PATCH SIMILARITY UNDER NON GAUSSIAN NOISE

Charles-Alban Deledalle, Florence Tupin

Institut Telecom, Telecom ParisTech
CNRS LTCI
Paris, France

Loïc Denis

Observatoire de Lyon, CNRS CRAL
UCBL, ENS de Lyon, Université de Lyon
Lyon, France

ABSTRACT

Many tasks in computer vision require to match image parts. While higher-level methods consider image features such as edges or robust descriptors, low-level approaches compare groups of pixels (patches) and provide dense matching. Patch similarity is a key ingredient to many techniques for image registration, stereo-vision, change detection or denoising. A fundamental difficulty when comparing two patches from "real" data is to decide whether the differences should be ascribed to noise or intrinsic dissimilarity. Gaussian noise assumption leads to the classical definition of patch similarity based on the squared intensity differences. When the noise departs from the Gaussian distribution, several similarity criteria have been proposed in the literature. We review seven of those criteria taken from the fields of image processing, detection theory and machine learning. We discuss their theoretical grounding and provide a numerical comparison of their performance under Gamma and Poisson noises.

Index Terms— Patch similarity, Likelihood ratio, Bayesian approach, Detection, Matching

1. INTRODUCTION

The similarity or dissimilarity between pixel values has been defined in many different ways, depending on the problem at hand (stereo-vision, registration, denoising, ...), the noise model and the *prior* knowledge. We focus in the following on how to compare noisy values, and how similarity criteria can be derived from a given noise distribution. The comparison of noise-free patches and the similarity between a noise-free and a noisy patch (template matching) are out of the scope of the paper.

By $\boldsymbol{x}$ we denote a patch, i.e., a collection of N observations (pixel values). We do not specify here a shape for the patch but consider that the values in vector $\boldsymbol{x}$ are ordered so that when two patches $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are compared, values with identical index are in correspondence.

We assume that the noise can be modeled by a (known) distribution so that a noisy patch $\boldsymbol{x}$ is a realization of an N-dimensional random variable $\boldsymbol{X}$. The vector of parameters $\boldsymbol{\theta}$ of the pdf is referred in the following as the noise-free patch. We will consider in our experiments white noise, i.e., $p(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{k=1}^{N} p(x_k;\theta_k)$, even if the definitions of all criteria are general enough to deal with correlated noise.

Patch similarity: a pair of (noisy) patches $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is considered similar (i.e., in-match) when $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are realizations of independent random variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ following the same distribution (of pdf $p(.;\boldsymbol{\theta}_{12})$). The evaluation of the similarity between noisy patches can then be rephrased as the following hypothesis test (a parameter test):

$$\mathcal{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \equiv \boldsymbol{\theta}_{12} \qquad \text{(null hypothesis)}, \tag{1}$$

$$\mathcal{H}_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \qquad \text{(alternative hypothesis)}. \tag{2}$$

For a given similarity criterion c, the *probability of false alarm* (to decide $\mathcal{H}_1$ under $\mathcal{H}_0$) and the *probability of detection* (to decide $\mathcal{H}_1$ under $\mathcal{H}_1$) are defined as:

$$P_{FA} = \mathbb{P}(c(\boldsymbol{X}_1, \boldsymbol{X}_2) < \tau; \boldsymbol{\theta}_{12}, \mathcal{H}_0), \tag{3}$$

$$P_D = \mathbb{P}(c(\boldsymbol{X}_1, \boldsymbol{X}_2) < \tau; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathcal{H}_1). \tag{4}$$

Note that the inequality symbols are reversed compared to usual definitions since we consider detection of dissimilarities based on the similarity measure c.

According to Neyman-Pearson theorem, the optimal criterion, i.e., the criterion which maximizes P_D for any given P_{FA}, is the likelihood ratio (LR) test:

$$L(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12}, \mathcal{H}_0)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathcal{H}_1)}. \tag{5}$$

The application of the likelihood ratio test requires the knowledge of the parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_{12}$ (the noise-free patches) which, of course, are unavailable. Our problem is thus a *composite hypothesis problem*. A criterion maximizing P_D for all P_{FA} and all values of the unknown parameters is said *uniformly most powerful* (UMP). Kendall and Stuart (1979) showed that no UMP detector exist in general for our *composite hypothesis problem* [1], so that any criteria can be defeated by another one at a specific P_{FA}. The research of a universal similarity criterion is then futile. We address in the following the question of how different criteria behave on patches extracted from natural images.

2. PATCH SIMILARITY CRITERIA

2.1. Euclidean distance

The usual way to define the similarity between two noisy patches is to consider their Euclidean distance. The use of an exponential kernel of bandwidth $h > 0$ leads to the following similarity criterion:

$$\mathcal{N}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2}{h}\right). \tag{6}$$

The Mahalanobis distance can be used instead if noise is correlated. As we shall see later, under the assumption of Gaussian noise, all the similarity criteria we consider boil down to this same expression. There are then more than one way to justify or interpret its expression in that case. The advantage to use a metric is that it involves good properties such as the triangle inequality. Under Gaussian assumptions, the distribution of $\mathcal{N}$ can be used to choose a threshold τ with a given P_{FA} value. It is a *constant false alarm rate detector* (CFAR), which means that a constant P_{FA} can be maintained with a same τ whatever the underlying noise-free patches.

The performance of this criterion however falls when the noise departs from Gaussian distribution. While h can be set from the noise variance, difficulties arise when the noise variance is signal-dependent, and then can vary between and inside patches. A classical approach to extend the applicability of Euclidean distance to non-Gaussian noises is to apply a transformation to the noisy patches. The transformation is chosen so that the transformed patches follow a (close to) Gaussian distribution with constant variance (hence their name: variance-stabilization transforms). This leads for instance to the homomorphic approach which maps multiplicative noise to additive noise with stationary variance. This is also the principle of Anscombe transform and its variants used for Poisson noise. These approaches are popular and frequently used for patch selection (or block-matching) in many denoising algorithms [2, 3, 4]. Given an application s which stabilizes the variance for a specific noise pdf, a similarity criterion is obtained using (6) on the output of s:

$$S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathcal{N}(s(\boldsymbol{x}_1), s(\boldsymbol{x}_2)). \tag{7}$$

Besides the problem of the existence of a suitable s for some noise distributions, an important limitation lies in the nonlinear distortion of noise-free patches. For instance, in the homomorphic approach, the logarithm transforms the contrast of noise-free patches; performances are affected accordingly.

2.2. Likelihood ratio extensions

Motivated by optimality guarantees of the LR test (5), similarity criteria can be defined from statistical detectors designed for *composite hypothesis problems*. The similarity criterion in eq. (8) is based on the Bayesian likelihood ratio (BLR)

$$L_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathcal{H}_0)}{p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathcal{H}_1)} = \frac{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_{12}=\boldsymbol{t})p(\boldsymbol{x}_2|\boldsymbol{\theta}_{12}=\boldsymbol{t})p(\boldsymbol{\theta}_{12}=\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}}{\int p(\boldsymbol{x}_1|\boldsymbol{\theta}_1=\boldsymbol{t}_1)p(\boldsymbol{\theta}_1=\boldsymbol{t}_1)\mathrm{d}\boldsymbol{t}_1 \int p(\boldsymbol{x}_2|\boldsymbol{\theta}_2=\boldsymbol{t}_2)p(\boldsymbol{\theta}_2=\boldsymbol{t}_2)\mathrm{d}\boldsymbol{t}_2}. \tag{8}$$

which considers noise-free patches as realizations of random vectors with known *prior* pdf. Given perfect knowledge of *prior* pdf $p(\boldsymbol{\theta}_1)$, $p(\boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_{12})$, eq. (8) leads to an optimal Neyman-Pearson detector. This criterion has been used in the context of classification: Minka [5] exhibits a relationship between BLR and the canonical distance measure minimizing errors in nearest neighborhood classifiers. He also linked BLR to mutual information: the more additional knowledge is brought by $\boldsymbol{x}_2$ compared to the observation of $\boldsymbol{x}_1$ alone, the more dissimilar the underlying parameters are [6].

Despite its theoretical performance, this approach suffers from two drawbacks in practice. First, it requires computation of integrals which, depending on the distributions, may not be known in closed form and therefore are time-consuming. Second, it requires knowledge of the *prior* pdf. In the absence of a statistical model of noise-free patches, a *non-informative prior* can be used. Jeffreys' *prior* is independent upon the choice of the noise-free patch space (e.g., testing that two gamma random vectors share identical standard deviations $\theta_{12,k} = \sigma_k$ or identical variances $\theta_{12,k} = \sigma_k^2$ leads to the same BLR when Jeffreys' *prior* are used).

Rather than modeling noise-free patches as random variables, the generalized LR (GLR) replaces $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_{12}$ in eq. (5) by their maximum likelihood estimates (MLE) under each hypothesis:

$$L_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\sup_{\boldsymbol{t}} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12} = \boldsymbol{t}, \mathcal{H}_0)}{\sup_{\boldsymbol{t}_1, \boldsymbol{t}_2} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_1 = \boldsymbol{t}_1, \boldsymbol{\theta}_2 = \boldsymbol{t}_2, \mathcal{H}_1)} = \frac{p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_{12})p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_{12})}{p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_1)p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_2)} \tag{9}$$

Asymptotically to the SNR, GLR is optimal due to the efficiency of MLE. Its asymptotic distribution is known and then the P_{FA} values associated to any given threshold τ: GLR is asymptotically CFAR. The GLR test is also invariant [7]: it does not depend on the arbitrary choice of the noisy patch space (e.g. considering an observed patch of amplitudes $\boldsymbol{x} = (A_1, \cdots, A_N)$ or intensities $\boldsymbol{x} = (A_1^2, \cdots, A_N^2)$ lead to the same criterion). While we mention that there is no UMP detectors for our composite hypothesis problem, GLR is asymptotically UMP among invariant tests [8]. Finally, compared to BLR, GLR is easy to implement, since it only requires to computes MLE, and does not require any *prior* knowledge.

The main drawback of GLR lies in the lack of knowledge on how it behaves in low SNR conditions (i.e., for too small patches according to the noise level). It is known that, for low SNR and specific applications, GLR can be defeated by other

invariant detectors [9]. This drawback lies in its dependency on MLE which behaves poorly for low SNR (e.g. the GLR that two Gaussian random vectors share an identical covariance matrix $\boldsymbol{\theta}_{12}$ is undefined since MLE of $\boldsymbol{\theta}_1$ from $\boldsymbol{x_1}$ only would not be positive definite).

2.3. Joint likelihood criteria

Other criteria use the joint likelihood of observations under $\mathcal{H}_0$ to evaluate similarities between noisy data. This leads to the Bayesian joint likelihood criteria [10, 11, 12, 13]:

$$\begin{aligned} Q_B(\boldsymbol{x}_1, \boldsymbol{x}_2) &= p(\boldsymbol{x}_1, \boldsymbol{x}_2; \mathcal{H}_0) \\ &= \int p(\boldsymbol{x}_1|\boldsymbol{\theta}_1 = \boldsymbol{t}) p(\boldsymbol{x}_2|\boldsymbol{\theta}_2 = \boldsymbol{t}) p(\boldsymbol{\theta}_{12} = \boldsymbol{t}) \, \mathrm{d}\boldsymbol{t} \end{aligned} \quad (10)$$

or, following the simplification of GLR, the maximum joint likelihood [14]:

$$\begin{aligned} Q_G(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \sup_{\boldsymbol{t}} p(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}_{12} = \boldsymbol{t}, \mathcal{H}_0) \\ &= p(\boldsymbol{x}_1; \boldsymbol{\theta}_1 = \hat{\boldsymbol{t}}_{12}) p(\boldsymbol{x}_2; \boldsymbol{\theta}_2 = \hat{\boldsymbol{t}}_{12}). \end{aligned} \quad (11)$$

Such criteria have been designed to measure the probability of sharing a common parameter. However, they evaluate instead the joint likelihood pdf under $\mathcal{H}_0$ which cannot provide information without knowledge under $\mathcal{H}_1$. This leads to non-invariance issues and the self recognition paradox [10]: two different noisy patches $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ can be more similar than two identical noisy patches ($\boldsymbol{x}_1 = \boldsymbol{x}_1$).

However, Q_B offers a useful property: it corresponds to an inner product [11] in the space of functions $\boldsymbol{\theta} \mapsto \mathbb{R}$, the feature of $\boldsymbol{x}$ being $(p(\boldsymbol{x}|\boldsymbol{\theta} = \boldsymbol{t}))_{\boldsymbol{t}}$. The "mutual information" kernel is based on this property.

2.4. Mutual information kernel

Given the Bayesian joint distribution $Q_B(\boldsymbol{x}_1, \boldsymbol{x}_2)$, Seeger [11] defines a covariance kernel linked to the sample mutual information between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and defined as:

$$K_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{Q_B(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{Q_B(\boldsymbol{x}_1, \boldsymbol{x}_1) Q_B(\boldsymbol{x}_2, \boldsymbol{x}_2)}}. \quad (12)$$

Since Q_B can be seen as an inner product in the feature space, K_B corresponds to a cosine in the feature space $K_B(\boldsymbol{x}_1, \boldsymbol{x}_2) = \dfrac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|}$. Seeger shows that it is a kernel covariance matrix and coins it the mutual information kernel. Algorithms can be adapted to the noise pdf using the so-called *kernel tricks*, i.e., by considering higher dimensional space while never mapping the data in practice. This leads for instance to non-linear support vector machines or non-linear principal component analysis. The mutual information kernel is also invariant. Note also that its *prior*-less extension using MLE would lead to GLR. Compared to GLR, the main limitation of the mutual information kernel is its dependency on the *prior* pdf and the lack of asymptotic performance results.

3. EVALUATION OF SIMILARITY CRITERIA

We evaluate the relative performances of the 7 aforementioned criteria on a dictionary composed of 196 noise-free patches of size $N = 8 \times 8$. The noise-free patches are obtained using the k-means on patches extracted from the classical 512×512 *Barbara* image. The noisy patches are noisy realizations of the noise-free patches under gamma or Poisson noise with an overall SNR of about 1 dB. All criteria have been derived[1] in the case of gamma or Poisson noise (table 1). They are evaluated for all pairs of noisy patches. The process is repeated 200 times with independent noise realizations.

In practice, Bayesian criteria are more difficult to obtain due to integrations over the noise-free patch space. While all criteria are equivalent for Gaussian noise, there are four different expressions for gamma noise and they are all different for Poisson noise. The distinction seems to emerge with the "complexity" induced by the noise distribution, (by considering that gamma noise is more challenging than Gaussian noise, and that Poisson noise is the most challenging).

Numerically, the performances of the similarity criteria are given in term of their *receiver operating characteristic* (ROC) curve, i.e., the curve of P_D with respect to P_{FA}. Results are given in Figure 1. For small P_{FA}, GLR is the most powerful followed by the mutual information kernel, BLR and the variance stabilization criteria. Other criteria behave poorly for such a low SNR. Such behaviors agree with the theoretical predictions. The poor performances of the joint likelihood based criteria can arise from their non-invariance and the induced self-similarity paradox. The low performance of $\mathcal{N}$ is certainly due to its non-adaptivity to either the target noise or the target noise variance. The variance stabilization criteria are always defeated by GLR, due to the distortions of the noise-free patches as well as the consideration of the noise variance only, instead of the full noise pdf. The lower performance of Bayesian criteria compared to criteria that use MLE may be due to the low quality of the *prior* pdf.

4. CONCLUSION

This paper compares seven similarity criteria designed for noisy data and used in different communities. It has been shown that on 8×8 patches extracted from a natural image and under a high level of gamma or Poisson noise, the GLR detector is the most powerful on low levels of false alarms. It is also easy to implement and theoretically well grounded. Based on this study, we would recommend a broader use of this criterion for measuring patch similarity in computer vision. Future work would be to provide comparisons on smaller patches where GLR is known to behave poorly. We also plan to study the impact of the choice of a similarity criterion on the performance of tasks such as stereo-matching or denoising.

[1] the complete derivations are available in `http://perso.telecom-paristech.fr/~deledall/patchsim.php`

name	pdf	Q_B	Q_G	L_B	L_G	K_B	S	$\mathcal{N}$
Gaussian	$\frac{e^{-\frac{(x-\theta)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$e^{-(x_1-x_2)^2}$						
Gamma	$\frac{L^L x^{L-1} e^{-\frac{Lx}{\theta}}}{\Gamma(L)\theta^L}$	$\frac{1}{x_1 x_2}\left(\frac{x_1 x_2}{(x_1+x_2)^2}\right)^L$		$\frac{x_1 x_2}{(x_1+x_2)^2}$			$e^{-\left(\log\frac{x_1}{x_2}\right)^2}$	
Poisson	$\frac{\theta^x e^{-\theta}}{x!}$	$\frac{\Gamma'(x_1+x_2)}{2^{x_1+x_2}x_1!x_2!}$	$\frac{(x_1+x_2)^{x_1+x_2}}{(2e)^{x_1+x_2}x_1!x_2!}$	$\frac{\Gamma'(x_1+x_2)}{2^{x_1+x_2}\Gamma'(x_1)\Gamma'(x_2)}$	$\frac{(x_1+x_2)^{x_1+x_2}}{2^{x_1+x_2}x_1^{x_1}x_2^{x_2}}$	$\frac{\Gamma'(x_1+x_2)}{\sqrt{\Gamma'(2x_1)\Gamma'(2x_2)}}$	$e^{-\left(\sqrt{x_1+a}-\sqrt{x_2+a}\right)^2}$	

Table 1. Instances of the seven criteria for Gaussian, gamma and Poisson noise (parameters σ and L are fixed and known). All Bayesian criteria are obtained with Jeffreys' *priors* (resp. $1/\sigma$, $\sqrt{L}/\theta$, $\sqrt{1/\theta}$). All constant terms which do not affect the detection performance are omitted. For clarity reason, we define $\Gamma'(x) = \Gamma(x + 0.5)$ and the Anscombe constant $a = 3/8$.

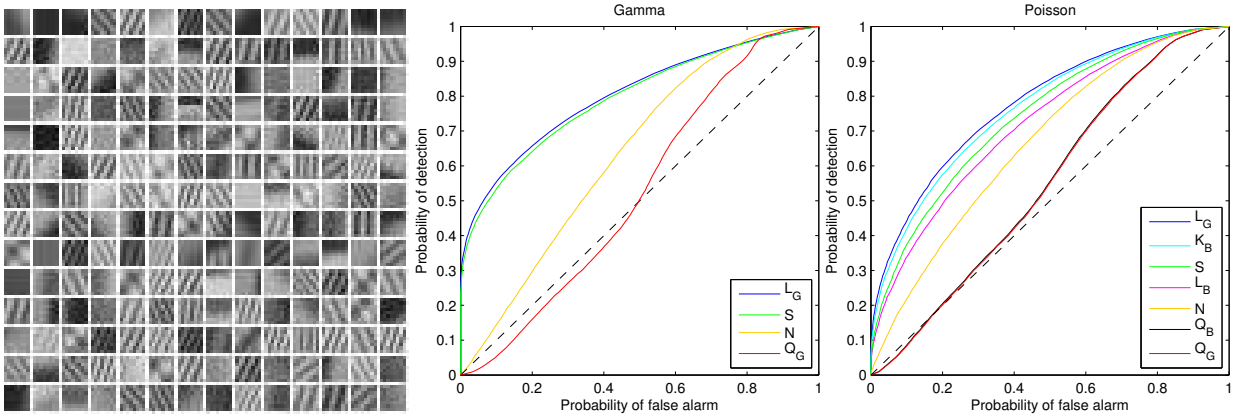

Fig. 1. (left) Patch dictionary. (center) ROC curve obtained under gamma noise and (right) ROC curve obtained under Poisson noise. In both experiments, the SNR over the whole dictionary is about 1 dB.

5. REFERENCES

[1] M. Kendall and A. Stuart, "The advanced theory of statistics. Vol. 2: Inference and relationship," 1979.

[2] M. Mäkitalo, A. Foi, D. Fevralev, and V. Lukin, "Denoising of single-look SAR images based on variance stabilization and nonlocal filters," in *ICMMET'10, Kiev, Ukraine*, 2010.

[3] M. Mäkitalo and A. Foi, "On the inversion of the Anscombe transformation in low-count Poisson image denoising," in *IEEE LNLA'2009, Tuusula, Finland*, 2009.

[4] J. Boulanger, C. Kervrann, P. Bouthemy, P. Elbau, J.B. Sibarita, and J. Salamero, "Patch-based nonlocal functional for denoising fluorescence microscopy image sequences," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 442–454, 2010.

[5] T.P. Minka, "Distance measures as prior probabilities," *Microsoft research report*, 2000.

[6] T.P. Minka, "Bayesian Inference, Entropy, and the Multinomial Distribution," *Microsoft research report*, 1998.

[7] S.M. Kay and J.R. Gabriel, "An invariance property of the generalized likelihood ratio test," *IEEE Signal Processing Letters*, vol. 10, no. 12, pp. 352–355, 2003.

[8] E.L. Lehmann, "Optimum invariant tests," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 881–884, 1959.

[9] H.S. Kim and A.O. Hero III, "Comparison of GLR and invariant detectors under structured clutter covariance," *IEEE TIP*, vol. 10, no. 10, pp. 1509–1520, 2001.

[10] P. Yianilos, "Metric learning via normal mixtures," *NEC Research Institute Technical Report*, 1995.

[11] M. Seeger, "Covariance kernels from Bayesian generative models," in *Advances in neural information processing systems*. MIT Press, 2002, p. 905.

[12] Y. Matsushita and S. Lin, "A Probabilistic Intensity Similarity Measure based on Noise Distributions," in *IEEE CVPR'07*, 2007, pp. 1–8.

[13] C.A. Deledalle, L. Denis, and F. Tupin, "Iterative Weighted Maximum Likelihood Denoising with Probabilistic Patch-Based Weights," *IEEE TIP*, vol. 18, no. 12, pp. 2661–2672, Dec. 2009.

[14] F. Alter, Y. Matsushita, and X. Tang, "An intensity similarity measure in low-light conditions," *Lecture Notes in Computer Science*, vol. 3954, pp. 267, 2006.