathrm{p}(\mathbf{x}) = \mathrm{p}(\|\mathbf{D}\mathbf{x}\|) = \mathcal{L}(0, \alpha\mathbf{Id})$ with $\alpha > 0$, that is, for all image pixels $i = , 1 \ldots, N$ the quantities $\|(\mathbf{D}\mathbf{x})_i\|$ are i.i.d. drawn from a Laplace distribution with scale parameter α. This choice, however, has been previously shown [3] to be too rigid to adapt to the actual distribution of natural image gradients. For this reason, a plethora of more tailored approaches has been proposed, relying, e.g., on the design of hyper-Laplacian distributions [4] or on non-stationary Markov random fields [5, 3] have been proposed. In this latter case, the idea is to assume that at each pixel $i = 1, \ldots, N$ the underlying prior p.d.f. parameters depend on the local image content so that: $\mathrm{p}(\mathbf{x}) = \mathrm{p}(\mathbf{x}; \boldsymbol{\alpha}) = \prod_{i=1}^{n} \mathrm{p}(x_i; \alpha_i)$, where hyper-parameters $\boldsymbol{\alpha} \in \mathbb{R}^N$ are now stored in a vector. When assuming a Laplace prior on $\|\mathbf{D}\mathbf{x}\|$, this choice thus corresponds to consider for all $i = 1, \ldots, N$

$$\mathrm{p}(\|(\mathbf{D}\mathbf{x})_i\|; \alpha_i) = \alpha_i \exp\left(-\alpha_i \|(\mathbf{D}\mathbf{x})_i\|\right), \quad (3)$$

where $\alpha_i > 0$ is the local scale parameter.

MAP estimation and space-variant modelling. An analogous formulation of problem (1) corresponds to perform standard Maximum A Posteriori (MAP) estimation in (2) by taking the negative logarithm and neglecting the normalization constant $\mathrm{p}(\mathbf{y})$, thus obtaining:

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}} \ -\ln \mathrm{p}(\mathbf{y}|\mathbf{x}) - \ln \mathrm{p}(\mathbf{x}).$$

Assuming, for instance, that $\mathrm{p}(\mathbf{y}|\mathbf{x}) = \mathcal{N}(0, \sigma^2\mathbf{Id})$ and that $\mathrm{p}(\mathbf{x})$ is chosen as in (3), standard calculations show that the problem becomes:

$$\arg\min_{\mathbf{x}} \ \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \mathrm{WTV}(\mathbf{x}) \quad \text{(HWTV-}\ell_2\text{)}$$

where $\mu := 1/\sigma^2 > 0$ and $\mathrm{WTV}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \|(\mathbf{D}\mathbf{x})_i\|$ is the space-variant version of TV which has been considered in previous works (see, e.g., [1, 3]). By dividing both terms above by μ, one can rewrite (HWTV-ℓ_2) as:

$$\arg\min_{\mathbf{x}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \sum_{i=1}^{N} \lambda_i \|(\mathbf{D}\mathbf{x})_i\|, \quad \text{(WTV-}\ell_2\text{)}$$

where, for each $i = 1, \ldots, N$ the parameter $\lambda_i := \alpha_i/\mu$ now depends now on both the hyperparameters associated to $\mathrm{p}(\mathbf{x})$ and to $\mathrm{p}(\mathbf{y}|\mathbf{x})$. Note, however, that differently from its space-invariant version (where $\lambda_i = \lambda = \alpha/\mu$ for all i), the model in (WTV-ℓ_2) is adaptive as the parameters λ_i can be adjusted to describe local image content. Their choice is crucial for obtaining a good denoising result as, locally, they balance the effect of the regularisations against the one of the data term.

Starting from formulation (WTV-ℓ_2), we propose in the following a deep-learning strategy based on the use of CNNs for estimating the vector $\boldsymbol{\lambda} \in \mathbb{R}^N$ from a given noisy observation $\boldsymbol{y} \in \mathbb{R}^N$. Alternative strategies previously considered in the standard literature of statistical/variational problems in imaging, rather considered the parameter estimation problem in the decoupled form (HWTV-ℓ_2) and estimated the parameters μ and $\boldsymbol{\alpha} = (\alpha_i)_i$ separately. For instance, in [1] a hybrid strategy for estimating μ via the standard discrepancy principle was combined with a maximum-likelihood approach for estimating the parameters $\boldsymbol{\alpha}$ for model (HWTV-ℓ_2) by exploiting the non-stationary form (3). Note that in order to provide suitable estimations of μ, this approach requires the prior knowledge (or estimation) of the noise distribution (Gaussian) and of its intensity value σ^2, which may be limiting in real-world applications.

From an optimization point of view, we solve problem (WTV-ℓ_2) by smoothing the WTV term with a parameter $0 < \varepsilon \ll 1$, thus considering for all $i = 1, \ldots, N$ the smoothed quantities $\|(\mathbf{D}\mathbf{x})_i\|_\varepsilon = \sqrt{(\mathbf{D}\mathbf{x})_{i,1}^2 + (\mathbf{D}\mathbf{x})_{i,2}^2 + \varepsilon}$, where $(\mathbf{D}\mathbf{x})_{i,1}$ and $(\mathbf{D}\mathbf{x})_{i,2}$ denote the gradient components along the horizontal and vertical direction, respectively. Upon such smoothing, a numerical solution of (WTV-ℓ_2) can be computed by standard gradient descent.

3. PARAMETER ESTIMATION VIA CNN LEARNING

We consider a training set composed of image patches of size 32×32, see Section 4.1 for more details on the dataset employed. For each image patch $\tilde{\mathbf{y}} \in \mathbb{R}^{32\times 32}$, we design a CNN with a light architecture to estimate an optimal parameter $\lambda > 0$. The network architecture is given in Figure 1 where the number of layers has been optimized w.r.t. to the mean square error. It consists simply of a convolutional layer, a max-pooling layer, a flattening layer and 2 dense layers. By tuning the model hyper-parameters, the following choices were made: batch size of 150 and a learning rate of 0.001. Concerning the number of epochs, we used early stopping to stop training once the model performance decreases on the validation dataset. Note that we have only to train three layers, one convolutional layer and two dense layers. Besides, the input matrix has a reduced dimension that is 32×32 independently of the initial image size. This is major advantage compared to end-to-end denoising CNNs based on the initial image as input.

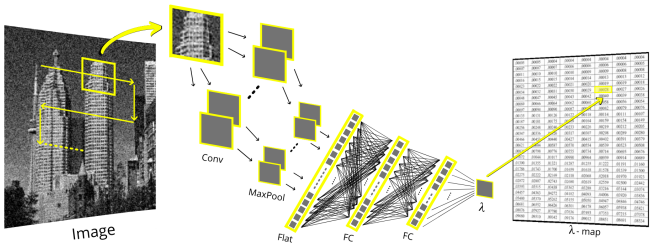


Fig. 1: CNN architecture for the estimation of the optimal parameter λ_j from a given image patch $\tilde{\mathbf{y}}_j \in \mathbb{R}^{32\times 32}$ extracted from the noisy image $\mathbf{y}$. The estimated parameters are collected on the λ-map $\boldsymbol{\lambda}_{\text{CNN}}$ containing at each pixel $j = 1, \dots, N$ the parameter λ_j .

Once the training is performed, for a new given noisy image $\mathbf{y} \in \mathbb{R}^N$, we extract image patches centered on each pixel $(\mathbf{y}_j)_{j=1}^D$ of size 32×32 by sliding a window over the whole image domain. The parameter estimation is then performed as above on each patch $\mathbf{y}_j \in \mathbb{R}^{32\times 32}$, providing a value $\lambda_j > 0$ as an output. The values λ_j are then collected in the array $\boldsymbol{\lambda}_{\text{CNN}} \in \mathbb{R}^N$, where, note, the values λ_j around the image boundary are computed by extrapolation. The estimation is performed independently on each patch, which allows for a parallel implementation. Having constructed the vector $\boldsymbol{\lambda}_{\text{CNN}}$, the denoising of the image $\mathbf{y}$ can thus be by solving (WTV-ℓ_2) using the estimated parameter map.

4. NUMERICAL RESULTS

4.1. Datasets construction

To develop a general model well-adapted to natural images with different image contents, textures and corrupted with different noise distributions, we considered a selection of 25 grayscale normalized images extracted from the Berkeley Segmentation Dataset[1]. Ground-truth images were corrupted with noise of different distributions (additive white Gaussian with variance $\sigma^2 \in \{0.02, 0.05, 0.07\}$ speckle noise with same variances and Poisson). For each of these images, we extracted 32×32 image patches $\mathbf{y} \in \mathbb{R}^{32\times 32}$ and compute the scalar optimal parameters $\hat{\lambda} > 0$ by brute-force estimation, i.e. we solve problem (WTV-ℓ_2) for a range of constant parameters $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, with $\lambda_{\min} < \lambda_{\max}$ and select as $\hat{\lambda}$ the value such that:

$$\hat{\lambda} \in \underset{\lambda\in[\lambda_{\min},\lambda_{\max}]}{\arg\max} \ \text{SSIM}(\mathbf{x}(\lambda), \tilde{\mathbf{x}}), \tag{4}$$

[1] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

where $\tilde{\mathbf{x}}$ is the ground-truth noise-free version of $\mathbf{y}$.

To mitigate the intensity bias due to Poisson noise corruption, we enriched our dataset with the set of noisy images with reverse intensities (i.e. negatives) along with the corresponding estimated parameters. The final training dataset is thus composed of $K = 19200$ patches of size 32×32 paired with the corresponding optimal value of the regularization parameter $\left\{(\mathbf{y}_j, \hat{\lambda}_j)\right\}_{j=1}^K$.

4.2. Denoising results and comparisons

Patch			
λ_{BF}	0.01	0.06	0.135
SSIM	0.68	0.73	0.57
λ_{CNN}	0.007	0.05	0.17
SSIM	0.68	0.71	0.56

Table 1: Comparisons between optimal estimated parameters λ_{BF} and λ_{CNN} on exemplar image patches via (4) and via the proposed CNN model, respectively.

In Table1, we compare the value λ_{BF} estimated as in (4) with the value λ_{CNN} estimated by the CNN model for three exemplar image patches. The first two patches are corrupted with Gaussian noise with zero mean and variance $\sigma^2 = 0.005$. The first one contains textured details, while the second one contains an extended homogeneous region. The third patch has the same content as the second one, but it is corrupted with Gaussian noise with lower variance $\sigma^2 = 0.02$. The two methods show comparable SSIM values. Note also that, as expected, smaller regularization is employed whenever textured details are present, while a larger smoothing is estimated in the case of stronger noise.

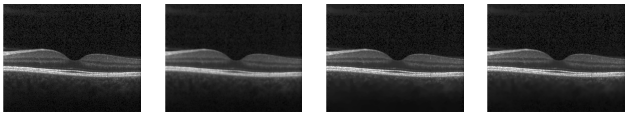
In Figure 2 we report the results by applying the proposed parameter estimation approach on the noisy image $\mathbf{y}$ in Figure 2a characterized by the presence of both homogeneous (e.g. sky) and textured (e.g. skyscraper) regions and corrupted with Gaussian noise of zero mean and variance $\sigma^2 = 0.01$. We compare in Figure 2b the denoising results obtained by using a standard TV-ℓ_2 denoising model (i.e. (WTV-ℓ_2) with a space-invariant parameter λ estimated as in (4), in Figure 2c the result obtained by applying the iterative approach considered in [1] for estimating in (HWTV-ℓ_2) both the parameter $\mu > 0$ by discrepancy principle and the parameters $\boldsymbol{\alpha}_{\text{ML}} = (\alpha_i)_i$ by maximum-likelihood (on a 32×32 window), and in Figure 2d the result obtained by solving (WTV-ℓ_2) where the parameters $\boldsymbol{\lambda}_{\text{CNN}} = (\lambda_i)_i$ are estimated as described above using a first CNN model learned only with Gaussian and Poisson noise. SSIM values w.r.t. to the ground-truth image are reported for comparisons. Generally speaking, we observe that compared to a scalar (i.e. global) parameter selection, allowing a local adjustment of the TV smoothing favors better

(a) Noisy image $\mathbf{y}$ (b) TV-ℓ_2 with estimated λ (0.767) (brute-force) (c) (HWTV-ℓ_2) with estimated $\mu, \boldsymbol{\alpha}_{\text{ML}}$ (0.797) (σ^2 required) (d) WTV-ℓ_2 with estimated $\boldsymbol{\lambda}_{\text{CNN}}$ (0.788) (automatic)

Fig. 2: Figure 2a: noisy image $\mathbf{y}$ corrupted with Gaussian noise with zero mean and variance $\sigma^2 = 0.01$. Figure 2b: TV-ℓ_2 denoising result with $\boldsymbol{\lambda} \equiv \lambda = 0.07$ optimizing SSIM by brute-force; Figure 2c: WTV-ℓ_2 denoising result with space-variant $\boldsymbol{\lambda}_{\text{ML}}$ estimated as in [1]; Figure 2d: TV-ℓ_2 denoising result with space-variant $\boldsymbol{\lambda}_{\text{CNN}}$ estimated by CNN learning. SSIM values are reported in brackets.

detail preservation. In terms of restoration quality, we see that the SSIM value of the denoised image in Figure 2c is slightly better than the one obtained by our approach. However, we highlight that while the approach in [1] relies on the prior knowledge of the noise distribution and intensity (i.e. the values σ^2 or its estimation), our approach does not require any assumption on the type of noise nor on the noise variance value as what is learned by the proposed CNN network is at each pixel the product $\lambda_i = \alpha_i \sigma^2$.

To test the proposed estimation strategy on a real-world problem where the noise distribution is unknown, we consider Optical Coherence Tomography (OCT) human data (see Figure 3) [2] The noise observed in OCT measurements is hardly Gaussian, as it is typically assumed to be signal-dependent and/or speckle-type. Due to the local adaptivity of our estimation approach to both local image content and noise intensity, we nonetheless considered models (HWTV-ℓ_2) coupled with the discrepancy- and Maximum-Likelihood-based approach [1], and (WTV-ℓ_2) with the proposed CNN-based parameter estimation strategy to denoise it. We observe that using the computed values $\boldsymbol{\lambda}_{\text{CNN}}$ allows for an improved image smoothing and to the reduction of noise artifacts in the background region.



(a) $\mathbf{y}$ (b) SB [6] (c) ML [1] (d) Proposed

Fig. 3: Denoising of an OCT image corrupted by speckle noise: comparison between state-of-the art statistical-based (SB) OCT-adapted denoising [6], solution of (HWTV-ℓ_2) with ML parameter estimation and proposed WTV-ℓ_2 denoising with estimated $\boldsymbol{\lambda}_{\text{CNN}}$.

[2] Kermany, D., Zhang, Kang Goldbaum, M., *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*, Mendeley Data, V2, 2018. Dataset: `https://www.kaggle.com/paultimothymooney/kermany2018`.

5. CONCLUSIONS

We proposed a hybrid approach combining a model-based image denoising model with a CNN strategy for hyperparameter learning. As the selected weighted-TV model embeds prior information on the desired solution, a small training set and a light CNN architecture can be used. The use of CNN learning avoids the requirement of a prior noise model and intensity for the hyperparameter estimation step. The proposed approach is thus very versatile as it can be used for general noise distributions. Compared to statistically-based state-of-the-art approaches we obtain comparable, although slightly lower, results in case of an additive Gaussian noise, with the advantage of avoiding the prior knowledge of the noise variance. Besides, our CNN-based approach better generalizes to non Gaussian noise as exemplified on real OCT images. A natural generalization is to extend the proposed approach to the case of deblurring, and estimate both the regularisation hyperparameters and the convolution kernel at the same time.

6. REFERENCES

[1] L. Calatroni, A. Lanza, M. Pragliola, and F. Sgallari, "Adaptive parameter selection for weighted-TV image reconstruction problems," in *J.Phys.: Conf. Series, NCMIP 2019*, 2020, vol. 1476, pp. 541–547.

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

[3] M. Pragliola, L. Calatroni, A. Lanza, and F. Sgallari, "On and beyond Total Variation regularisation in imaging: the role of space variance," 2021, arXiv preprint 2104.03650.

[4] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Advances in Neural Information Processing Systems*. 2009, vol. 22, Curran Associates, Inc.

[5] X. Descombes, M. Sigelle, and F. Preteux, "Estimating Gaussian Markov random field parameters in a nonstationary framework: application to remote sensing imaging," *IEEE Trans. Image Proc.*, vol. 8, no. 4, pp. 490–503, 1999.

[6] Muxingzi Li, Ramzi Idoughi, Biswarup Choudhury, and Wolfgang Heidrich, "Statistical model for OCT image denoising," *Biomed. Opt. Express*, vol. 8, no. 9, pp. 3903–3917, 2017.