CLASS-SPECIFIC IMAGE DENOISING USING IMPORTANCE SAMPLING

Milad Niknejad, José M. Bioucas-Dias, Mário A. T. Figueiredo

Instituto de Telecomunicações, and
Instituto Superior Técnico, Universidade de Lisboa, Portugal

ABSTRACT

In this paper, we propose a new image denoising method, tailored to specific classes of images, assuming that a dataset of clean images of the same class is available. Similarly to the *non-local means* (NLM) algorithm, the proposed method computes a weighted average of non-local patches, which we interpret under the importance sampling framework. This viewpoint introduces flexibility regarding the adopted priors, the noise statistics, and the computation of Bayesian estimates. The importance sampling viewpoint is exploited to approximate the *minimum mean squared error* (MMSE) patch estimates, using the true underlying prior on image patches. The estimates thus obtained converge to the true MMSE estimates, as the number of samples approaches infinity. Experimental results provide evidence that the proposed denoiser outperforms the state-of-the-art in the specific classes of face and text images.

Index Terms— Patch-based image denoising, class-adapted denoising, non-local means, minimum mean squared error, importance sampling.

1. INTRODUCTION

Image denoising is one of the classical and fundamental problems in image processing and computer vision. In the past decade, the state-of-the-art has been dominated by patch-based methods, not only in image denoising, but also in more general inverse problems. In some approaches (called *internal*), the image is denoised using information exclusively extracted from the noisy image. For example, denoising is carried out by averaging similar patches (as in NLM [1]), by collaboratively filtering sets of similar patches (as in BM3D [2]), by learning a *Gaussian mixture model* (GMM) from the noisy image and then using it a prior to obtain MMSE patch estimates [3], or by obtaining *maximum a posteriori* (MAP) patch estimates using a Gaussian prior estimated from a collection a similar patches [4]. The so-called *external* methods take advantage of a dataset of clean image patches, which can be used in different ways: to denoise each noisy patch by computing weighted averages of clean patches [5]; to learn a prior for clean patches (*e.g.*, a GMM [6]), which is subsequently used to denoise the noisy patches [7]. Hybrid external/internal methods have also been proposed [8].

Let $\mathbf{y} = \mathbf{x} + \mathbf{n} \in \mathbb{R}^p$, where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is additive Gaussian noise of variance σ^2, denote a noisy observed image patch and $\mathbf{x} \in \mathbb{R}^p$ the corresponding clean patch. Some external image denoising methods estimate $\mathbf{x}$ via the non-parametric weighted average

$$\widehat{\mathbf{x}} = \frac{\sum_{j=1}^{n} w_j \mathbf{z}_j}{\sum_{j=1}^{n} w_j}, \tag{1}$$

where $\{\mathbf{z}_j, j = 1, .., n\}$ is a set of clean patches selected from an external dataset, $w_j = \exp(-\frac{1}{2\sigma^2}||\mathbf{y}_i - \mathbf{z}_j||_2^2)$. If $\{\mathbf{z}_j, j = 1, .., n\}$ is a set of samples from a prior p_X, then

$$\lim_{n\to\infty} \widehat{\mathbf{x}} = \mathbb{E}[\mathbf{x}|\mathbf{y}],$$

i.e., as $n \to \infty$, $\widehat{\mathbf{x}}$ converges to the MMSE estimate of $\mathbf{x}$ under that prior [9]. However, computing (1) using all the patches in some large external dataset is computationally very demanding. In order to mitigate this computational hurdle, k *nearest neighbours* (k-NN) clustering has been used [5] to find similar patches and thus to reduce the number of patches averaged in (1). However, given that the clustering is performed on noisy patches, its quality is often questionable.

The success of patch-based denoising methods relies crucially on the suitability of the priors used. Although several studies provide evidence that leptokurtic multivariate distributions are a good fit to image patches [10], those densities have seldom been used for denoising, due to algorithmic hurdles raised by the learning procedure and posterior inference.

In many applications, the noisy image is known to belong to a specific class, such as text, face, or fingerprints, and this knowledge should be exploited by the denoising method. One approach to implement this idea is to use an external method, based on a dataset of clean images from the specific class in hand, rather than a general-purpose dataset of natural images [11, 12]. The obvious rationale is that more similar patches can be found in the external set of images from the same class than in a generic dataset, and the statistical properties of the patches derived from the class-specific dataset are also better adapted to the underlying clean image.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement 607290 (SpaRTaN).

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

In this paper, we first show that the non-parametric formula in (1) can be derived from the *importance sampling* (IS) framework [13], which is a method of the Monte-Carlo family [14]. Then, based on the IS perspective, we propose an image denoising method using class-specific external datasets, with two stages: in the first stage, a set of multivariate *generalized Gaussian* (GG) distributions is learned from the external clean patches; then, noisy patches are denoised by first assigning each to one of the learned GG distributions, and then approximating the MMSE estimate via (1). Under the IS framework, the MMSE patch estimates are approximated by sampling directly from the patches from which the GG distributions were estimated. The obtained results show that the proposed method outperforms other state-of-the-art general and class-specific denoisers.

In the following sections, we first describe the IS viewpoint for (1). Then, the proposed method for class-specific image denoising is described. Finally, experimental comparisons with the state-of-the-art methods are conducted.

2. IMPORTANCE SAMPLING

A fundamental step in a patch-based denoising algorithm is the estimation of the clean patches from the noisy ones. A classical result in Bayesian point estimation is that the MMSE estimate is given by the posterior expectation [15]:

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = \int \mathbf{x}\, p_{X|Y}(\mathbf{x}|\mathbf{y})\; d\mathbf{x} = \int \mathbf{x}\, \frac{p_{Y|X}(\mathbf{y}|\mathbf{x})\; p_X(\mathbf{x})}{p_Y(\mathbf{y})}\; d\mathbf{x}, \tag{2}$$

where p_X is the prior and $p_{Y|X}$ is the likelihood function, and the second equality is simply Bayes' law. Computing (2) is usually far from trivial, except when a conjugate prior is used [15]; a famous example is the Gaussian (or mixture of Gaussians) prior with a Gaussian likelihood, for which the posterior expectation has a simple closed-form.

One way to approximate (2) is to simply average random samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim p_{X|Y}$. However, sampling from $p_{X|Y}$ may not be a simple task. In particular, its normalization constant $p_Y(\mathbf{y})$ is itself hard (or impossible) to compute, as it is itself an integral that is intractable for arbitrary priors.

One way to circumvent the difficulty in sampling from $p_{X|Y}$ is to resort to *importance sampling* (IS) [13, 14]. By invoking the law of large numbers, $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ can be approximated by averaging $\mathbf{x}_i \frac{p_{Y|X}(\mathbf{y}|\mathbf{x}_i)}{p_Y(\mathbf{y})}$ using random samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim p_X$. Since the marginal density $p_Y(\mathbf{y})$ is still unknown, we may resort to the so-called *self-normalized IS* (SNIS), which does not require knowledge of the normalization constants of target density $p_{X|Y}$ [13, 16]:

$$\widehat{\mathbb{E}}_n[\mathbf{x}|\mathbf{y}] = \frac{\displaystyle\sum_{j=1}^{n} \mathbf{x}_j\, p_{Y|X}(\mathbf{y}|\mathbf{x}_j)}{\displaystyle\sum_{j=1}^{n} p_Y(\mathbf{y}|\mathbf{x}_j)}, \tag{3}$$

where $\mathbf{x}_1, ..., \mathbf{x}_n$ is a set of independent samples drawn from p_X. It can be shown that $\lim_{n\to\infty} \widehat{\mathbb{E}}_n[\mathbf{x}|\mathbf{y}] = \mathbb{E}[\mathbf{x}|\mathbf{y}]$ [13].

Notice that (1) is formally equivalent to (3), as long as

$$p_{Y|X}(\mathbf{y}|\mathbf{x}) \propto \exp(-\tfrac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|_2^2),$$

i.e., if the noise is zero-mean Gaussian with variance σ^2, and the set $\{\mathbf{z}_j, j = 1, .., n\}$ in (1) contains samples from the prior p_X. A special case of this denoiser, which was used to obtain a lower bound that denoising algorithms can achieve [9], just averages the central pixel of the patch; this corresponds to replacing $\mathbf{x}_j$ with $x_{j,c}$ in both the left and right hand sides of (3), where $x_{j,c}$ denotes the central pixel of patch $\mathbf{x}_i$.

In [17], it was shown that, for a fixed number of samples n, the MSE of the estimator (3) for the central pixel is reduced if the variance of samples, given the noisy patch, is decreased. Since the use of patch samples from a clean dataset, with a given (but unknown) distribution, tends to have a large variance, n has to be large in order to obtain a good approximation of $\mathbb{E}[\mathbf{x}|\mathbf{y}]$, as reported in [9]. Aiming at reducing the sample size in (3), some authors use only a subset of k clean patches that are the most similar to the noisy patch [5]. However, because this approach compares a noisy patch with clean patches, this subset is not guaranteed to contain a proper set of similar and correlated patches.

In this paper, the large sample size hurdle is alleviated by sampling from clusters of clean patches obtained from class-specific datasets. The obtained clusters, based on GG densities, have low intra-cluster variance, which is equivalent to strong correlation among the samples in a given cluster.

3. PROPOSED METHOD

3.1. Learning patch priors

Learning image priors is an important step in many image denoising algorithms. In [7], a Markov random field is learned from a set of natural images whose potentials are modelled as a *product of experts* (PoE). In the EPLL approach [6], a mixture of multivariate Gaussians is learned from the clean patches in an external dataset. In [3, 18, 19], a mixture of Gaussians is learned from the patches of the noisy image. In this work, we fit a set of M multivariate GG densities to a set of clean patches of the class-specific external dataset. The GG density with parameters $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta\}$ has the form [20]

$$p_X(\mathbf{x}; \Theta) = \frac{\beta\Gamma(\frac{p}{2})}{\pi^{\frac{p}{2}}\Gamma(\frac{p}{2\beta})2^{\frac{p}{2\beta}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))^{\beta}}, \tag{4}$$

where $\beta > 0$ is the shape parameter, $\Gamma(.)$ represents the Gamma function, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix, respectively.

We take the following iterative procedure, after some initialization, and until some stopping criterion is satisfied:

1. cluster the clean patches using the *maximum likelihood* (ML) criterion
$$\widehat{m}_j = \underset{m\in\{1,\ldots,M\}}{\arg\max}\, p_X(\mathbf{x}_j|\hat{\Theta}_m),$$
where $\hat{\Theta}_m$ is the current estimate of the parameter vector of the m-th cluster;
2. update the parameter estimates
$$\widehat{\Theta}_m = \underset{\Theta_m}{\arg\max} \prod_{j:\widehat{m}_j=m} p_X(\mathbf{x}_j|\Theta_m),$$
for $m = 1,\ldots,M$, using the method presented in [20].

Notice that, although the above learning procedure might have some computational complexity, it needs to be applied just once, for a given class-specific image dataset.

3.2. Image denoising

In order to denoise the degraded image, each noisy patch is assigned to one of the distributions learned from the external dataset by computing

$$\widehat{m} = \arg\max_m p_Y(\mathbf{y}|m), \tag{5}$$

where $p_Y(\mathbf{y}|m)$ is the density of $\mathbf{y} = \mathbf{x} + \mathbf{n}$, under the prior $p_X(\mathbf{x};\widehat{\Theta}_m)$. If the cluster densities of the clean patches were Gaussian, then $p_Y(\cdot|m) \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \sigma^2\mathbf{I})$ [6]. However, since we are using GG densities, the densities $p_Y(\mathbf{y}|m)$ do not have a simple expression, making it impractical to compute the ML assignments (5). At this point, we take a pragmatic decision: we fit Gaussian densities to the clusters, and classify each noisy patch into one of the clusters via ML using these Gaussian approximations. Compared with GG densities, the Gaussian densities are a weaker fit; we remark, however, that they are used only for assigning the noisy patches to clusters, not in the clustering procedure itself, neither for computing the final patch estimates.

After determining the patch distribution, the MMSE patch estimates are obtained via sampling according to (1). Although it is possible to sample efficiently from a GG density [20], we follow an alternative approach: we use as samples in (1) a set of randomly chosen clean patches from the cluster of clean patches to which it is assigned as explained in the previous paragraph. In this way, we are sampling from the underlying patch distribution, rather than from any fitted parametric density. As already mentioned, the reason for clustering is to reduce the variance of the samples used in the importance sampling formula.

3.3. More improvements

Although the importance sampling viewpoint for the formula in (1) brings flexibility to the estimation of the clean patches, it also has well known shortcomings. One of them is the large variation of the importance weights w_j's, which is associated with samples of very low representativity; this shortcoming is known as degeneracy of the weights [21]. In order to alleviate it, we use the method recently proposed in [22], which simply applies the hard thresholding operator on the importance weights w_j's before computing the sums in (1). This prevents the samples with low values of w_j to contribute to the estimate $\widehat{\mathbf{x}}_i$, and, thus, reduces the variance of the weights.

Finally, the patches are returned the original position in the image and averaged in the overlapped pixels to reconstruct the whole image. In order to further improve the algorithm performance, in the denoising step, we implement two iterations of our algorithm with a boosted image as an input of the second iteration. Boosting is a known strategy, which brings back the image details missing after the first step of denoising and that has been often used in image denoising (*e.g.*, [23, 24]). In this strategy, denoting the image obtained in the first stage by $\mathbf{X}^{(1)}$, the boosted image is obtained by $\mathbf{Y}^{(1)} = \mathbf{X}^{(1)} + r(\mathbf{Y} - \mathbf{X}^{(1)})$, where $r < 1$ is a constant. The noise level for the second iteration is computed as $\sigma_2^2 = \sigma^2 - \frac{1}{N^2}||\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}||_F^2$, which is exactly the same formula used in [24] for the same purpose.

4. EXPERIMENTAL RESULTS

In this section, we compare the proposed method with other internal and external image denoising methods for text and face images. We take BM3D as benchmark for internal image denoising methods. For a fair comparison, EPLL is trained with image datasets from the same class, and it is here called class-adapted EPLL. We also consider the state-of-the-art denoiser proposed by Luo et. al [11], which was specifically designed for class-specific datasets. To compare with methods using the non-parametric formula (1), we consider the external non-local means denoiser [11].

Regarding the setting of the parameters, in the prior learning step, the initialization is obtained by the k-means algorithm with 20 clusters. The parameter β in GG was empirically set to 0.9. The number of patch samples used in (3) was set to $n = 500$. The threshold for the importance weights was set to 5×10^{-60}.

The Gore face image dataset [25] was used as the face image dataset. For the text dataset, we extracted images from the different parts of a text document with different font sizes. This way, we considered both high quality (low variance) and low-quality (high variance) image datasets. In all the experiments, 5 images of the respective dataset were randomly selected for test and the remaining ones for training. Each test image was contaminated with white additive Gaussian noise,

Table 1: Denoising results (average PSNR over 5 test images) for the Gore face dataset [25] and the text dataset.

	$\sigma = 20$		$\sigma = 30$		$\sigma = 40$		$\sigma = 50$	
	face	text	face	text	face	text	face	text
BM3D	31.88	28.13	29.64	24.95	27.57	22.55	26.98	20.91
EPLL (generic)	31.66	28.15	29.43	25.21	27.74	23.15	26.58	21.72
Class adapted EPLL	32.34	29.23	30.16	26.39	28.49	24.40	27.28	23.01
External Non-local means	31.81	25.93	30.08	25.27	28.75	23.90	27.48	22.50
Luo et. al. [11]	32.98	27.52	30.89	27.44	29.24	26.29	28.01	25.02
Proposed method (Gaussian)	32.63	30.12	30.79	27.81	29.11	26.41	27.75	25.22
Proposed method (GG)	**33.09**	**30.93**	**30.99**	**28.20**	**29.48**	**26.75**	**28.08**	**25.79**

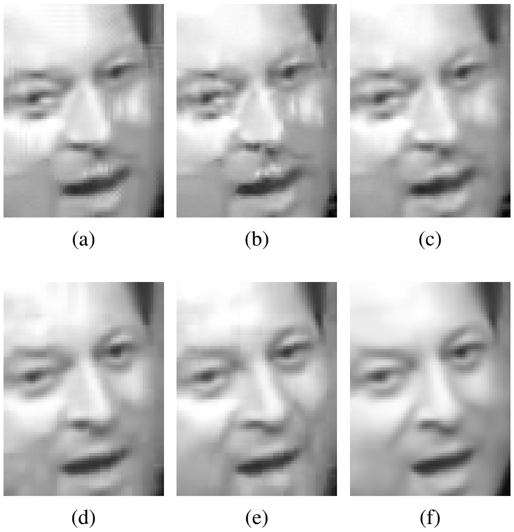

Fig. 1: An example of Denoising for a face image in the Gore dataset ($\sigma = 30$): (a) BM3D (PSNR=29.46) (b) EPLL (PSNR=28.97) (c) Class specific EPLL (PSNR=29.91); (d) External non-local means (PSNR=31.97) (e) Luo et. al. (PSNR=32.20) [11]; This work (PSNR=33.02).

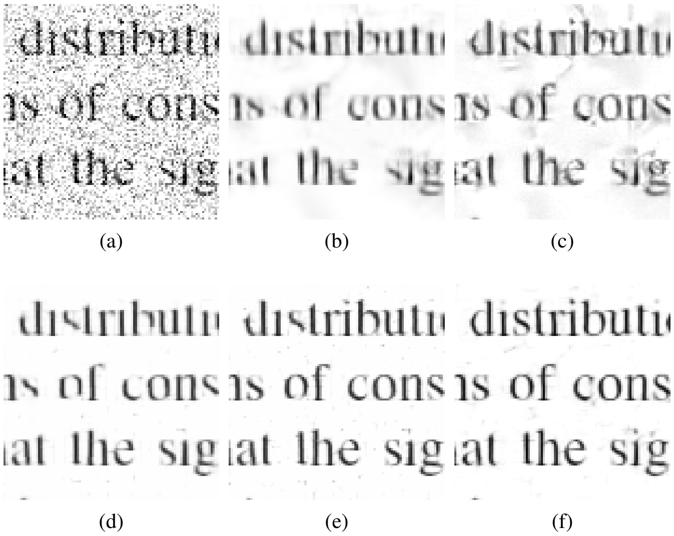

Fig. 2: An example of Denoising for a part of text image (a) Noisy ($\sigma = 60$) (b) BM3D (PSNR=20.14) (c) Class specific EPLL (PSNR=20.85); (d) External non-local means (PSNR=21.79) (e) Luo et. al. (PSNR=24.12) [11]; This work (PSNR=25.44).

and then denoised using the methods described before. Table 1 reports average PSNR over the five restored images.

From those results we may extract three conclusions: a) the proposed method outperforms the competitors, although the advantage over the denoiser in [11] is small for face; b) the clustering using GG prior yields better denoising results than the Gaussian one; c) considering our method in the multivariate Gaussian case, the approximate solution to the exact distribution (our method) performs better than the exact solution to the approximate fitted distribution (EPLL method).

Two examples of denoised images in the mentioned two experiments are shown in Fig. 1 and Fig. 2.

5. CONCLUSION

In this paper, we propose importance sampling to approximate the MMSE estimates of clean patches in which the samples are drawn from datasets of clean images from the same class. The clean patches were clustered under the assumption that each cluster follows a generalized Gaussian distribution. The experimental results provide evidence that our method outperforms the state-of-the-art denoisers based on class-specific datasets. Considering other priors for image patch clustering, using the importance sampling for generic images or other noise distributions, and applying other improvements for importance sampling are subjects for future works.

6. REFERENCES

[1] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 2, pp. 60–65.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-d transform-domain collaborative filtering,” *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[3] A. Teodoro, M. Almeida, and M. Figueiredo, “Single-frame image denoising and inpainting using Gaussian mixtures,” in *4th International Conference on Pattern Recognition Applications and Methods*, 2015.

[4] M. Niknejad, H. Rabbani, and M. Babaie-Zadeh, “Image restoration using Gaussian mixture models with spatially constrained patch clustering,” *IEEE Transactions on Image Processing*, vol. 24, pp. 3624–3636, 2015.

[5] M. Zontak and M. Irani, “Internal statistics of a single natural image,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 977–984.

[6] D. Zoran and Y. Weiss, “From learning models of natural image patches to whole image restoration,” in *International Conference on Computer Vision (ICCV)*, 2011, pp. 479–486.

[7] S. Roth and M. Black, “Fields of experts: A framework for learning image priors,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 2, pp. 860–867.

[8] I. Mosseri, M. Zontak, and M. Irani, “Combining the power of internal and external denoising,” in *IEEE International Conference on Computational Photography*, 2013, pp. 1–9.

[9] A. Levin and B. Nadler, “Natural image denoising: Optimality and inherent bounds,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2833–2840.

[10] H. Gerhard, L. Theis, and M. Bethge, “Modeling natural image statistics,” in *Biologically-inspired Computer Vision: Fundamentals and Applications*. 2015, Wiley.

[11] E. Luo, S. Chan, and T. Nguyen, “Adaptive image denoising by targeted databases,” *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2167–2181, 2015.

[12] A. Teodoro, J. Bioucas-Dias, and M. Figueiredo, “Image restoration and reconstruction using variable splitting and class-adapted image priors,” in *IEEE International Conference on Image Processing*, 2016.

[13] T. Hesterberg, “Weighted average importance sampling and defensive mixture distributions,” *Technometrics*, vol. 37, no. 2, pp. 185–192, 1995.

[14] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 1999.

[15] C. Robert, *The Bayesian Choice: A Decision Theoretic Motivation*, Springer-Verlag, New York, 1994.

[16] A. Owen, *Monte Carlo theory, methods and examples*, 2013, available at `http://statweb.stanford.edu/~owen/mc/`.

[17] A. Levin, B. Nadler, F. Durand, and W. Freeman, “Patch complexity, finite pixel correlations and optimal denoising,” in *European Conference on Computer Vision*, 2012, pp. 73–86.

[18] 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*, vol. 21, no. 5, pp. 2481–2499, 2012.

[19] M. Niknejad, H. Rabbani, M. Babaie-Zadeh, and C. Jutten, “Image interpolation using gaussian mixture models with spatially constrained patch clustering,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1613–1617.

[20] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, “Parameter estimation for multivariate generalized Gaussian distributions,” *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5960–5971, 2013.

[21] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and computing*, vol. 10, pp. 197–208, 2000.

[22] E. Koblents and J. Míguez, “A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models,” *Statistics and Computing*, vol. 25, pp. 407–425, 2015.

[23] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, “An iterative regularization method for total variation-based image restoration,” *Multiscale Modeling & Simulation*, vol. 4, pp. 460–489, 2005.

[24] W. Dong, G. Shi, and X. Li, “Nonlocal image restoration with bilateral variance estimation: a low-rank approach,” *IEEE Transactions on Image Processing*, vol. 22, pp. 700–711, 2013.

[25] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, “RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2233–2246, 2012.