Bayesian Inference on Multiscale Models for Poisson Intensity Estimation: Applications to Photon-Limited Image Denoising

Stamatios Lefkimmiatis, *Student Member, IEEE*, Petros Maragos, *Fellow, IEEE*, and George Papandreou, *Student Member, IEEE*

***Abstract*—We present an improved statistical model for analyzing Poisson processes, with applications to photon-limited imaging. We build on previous work, adopting a multiscale representation of the Poisson process in which the ratios of the underlying Poisson intensities (rates) in adjacent scales are modeled as mixtures of conjugate parametric distributions. Our main contributions include: 1) a rigorous and robust regularized expectation-maximization (EM) algorithm for maximum-likelihood estimation of the rate-ratio density parameters directly from the noisy observed Poisson data (counts); 2) extension of the method to work under a multiscale hidden Markov tree model (HMT) which couples the mixture label assignments in consecutive scales, thus modeling interscale coefficient dependencies in the vicinity of image edges; 3) exploration of a 2-D recursive quad-tree image representation, involving Dirichlet-mixture rate-ratio densities, instead of the conventional separable binary-tree image representation involving beta-mixture rate-ratio densities; and 4) a novel multiscale image representation, which we term Poisson-Haar decomposition, that better models the image edge structure, thus yielding improved performance. Experimental results on standard images with artificially simulated Poisson noise and on real photon-limited images demonstrate the effectiveness of the proposed techniques.**

***Index Terms*—Bayesian inference, expectation-maximization (EM) algorithm, hidden Markov tree (HMT), photon-limited imaging, Poisson-Haar decomposition, Poisson processes.**

I. Introduction

PHOTON detection is the basis of image formation for a great number of imaging systems used in a variety of applications, including medical [1] and astronomical imaging [2]. In such systems, image acquisition is accomplished by counting photon detections at different spatial locations of a sensor, over a specified observation period. For low intensity levels, one of the dominant noise sources responsible for the degradation of the quality of the captured images is the so-called quantum or shot noise. Quantum noise [3] is due to fluctuations on the number of detected photons, an inherent limitation of the discrete nature of the detection process, and degrades such images both qualitatively and quantitatively. The resulting degradation can prove to be a major obstacle preventing image analysis and information extraction. Thus, the development of methods and techniques to alleviate the arising difficulties is of fundamental importance.

The basic photon-imaging model assumes that the number of detected photons at each pixel location is Poisson distributed. More specifically, under this model the captured image is considered as a realization of an inhomogeneous Poisson process. This Poisson process is characterized by a 2-D spatially varying rate function which equals the process mean and corresponds to the noise-free intensity image we want to recover. The variability of the counts about the mean can be interpreted as noise. Since the Poisson process variance equals the rate function/noise-free image, the noise appearing in the acquired image is spatially varying and signal-dependent. This restrains us from using a variety of well-studied methods and tools that have been developed for coping with additive homogeneous noise. From a statistical viewpoint, denoising the captured image corresponds to estimating the discretized underlying intensity from a single realization of the Poisson process. For a 2-D Poisson process the discretized intensity image can be represented as a 2-D array $\boldsymbol{\lambda}$, while the observed noisy image as a 2-D array $\mathbf{x}$. Then at each location (i, j) the observed counts $x(i, j)$ can be considered as a realization of the random variable[1] $X(i, j)$ which follows a Poisson distribution with intensity $\lambda(i, j)$.

A host of techniques have been proposed in the literature to account for the Poisson intensity estimation problem. A classical approach followed by many researchers includes preprocessing of the count data by a variance stabilizing transform (VST) such as the Anscombe [4] and the Fisz [5] transforms, applied either in the spatial [6] or in the wavelet domain [7]. The transformation reforms the data so that the noise approximately becomes Gaussian with a constant variance. Standard techniques for independent identically distributed (i.i.d.) Gaussian noise are then used for denoising. A recent method of this family is proposed in [8], where the authors first apply a new multiscale VST transform followed by conventional denoising

Manuscript received September 07, 2008; revised March 26, 2009. First published May 02, 2009; current version published July 10, 2009. This work was supported in part by the projects ΠΕΝΕΔ-2003 ΕΔ-554 & ΕΔ-865, which are co-financed by the E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%), and in part by the European FP6 FET research project ASPI (IST-FP6-021324). It was also supported by the Bodossaki Foundation and the Onassis Public Benefit Foundation through scholarships to SL and GP, respectively. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yongyi Yang.

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece (e-mail: sleukim@cs.ntua.gr; maragos@cs.ntua.gr; gpapan@cs.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2022008

[1]In this paper, we denote random variables with upper case letters and their realizations with lower case letters.

based on the curvelet [9] and ridgelet [10] transformations. An alternative approach includes wavelet-domain methods modified to account for the Poisson image statistics, thus avoiding the preprocessing of the count data. Kolaczyk in [11] has developed a shrinkage method for the Haar wavelet transform with "corrected" thresholds accounting for the specific characteristics of the Poisson noise, while Nowak and Baraniuk [12] proposed a wavelet-domain filtering approach based on a cross-validation estimator.

Another major category includes methods adopting a multiscale Bayesian framework specifically tailored for Poisson data, independently initiated by Timmerman and Nowak [13] and Kolaczyk [14]. One of the key advantages of Bayesian methods is that they allow incorporation of prior knowledge into the estimation procedure. In addition, Bayesian methods combined with multiscale analysis are becoming increasingly popular since they can significantly simplify the estimation problem. In particular, the framework of [13] and [14] involves a decomposition of the Poisson process likelihood function across scales, allowing computationally efficient intensity estimation by means of a scale-recursive scheme. Additionally, this multiscale likelihood decomposition is naturally accompanied by a factorized prior model for the image intensity, in which the ratios of the underlying Poisson intensities (rates) at adjacent scales are modeled as mixtures of conjugate parametric distributions. This prior model has been proven quite effective in image modeling.

Despite the strong potential of the multiscale Bayesian framework for analyzing Poisson processes, previous work in this area [13]–[16] has certain shortcomings which hinder its wider applicability. In this paper, we deal with the following problems in the aforementioned framework.

1) *Image representation*: Multiscale decomposition strategies for handling intrinsically 2-D image data.
2) *Parameter estimation*: Estimation of the prior model parameters directly from the observed Poisson noisy data to accurately match the statistics of the source image.
3) *Interscale dependencies*: Modeling interscale dependencies among intensity/rate ratios arising in natural images.

Our main contributions are: (1) Regarding the problem of image representation, besides the conventional separable binary-tree image representation involving beta-mixture rate-ratio densities, we explore a recursive quad-tree image representation, explicitly tailored to 2-D data, involving Dirichlet-mixture rate-ratio densities; a similar single-component Dirichlet quad-tree image representation was first studied by [17] in the context of image deconvolution. Further, we propose a novel directional multiscale image representation, termed Poisson-Haar decomposition due to its close relation with the 2-D Haar wavelet transform, which better captures the edge detail structure of images, thus providing improved results in image modeling and consequently in the intensity estimation problem. (2) Regarding the problem of parameter estimation, we propose an Expectation-Maximization (EM) technique for maximum-likelihood estimation of the prior distribution parameters directly from the observed Poisson data, so that the model accurately matches the statistics of the source image. The robustness of the technique is enhanced by incorporating an appropriate regularization term. This term is interpretable as specifying a conjugate prior for the density parameters to be estimated and leads to a maximum *a posteriori* (MAP) instead of the standard maximum likelihood (ML) estimation. The proposed EM-based method can be equally well applied to either the conventional separable model, the nonseparable quad-tree, or the proposed Poisson-Haar decomposition. A preliminary version of this method was first presented in [18]. (3) The statistical framework which treats each scale of analysis as independent can be unrealistic for many real-world applications. Hidden Markov tree (HMT) structures as those proposed in [19]–[22] can efficiently model interscale dependencies between rate-ratio density mixture assignments and, thus, are more appropriate. The parameter estimation issue is even more pronounced in this case due to the extra HMT-specific parameters that need to be fitted. This problem is addressed by extending our EM parameter estimation method to also cover the HMT case; this leads to further benefits in the intensity estimation problem. We experimentally validate the effectiveness of all proposed contributions by comparing our results with those of previous techniques on photon-limited versions of standard test images. We also apply the proposed methods on denoising photon-limited biomedical and astronomical images.

The paper is organized as follows. In Section II we discuss the multiscale Bayesian framework for Poisson intensity estimation. In this context we briefly review the image partitioning scheme of [13], [14], we explore the nonseparable quad-tree image partitioning and introduce the novel Poisson-Haar multiscale image representation. In Section III we derive Bayesian estimators for the discussed models, while in Sections IV and V we describe our EM-based parameter estimation techniques for the independent and HMT models, respectively. Experimental results and comparisons on photon-limited imaging applications are presented in Section VI.

II. Multiscale Modeling of Poisson Processes

A. Problem Formulation

To simplify the analysis we initially assume that the image of interest is a 1-D signal of length N, $\boldsymbol{\lambda} = [\lambda(0), \lambda(1), \ldots, \lambda(N-1)]$, where $\lambda(k)$, $k = 0, \ldots, N-1$ refers to each individual pixel. Estimation of the intensity image $\boldsymbol{\lambda}$ is based on the corresponding observed noisy photon counts $\mathbf{x} = [x(0), x(1), \ldots, x(N-1)]$. Under the basic photon-imaging model, the vector $\mathbf{x}$ consists of the observation samples of N random variables $X(k)$ which are conditionally independent upon $\boldsymbol{\lambda}$, and each one follows a Poisson distribution with rate parameter $\lambda(k)$, denoted by $X(k)|\lambda(k) \sim \mathrm{Pois}(\lambda(k))$. See Table I for definitions of the Poisson and other probability distributions used in this paper. In the Bayesian framework, $\boldsymbol{\lambda}$ is not considered any more as deterministic but instead is treated as a vector containing the samples of a random sequence $\boldsymbol{\Lambda}$, whose particular realization we must estimate. To obtain a Bayesian estimator for this problem we first have to choose an appropriate prior probability model, $p(\boldsymbol{\lambda})$, for the random sequence. After specifying the model, we can devise strategies to find a statistically optimal estimator. If we select the mean squared error (MSE) as the criterion to minimize, then we

TABLE I
PROBABILITY DISTRIBUTIONS USED IN THE PAPER. THE GAMMA FUNCTION IS DEFINED BY $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$, D IS THE DIMENSION OF THE RANDOM VECTORS, AND $n = \sum_{k=1}^{D} x_k$

Univariate Distributions	
Poisson	$\mathrm{Pois}\,(x\|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$
Gamma	$\mathcal{G}\,(\lambda\|\alpha,\beta) = \lambda^{\alpha-1}\frac{e^{-\lambda/\beta}}{\beta^\alpha\Gamma(\alpha)}$
Binomial	$\mathrm{Bin}\,(x\|n,\theta) = \binom{n}{x}\theta^x\,(1-\theta)^{n-x}$
Beta	$\mathrm{Beta}\,(\theta\|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}$
Multivariate Distributions	
Multinomial	$\mathrm{Mult}\,(\mathbf{x}\|n,\boldsymbol{\theta}) = \frac{n!}{\prod_{k=1}^{D} x_k!}\cdot\prod_{k=1}^{D}\theta_k^{x_k}$
Dirichlet	$\mathrm{Dir}\,(\boldsymbol{\theta}\|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{D}\alpha_k\right)}{\prod_{k=1}^{D}\Gamma(\alpha_k)}\cdot\prod_{k=1}^{D}\theta_k^{\alpha_k-1}$
Polya	$\mathrm{Polya}\,(\mathbf{x}\|n,\boldsymbol{\alpha}) = \frac{n!}{\prod_{k=1}^{D} x_k!}\cdot\frac{\Gamma\left(\sum_{k=1}^{D}\alpha_k\right)}{\Gamma\left(n+\sum_{k=1}^{D}\alpha_k\right)}\cdot\prod_{k=1}^{D}\frac{\Gamma(x_k+\alpha_k)}{\Gamma(\alpha_k)}$

arrive at the minimum mean squared error (MMSE) estimator also termed posterior mean estimator [23] and is given by

$$\hat{\boldsymbol{\lambda}} = E[\boldsymbol{\lambda}|\mathbf{x}] = \int_{\mathcal{R}^N} \boldsymbol{\lambda} p(\boldsymbol{\lambda}|\mathbf{x}) d\boldsymbol{\lambda} \tag{1}$$

which can be written, by applying Bayes' rule, as[2]

$$\hat{\boldsymbol{\lambda}} = \frac{\int \boldsymbol{\lambda} p(\mathbf{x}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}}{\int p(\mathbf{x}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) d\boldsymbol{\lambda}}. \tag{2}$$

The marginal distribution of the noisy image $p(\mathbf{x})$ appearing in the denominator of (2) belongs to the mixed Poisson distribution family [24]; parameter estimation for an interesting subclass of this family is discussed in [25]. The Bayesian framework is attractive for our problem, but poses two challenges: One is the specification of an appropriate multivariate prior distribution for the intensity image, the other is the solution of the estimation problem in a multidimensional space which can be a formidable task.

A multiscale analysis of Poisson processes has been recently proposed independently in [13] and [14] which yields a tractable solution for Bayesian estimation. In the rest of this section we briefly review the multiscale framework suited for 1-D signals and its separable extension for the 2-D case, first proposed in [13]. Then we explore a multiscale quad-tree analysis explicitly designed for 2-D data and finally we propose our novel Poisson-Haar scheme which features extra benefits with respect to the other 2-D decomposition models.

B. Recursive Dyadic Partitioning

By denoting with $\mathbf{x}_0 = \mathbf{x}$ and $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}$ the finest scale representations of $\mathbf{x}$ and $\boldsymbol{\lambda}$, respectively, a multiscale analysis is obtained through recursions resembling those that yield the unnormalized Haar scaling coefficients [13]

$$x_j(k) = x_{j-1}(2k) + x_{j-1}(2k+1) \tag{3}$$

[2]For notational simplicity, we indicate the probability law associated either with a continuous or a discrete random variable X as $p(x)$, to be interpreted as either probability density or mass function depending on the case.

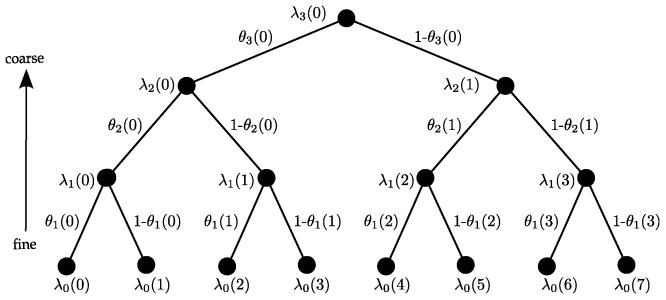

Fig. 1. Binary data tree obtained by a multiscale (fine-to-coarse) analysis of $\boldsymbol{\lambda}$ over three scales. Knowing the values of the splitting factors $\theta_j(k)$ and the total intensity of the image, $\lambda_J(0)$, we can fully recover our initial representation of $\boldsymbol{\lambda}$.

$$\lambda_j(k) = \lambda_{j-1}(2k) + \lambda_{j-1}(2k+1) \tag{4}$$

for $j = 1,\ldots,J$, $J = \log_2(N)$, and $k = 0,\ldots,N/2^j - 1$, as illustrated in Fig. 1. In the above equations j denotes the scale of analysis (J is the coarsest scale) and k the position in the corresponding vector $\mathbf{x}_j$. For simplicity we assume that the length N of the signal is a power of two; otherwise the analysis still holds by zero-padding the signal to make its length a power of two. This decomposition is motivated by two fundamental properties of Poisson processes [26]: (1) the counts over nonoverlapping intervals are independent, given the underlying intensities, and (2) the sum of independent Poisson random variables remains Poisson. Thus, the random variable $X_j(k)$, obtained as the sum of the Poisson random variables $X_{j-1}(2k)$ and $X_{j-1}(2k+1)$, will remain Poisson distributed with intensity $\lambda_j(k)$. Moreover, it holds that for two independent Poisson random variables, $X|\lambda_x \sim \mathrm{Pois}(\lambda_x)$ and $Y|\lambda_y \sim \mathrm{Pois}(\lambda_y)$, the conditional distribution of X given X + Y is binomial, namely $p(x|x+y,\lambda_x,\lambda_y) = \mathrm{Bin}(x|x+y,\lambda_x/(\lambda_x+\lambda_y))$ [27]. This statistical relation over adjacent scales permits the following factorization of the Poisson process likelihood function [14]:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\lambda}) &= \prod_{k=0}^{N-1} p\left(x_0(k)|\lambda_0(k)\right) \\ &= p\left(x_J(0)|\lambda_J(0)\right) \\ &\quad \times \prod_{j=1}^{J}\prod_{k=0}^{\frac{N}{2^j}-1} \mathrm{Bin}\left(x_{j-1}(2k)|x_j(k),\theta_j(k)\right) \end{aligned} \tag{5}$$

where $x_J(0)$ is the total count of the observed image $\mathbf{x}$, $\lambda_J(0)$ is the total intensity sum of the clean image $\boldsymbol{\lambda}$, and the rate-ratio parameter $\theta_j(k) \triangleq \lambda_{j-1}(2k)/\lambda_j(k)$ (success rate of the binomial distribution) can be interpreted as a splitting factor [20] that governs the multiscale refinement of the intensity $\boldsymbol{\lambda}$.

To proceed within the multiscale Bayesian framework we also have to express the prior distribution $p(\boldsymbol{\lambda})$ in a factorized form. With reference to Fig. 1, the intensity vector $\boldsymbol{\lambda}$ can be equivalently re-parameterized as $\boldsymbol{\lambda}_e = [\lambda_J(0),\theta_J(0),\ldots,\theta_1(0)\ldots,\theta_1(N/2-1)]$. This re-parametrization is closely related to the 1-D Haar wavelet transform, with λ_j corresponding to wavelet scaling coefficients and θ_j to wavelet detail coefficients [13]. More specifically, $\lambda_j(k)$ is a re-scaled version of the Haar scaling coefficient

$s_j(k)$, $\lambda_j(k) = 2^{j/2}s_j(k)$, while $\theta_j(k)$ is a divisibly normalized version of the Haar wavelet detail coefficient $w_j(k)$, also shifted by 0.5, $\theta_j(k) = (w_j(k)/2s_j(k)) + 0.5$. We show in Section II-D that in the 2-D case an analogous link exists between our proposed Poisson-Haar image decomposition and the 2-D Haar wavelet transform. At first sight the expression for θ in the 1-D case bears some resemblance to the Haar-Fisz transform [7], where the Haar wavelet coefficients of the noisy data are scaled by the square root of their corresponding scaling coefficient. However, in the Haar-Fisz case the normalization is applied to the wavelet coefficients of the noisy data to make them approximately Gaussian, while in this case the relation refers to the original intensities transformation. If we consider $\lambda_J(0)$ and $\theta_j(k)$ as observation samples of the random variables $\Lambda_J(0)$ and $\Theta_j(k)$, respectively, and assume statistical independence among these variables, then the prior distribution for $\boldsymbol{\lambda}_e$ can be expressed in the factorized form

$$p(\boldsymbol{\lambda}_e) = p\left(\lambda_J(0)\right) p(\boldsymbol{\theta}) = p\left(\lambda_J(0)\right) \prod_{j=1}^{J} \prod_{k=0}^{\frac{N}{2^j}-1} p\left(\theta_j(k)\right). \quad (6)$$

Since the beta distribution is conjugate to the binomial [28], it is convenient to complement the binomial terms in the likelihood function (5) with a beta-mixture prior

$$p\left(\theta_j(k)\right) = \sum_{m=1}^{M} \pi_{j,m} \mathrm{Beta}\left(\theta_j(k) | \alpha_{j,m}, \beta_{j,m}\right) \quad (7)$$

where $\pi_{j,m}$ is the mixture weight for the mth mixture component in jth scale of analysis, $\alpha_{j,m}$ and $\beta_{j,m}$ are the parameters of this beta mixture component and M is the total number of mixtures utilized at each scale. By using in (7) a mixture of beta densities instead of a single component, we can more accurately fit the image statistics. The increased accuracy of this approach is illustrated in Fig. 2, where the intensity rate-ratio histogram of a clean image is fitted by a single- versus a three-component distribution. Another way to justify modeling the random variables $\Theta_j(k)$ with beta mixture densities is to assume that, at every scale of analysis, each random variable $\Lambda_j(k)$ obeys a gamma distribution where the choice of gamma is due to its conjugacy with the Poisson distribution [26]. Specifically, if $\Lambda_{j-1}(2k) \sim \mathcal{G}(\alpha_j, \delta)$ and $\Lambda_{j-1}(2k+1) \sim \mathcal{G}(\beta_j, \delta)$, then the random variable $\Theta_j(k) = (\Lambda_{j-1}(2k))/(\Lambda_j(k))$ will be distributed as $\mathrm{Beta}(\alpha_j, \beta_j)$ [27]. Beyond the computational tractability arguments, the adoption of such a multiscale image prior induces an image intensity correlation structure corresponding to stochastic processes with $1/f$-like spectral behavior [13]. These processes correspond to fractal-based image descriptions and capture certain key properties of natural images, such as long-range intensity dependencies. In Fig. 3, we illustrate an example image sampled from this class of prior models.

The described statistical framework involving the multiscale factorization of the likelihood function and the prior density is appropriate for 1-D signals. For the case of our interest, $\mathbf{x}$ and $\boldsymbol{\lambda}$ will be 2-D images comprising the discrete observation samples and true values of the intensity function, respectively. One way

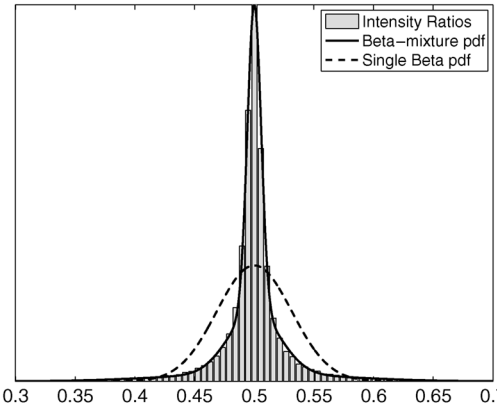

Fig. 2. Fitting the clean Lena image histogram of the intensity ratios $\theta_1^{(2)}(k)$, see (8), by either a single (dotted line) or a three (solid line) component symmetric ($\alpha_{jm} = \beta_{jm}$) beta-mixture distribution. The model parameters are fitted by maximum likelihood, using EM in the multiple mixture case.

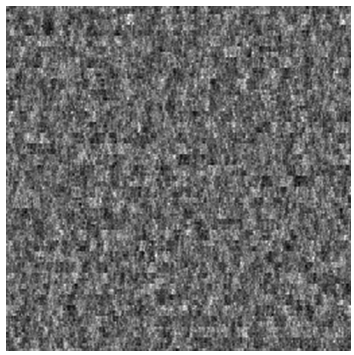

Fig. 3. Example image sampled from the proposed Poisson-Haar multiscale image prior model of Section II-D, in which dependencies across scales are captured by the HMT model of Section V.

to extend the multiscale analysis for 2-D images is to apply the 1-D model separably as proposed in [13]; each decomposition level will consist of one decimation step across the horizontal (or vertical) direction and then one decimation step across the vertical (resp. horizontal) direction, as illustrated in Fig. 4. Hereafter, we will refer to this 2-D multiscale framework as the *separable* model. Note that the separable model for a 2×2 neighborhood at a scale $j-1$, produces the following three rate-ratio subbands

$$\begin{aligned}
\theta_j^{(1)}(k,\ell) &= \frac{\lambda_{j-1}(2k,2\ell) + \lambda_{j-1}(2k,2\ell+1)}{\lambda_j(k,\ell)} \\
\theta_j^{(2)}(k,\ell) &= \frac{\lambda_{j-1}(2k,2\ell)}{\lambda_{j-1}(2k,2\ell) + \lambda_{j-1}(2k,2\ell+1)} \\
\theta_j^{(3)}(k,\ell) &= \frac{\lambda_{j-1}(2k+1,2\ell)}{\lambda_{j-1}(2k+1,2\ell) + \lambda_{j-1}(2k+1,2\ell+1)}
\end{aligned} \quad (8)$$

which, in contrast to the 1-D case, do not correspond one-to-one to the three subbands of the 2-D Haar wavelet transform.

C. Recursive Quad-Tree Partitioning

Besides the separable image model, we study a nonseparable 2-D quad-tree multiscale decomposition for the observation and intensity images $\mathbf{x}$ and $\boldsymbol{\lambda}$ of size $N_1 \times N_2$. Denoting for each

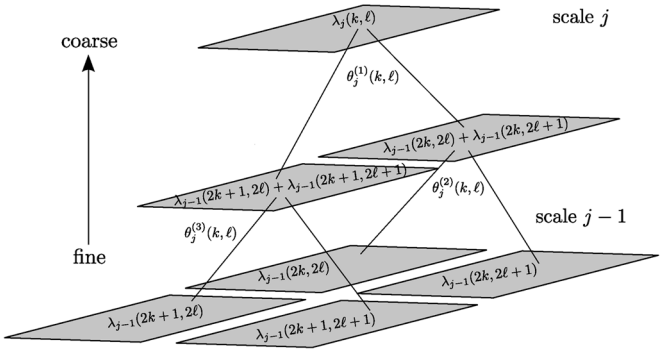

Fig. 4. Multiscale 2-D image analysis using the separable decomposition scheme of [13] and [20].

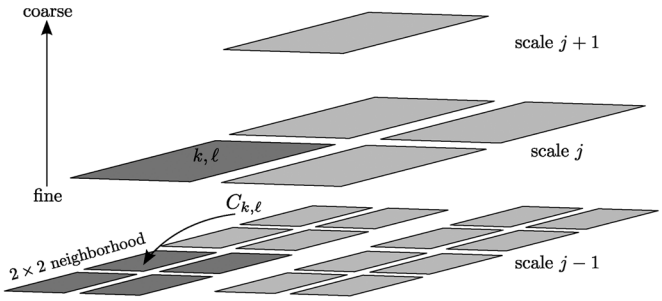

Fig. 5. Multiscale 2-D image analysis using the quad decomposition scheme. Each data node $\lambda_j(k,\ell)$ or $x_j(k,\ell)$ at a coarser scale, is the sum of the corresponding data nodes in the 2×2 neighborhood $C_{k,\ell}$ at the next finer scale.

jth-scale pixel location (k,ℓ) the 2×2 set of children pixel locations at the next finer scale as

$$C_{k,l}=\{(2k,2\ell),(2k,2\ell+1),(2k+1,2\ell),(2k+1,2\ell+1)\}$$

the decomposition formulas similar to (3) and (4) will be

$$x_j(k,\ell)=\sum_{(k',\ell')\in C_{k,\ell}} x_{j-1}(k',\ell') \tag{9}$$

$$\lambda_j(k,\ell)=\sum_{(k',\ell')\in C_{k,\ell}} \lambda_{j-1}(k',\ell') \tag{10}$$

for $j=1,\dots J$, $k=0,\dots,N_1/2^j-1$, $\ell=0,\dots,N_2/2^j-1$ and $J=\min\{\log_2(N_1),\log_2(N_2)\}$. From this multiscale analysis we obtain the quad-tree decomposition of the observation and intensity images $\mathbf{x}$ and $\boldsymbol{\lambda}$, as illustrated in Fig. 5. Similarly to the separable decomposition, since the random variables $X_{j-1}(k',l')$, with $(k',\ell')\in C_{k,\ell}$, are assumed conditionally independent, the random variable $X_j(k,\ell)$ will remain Poisson distributed with intensity $\lambda_j(k,\ell)$. Moreover it will similarly hold that, given $X_j(k,\ell)$, the joint conditional distribution of the four children's random vector $\mathbf{X}_j^c(k,\ell)=\{X_{j-1}(k',\ell'):(k',\ell')\in C_{k,\ell}\}$ is multinomial [26]

$$p\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right) = \mathrm{Mult}\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right) \tag{11}$$

with

$$\boldsymbol{\theta}_j(k,\ell)=\left\{\frac{\lambda_{j-1}(k',\ell')}{\lambda_j(k,\ell)}:(k',\ell')\in C_{k,\ell}\right\}. \tag{12}$$

The likelihood function can then be factorized as (see Appendix A for the derivation)

$$p(\mathbf{x}|\boldsymbol{\lambda})=p\left(x_J(0,0)|\lambda_J(0,0)\right) \cdot\prod_{j=1}^{J}\prod_{k=0}^{\frac{N_1}{2^j}-1}\prod_{\ell=0}^{\frac{N_2}{2^j}-1}\mathrm{Mult}\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right). \tag{13}$$

Similarly to the separable model, we also use a factorized multiscale prior on the re-parameterized intensity $\boldsymbol{\lambda}_e$

$$p(\boldsymbol{\lambda}_e)=p\left(\lambda_J(0,0)\right)\prod_{j=1}^{J}\prod_{k=0}^{\frac{N_1}{2^j}-1}\prod_{\ell=0}^{\frac{N_2}{2^j}-1}p\left(\boldsymbol{\theta}_j(k,\ell)\right). \tag{14}$$

Once more, since the Dirichlet distribution is conjugate to the multinomial [28], we adopt mixture of Dirichlet prior factors

$$p\left(\boldsymbol{\theta}_j(k,\ell)\right)=\sum_{m=1}^{M}\pi_{j,m}\mathrm{Dir}\left(\boldsymbol{\theta}_j(k,\ell)|\boldsymbol{\alpha}_{j,m}\right) \tag{15}$$

where $\boldsymbol{\alpha}_{j,m}$ is the parameter vector of the mth mixture component in the jth scale of analysis. An alternative justification is provided by noting that the Dirichlet like the beta distribution admits a representation in terms of gamma variables. We will refer hereafter to this decomposition as the *quad* model. A similar decomposition scheme was first mentioned but not further pursued in [20]. Later, it was studied in [17] in the context of image deblurring; however, the authors in [17] only considered the simpler case of a single-component Dirichlet prior distribution.

D. Multiscale Poisson-Haar Image Partitioning

As we have already seen, the separable and quad models can provide convenient multiscale representations for images degraded by Poisson noise. The re-parameterization of the intensity image into $\boldsymbol{\lambda}_e$ which comprises a single point $\lambda_J(0,0)$, encompassing the total intensity of the image, and the rate-ratio array $\boldsymbol{\theta}$ has enabled us to choose effective image priors. The array elements of $\boldsymbol{\theta}$ are defined in both cases as ratios of intensities at adjacent scales; see (8) for the separable and (12) for the quad model. These rate-ratios can be interpreted as intensity splitting factors and correspond to the percentage of the parent intensity at a coarse scale node which is distributed to each child at the next finer scale and, as we have discussed in Section II-B, they are closely related to wavelet detail coefficients.

This suggests interpreting the $\Theta_j(k,\ell)$ variables as representing the edge detail structure of the image at scale j. For the separable model of Fig. 4, with rate-ratio variables defined in (8), the value of the $\theta_j^{(1)}(k,\ell)$ variable and the value of the $\{\theta_j^{(2)}(k,\ell),\theta_j^{(3)}(k,\ell)\}$ variables characterizes the image edge content in the horizontal and vertical orientations, respectively, for that particular scale and position. In particular, existence of

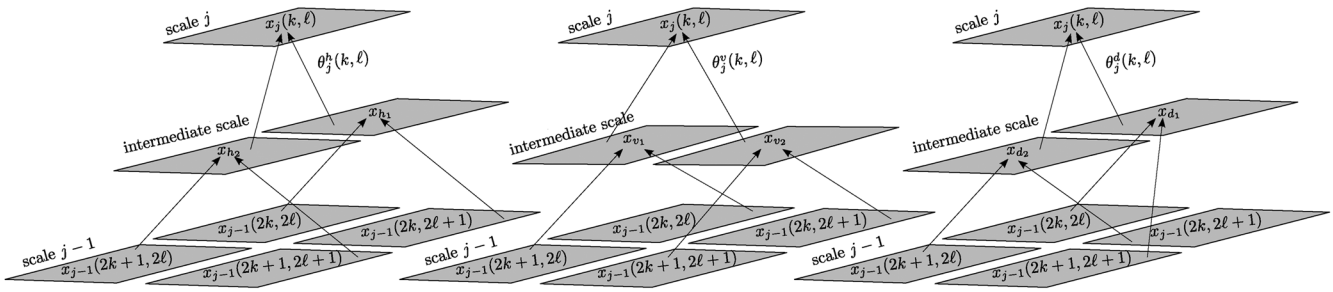

Fig. 6. Poisson-Haar image multiscale decomposition. All three trees result in the same parent value $x_j(k,\ell)$ at scale j. The difference among them lies in the formation of the intermediate scale, designed so that the 3 splitting factors $\theta^h, \theta^v, \theta^d$ of (16) are sensitive to image edges at different orientations (horizontal, vertical, or diagonal).

a step edge in each direction is indicated by significant departure from 0.5 of the corresponding rate-ratio coefficient. Under this view, a shortcoming of the separable 2-D model is that it presents poor orientation selectivity in capturing diagonal image edges. Unfortunately, the quad model also performs poorly in this task. More specifically, if all the quad intensity ratios $\boldsymbol{\theta}_j(k,\ell)$ defined in (12) are equal to 0.25, then the image is smooth in this region, since the parent splits its intensity equally to all children at the finer scale. Otherwise, it is not trivial to succinctly characterize the existence and orientation of image edges.

To achieve improved orientation selectivity in comparison to the previous two approaches, we propose a novel multiscale decomposition for 2-D Poisson processes designed to also explicitly model diagonal image edges, which we term *Poisson-Haar* decomposition due to its close link to the 2-D Haar wavelet transform. In the proposed decomposition, similarly to the separable scheme, we assume that among two scales, $j-1$ and j, there is an intermediate level. The parent observation $x_j(k,\ell)$ is then considered as the result of two consecutive steps. First, from the four children of scale $j-1$ two intermediate terms are produced, which in the second step are summed to give the parent observation, as shown in Fig. 6. This approach resembles the way the separable model decomposes the image; however, there is a critical difference: While in the separable scheme a single pair of intermediate terms is considered, see Fig. 4, in the proposed scheme we examine all three possible pair combinations at scale $j-1$ to form the intermediate terms. Each combination corresponds to a different image decomposition tree and can capture image edges at different orientations (horizontal, vertical, or diagonal). The advantage of our scheme is that it symmetrically considers all 3 orientations, while the separable scheme can only account for horizontal or vertical orientations.

More specifically, the proposed multiscale scheme is defined recursively as follows. Assume that we observe the counts $\mathbf{x}_j^c(k,\ell)$. Then the parent observation $x_j(k,\ell)$ can be equally obtained as the sum of any one of the three distinct pairs of intermediate terms, $\{x_{h_1}, x_{h_2}\}$, $\{x_{v_1}, x_{v_2}\}$ and $\{x_{d_1}, x_{d_2}\}$, as illustrated in Fig. 6. The three pairs of terms are produced by convolving $\mathbf{x}_j^c(k,\ell)$ with the kernel pairs $\left\{\begin{pmatrix}0&0\\1&1\end{pmatrix}, \begin{pmatrix}1&1\\0&0\end{pmatrix}\right\}$, $\left\{\begin{pmatrix}0&1\\0&1\end{pmatrix}, \begin{pmatrix}1&0\\1&0\end{pmatrix}\right\}$ and $\left\{\begin{pmatrix}1&0\\0&1\end{pmatrix}, \begin{pmatrix}0&1\\1&0\end{pmatrix}\right\}$, respectively, and then decimating the result by 2 in each direction. Since every intermediate term is derived as the sum of independent Poisson random variables, it will remain Poisson distributed. Thus, the conditional distributions given the parent value $p(x_{o_1}(k,\ell)|x_j(k,\ell))$, where o is one of (h,v,d) and denotes the orientation of each sum in Fig. 6, will be binomial with corresponding intensity-ratios

$$\begin{aligned}\theta_j^h(k,\ell) &= \frac{\lambda_{j-1}(2k,2\ell)+\lambda_{j-1}(2k,2\ell+1)}{\lambda_j(k,\ell)}\\ \theta_j^v(k,\ell) &= \frac{\lambda_{j-1}(2k,2\ell)+\lambda_{j-1}(2k+1,2\ell)}{\lambda_j(k,\ell)}\\ \theta_j^d(k,\ell) &= \frac{\lambda_{j-1}(2k,2\ell)+\lambda_{j-1}(2k+1,2\ell+1)}{\lambda_j(k,\ell)}.\end{aligned} \tag{16}$$

Comparing these rate-ratios to the ones of the 2-D separable model of (8), we notice that our Poisson-Haar scheme explicitly represents not only vertical and horizontal, but also diagonal edges. Moreover in this case, as opposed to the separable and quad schemes there is a direct relation between the θ rate-ratios with the 2-D Haar wavelet detail coefficient w in the corresponding subband, similarly to the 1-D case discussed in Section II-B

$$\theta_j^o(k,\ell) = \frac{w_j^o(k,\ell)}{2s_j(k,\ell)} + \frac{1}{2}.$$

In addition, $\lambda_j(k,\ell)$ is a re-scaled version of the 2-D Haar scaling coefficient $\lambda_j(k,\ell) = 2^j s_j(k,\ell)$. Equation (16) implies that we can employ an equivalent vector representation for the intensity image $\boldsymbol{\lambda}$, as follows:

$$\begin{aligned}\boldsymbol{\lambda}_e = [&\lambda_J(0,0), \boldsymbol{\theta}_J(0,0), \ldots, \boldsymbol{\theta}_1(0,0), \ldots,\\ &\boldsymbol{\theta}_1(N_1/2-1, N_2/2-1)]\end{aligned} \tag{17}$$

where the vectors $\boldsymbol{\theta}$ are of the form: $\boldsymbol{\theta}_j(k,\ell) = (\theta_j^h(k,\ell), \theta_j^v(k,\ell), \theta_j^d(k,\ell))$. Then each element of $\boldsymbol{\theta}_j(k,\ell)$ will be considered as an independent realization of one of the random variables $\Theta_j^h(k,\ell)$, $\Theta_j^v(k,\ell)$ and $\Theta_j^d(k,\ell)$. These variables will be modeled as beta-mixture distributions (7) following the same reasoning as for the separable scheme. The image prior distribution, similarly to the quad model, admits the multiscale factorization of (14), with

$$p\left(\boldsymbol{\theta}_j(k,\ell)\right) = p\left(\theta_j^h(k,\ell)\right) p\left(\theta_j^v(k,\ell)\right) p\left(\theta_j^d(k,\ell)\right). \tag{18}$$

The choice of an independent prior $p(\boldsymbol{\theta})$, across different orientations can be considered as analogous to the independent treatment of the wavelet detail subbands, which is a usual assumption in the wavelets literature.

The reconstruction of $\boldsymbol{\lambda}$ from $\boldsymbol{\lambda}_e$ (17) is one-to-one and can be achieved through a coarse-to-fine scale-recursive procedure, for $j = J, J-1, \ldots, 1$

$$\begin{aligned}
\lambda_{j-1}(2k,2\ell) &= \kappa\left(\theta_j^h(k,\ell)+\theta_j^v(k,\ell)+\theta_j^d(k,\ell)-1\right)\\
\lambda_{j-1}(2k,2\ell+1) &= \kappa\left(\theta_j^h(k,\ell)-\theta_j^v(k,\ell)-\theta_j^d(k,\ell)+1\right)\\
\lambda_{j-1}(2k+1,2\ell) &= \kappa\left(\theta_j^v(k,\ell)-\theta_j^h(k,\ell)-\theta_j^d(k,\ell)+1\right)\\
\lambda_{j-1}(2k+1,2\ell+1) &= \kappa\left(\theta_j^d(k,\ell)-\theta_j^h(k,\ell)-\theta_j^v(k,\ell)+1\right)
\end{aligned} \tag{19}$$

with $\kappa = \lambda_j(k,\ell)/2$.

III. Bayesian Intensity Estimation

In this section we derive an optimal Bayes estimator for the re-parameterized intensity vector $\boldsymbol{\lambda}_e$. This estimator minimizes the Bayesian MSE criterion [23] in the transformed intensity domain. Having estimated $\boldsymbol{\lambda}_e$, it is then straightforward to recover the intensity image $\boldsymbol{\lambda}$ of interest by a scale-recursion, as is (19) for the proposed Poisson-Haar decomposition scheme (analogous recursions hold for the separable and quad schemes). Note that other estimators could also be derived under the presented framework by adopting alternative optimality criteria, e.g., the MAP criterion.

A. Multiscale Posterior Factorization

Having factorized the prior distributions (6), (14), (18) and likelihood functions (5), (13) over multiple scales, it is straightforward to show that the posterior distribution $p(\boldsymbol{\lambda}_e|\mathbf{x}) = p(\lambda_J(0,0)|x_J(0,0))p(\boldsymbol{\theta}|\mathbf{x})$ also admits a multiscale factorization. The factorization of the posterior implies that the intensity estimation problem can be solved individually at each scale and position, instead of requiring a complicated high dimensional solution. At this point we are going to exploit the conjugacy between the likelihood function and the prior distribution to derive the posterior distribution.

By combining the likelihoods (5), (13) and the priors (6), (14), we obtain the posteriors for the binary and the quad tree decomposition models in the following forms:

$$\begin{aligned}
p(\boldsymbol{\lambda}_e|\mathbf{x}) &= p\left(\lambda_J(0)|x_J(0)\right)\\
&\quad\cdot\prod_{j=1}^{J}\prod_{k=0}^{\frac{N}{2^j}-1} p\left(\theta_j(k)|x_j(k),x_{j-1}(2k)\right)
\end{aligned} \tag{20}$$

$$\begin{aligned}
p(\boldsymbol{\lambda}_e|\mathbf{x}) &= p\left(\lambda_J(0,0)|x_J(0,0)\right)\\
&\quad\cdot\prod_{j=1}^{J}\prod_{k=0}^{\frac{N_1}{2^j}-1}\prod_{\ell=0}^{\frac{N_2}{2^j}-1} p\left(\boldsymbol{\theta}_j(k,\ell)|\mathbf{x}_j^c(k,\ell)\right).
\end{aligned} \tag{21}$$

Using the identity (can be easily verified by direct substitution)

$$\mathrm{Mult}(\mathbf{x}|n,\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Polya}(\mathbf{x}|n,\boldsymbol{\alpha})\mathrm{Dir}(\boldsymbol{\theta}|\mathbf{x}+\boldsymbol{\alpha}) \tag{22}$$

where the Polya distribution [29] (also called Dirichlet-multinomial) is defined in Table I, we can write the posterior factors as mixtures of Dirichlet densities

$$\begin{aligned}
p\left(\boldsymbol{\theta}_j(k,\ell)|\mathbf{x}_j^c(k,\ell)\right) &= \sum_{m=1}^{M}\gamma_m\left(z_j(k,\ell)\right)\\
&\quad\times\mathrm{Dir}\left(\boldsymbol{\theta}_j(k,\ell)|\mathbf{x}_j^c(k,\ell)+\boldsymbol{\alpha}_{j,m}\right).
\end{aligned} \tag{23}$$

Here $z_j(k,\ell)$ indicates which mixture component is responsible for generating the observation $\mathbf{x}_j^c(k,\ell)$ and the corresponding posterior mixture assignment weights equal

$$\gamma_m\left(z_j(k,\ell)\right) = \frac{\pi_{j,m}\mathrm{Polya}\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\alpha}_{j,m}\right)}{\sum_{n=1}^{M}\pi_{j,n}\mathrm{Polya}\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\alpha}_{j,n}\right)}. \tag{24}$$

Note that the expressions (23) and (24) also cover the separable model, in which case the Dirichlet reduces to beta distribution.

In all considered multiscale models, the posterior distribution of the total intensity $\lambda_J(0,0)$ could be modeled as gamma distribution, $p(\lambda_J(0,0)|x_J(0,0)) = \mathcal{G}(x_J(0,0)+\gamma,(\delta/(\delta+1)))$, whose mean is $\hat{\lambda}_J(0,0) = (x_J(0,0)+\gamma)\delta/(\delta+1)$. This result is obtained by placing a gamma prior on the total intensity, i.e., $\Lambda_J(0,0) \sim \mathcal{G}(\gamma,\delta)$, due to the conjugacy of the Poisson with the gamma distribution. Such an approach is followed by the authors in [15]. However, this modeling is not crucial for the efficiency of the final estimator. The local SNR of the image, defined as the ratio of the mean pixel value to the standard deviation of the pixel value, will equal to $SNR = \sqrt{\lambda_j(k,\ell)}$ for a 2-D Poisson process at the jth scale. As we are reaching the coarsest scales of analysis, the intensity $\lambda_j(k,\ell)$ increases and so does the SNR. Thus, since the number of counts at the coarsest scales is large in practice, we have chosen to use the observed total intensity as a good and reliable estimation of the underlying total intensity $\lambda_J(0,0)$, that is $\hat{\lambda}_J(0,0) \approx x_J(0,0)$.

B. Posterior Mean Estimation

Having at hand the posterior distribution, we can readily obtain the posterior mean estimator of the re-parameterized intensity $\boldsymbol{\lambda}_e$. First, we will derive the posterior mean estimator for the quad model and then we will present the corresponding estimates for the separable and the Poisson-Haar multiscale partitioning schemes, as they are found to be special cases of the former. Due to the factorization of the posterior distribution (21) and the independence assumption of the rate-ratio coefficients $\boldsymbol{\theta}$ in the prior model of (14), we can find the optimal estimator for each element of $\boldsymbol{\theta}_j(k,\ell)$ separately as

$$\begin{aligned}
\hat{\theta}_j^i(k,\ell) &= E\left[\theta_j^i(k,\ell)|\mathbf{x}_j^c(k,\ell)\right]\\
&= \int_{\mathcal{R}^D}\theta_j^i(k,\ell)p\left(\boldsymbol{\theta}_j(k,\ell)|\mathbf{x}_j^c(k,\ell)\right)d\boldsymbol{\theta}_j(k,\ell)\\
&= \sum_{m=1}^{M}\gamma_m\left(z_j(k,\ell)\right)\left(\frac{x_j^{c,i}(k,\ell)+\alpha_{j,m}^i}{x_j(k,\ell)+\alpha_{j,m}^0}\right)
\end{aligned} \tag{25}$$

where $\theta_j^i(k,\ell)$ is the ith element of the vector $\boldsymbol{\theta}_j(k,\ell)$ and $\alpha_{j,m}^0 = \sum_{t=1}^{D} \alpha_{j,m}^t$.

For the 1-D decomposition the MMSE estimator will be a special case of (25). The separable estimator for the 2-D image case is of the same form with the 1-D case and can be obtained by successively processing the rows and the columns of the image. For more details see [13]. Finally, the optimal estimates for the variables $(\Theta_j^h(k,\ell), \Theta_j^v(k,\ell), \Theta_j^d(k,\ell))$ arising in the proposed Poisson-Haar multiscale model will equal to

$$\hat{\theta}_j^o(k,\ell) = \sum_{m=1}^{M} \gamma_m\left(z_j^o(k,\ell)\right)\left(\frac{x_{o_1}(k,\ell) + \alpha_{j,m}^o}{x_j(k,\ell) + \alpha_{j,m}^o + \beta_{j,m}^o}\right) \tag{26}$$

with

$$\gamma_m\left(z_j^o(k,\ell)\right) = \frac{\pi_{j,m}\mathrm{Polya}\left(x_{o_1}(k,\ell)|x_j(k,\ell), \alpha_{j,m}^o, \beta_{j,m}^o\right)}{\sum_{n=1}^{M} \pi_{j,n}\mathrm{Polya}\left(x_{o_1}(k,\ell)|x_j(k,\ell), \alpha_{j,n}^o, \beta_{j,n}^o\right)} \tag{27}$$

where $x_{o_1}(k,\ell)$ denotes the first term of each pair of intermediate terms discussed in Section II-D and o denotes the selected orientation.

C. Recursive-Estimation and Cycle Spinning

Choosing any one of the three partitioning schemes, allows us to end up with an optimal estimation for the vector $\boldsymbol{\lambda}_e$. Since $\boldsymbol{\lambda}_e$ is a transformation of the intensity image $\boldsymbol{\lambda}$, we obtain our final estimate for the image of interest, by employing a scale-recursive algorithm to recover the original image intensity parameterization. Note that even though the applied transform is nonlinear, the resulting estimator is still optimal, in the MMSE sense, and in the image domain (see Appendix B for the proof). The algorithm starts at the coarser scale of analysis where $\hat{\lambda}_J(0,0)$ and $\hat{\boldsymbol{\theta}}_J(0,0)$ are used to obtain the intensity estimators for the next finer scale. For the proposed Poisson-Haar decomposition scheme the equations which lead us to the next scale are those in (19) but in place of the ground-truth $\boldsymbol{\theta}$ we use their estimates. Analogous equations hold for the other two schemes. This procedure is repeated until we reach the finest scale of analysis, $j = 0$, where the image of interest lies.

A shortcoming of all discussed tree partitioning schemes is that the resulting tree structure induces nonstationarity in space. This is due to the fact that neighboring image pixels might only share ancestors at very coarse scales of the tree [13], [21], [22]. The induced nonstationarity typically results in block artifacts in the final estimates. Such effects are also frequently met in the critically sampled wavelet transform. To alleviate these artifacts a commonly used technique is cycle spinning [30], which can yield significant improvements in the quality of the intensity estimates. This technique entails circularly shifting the observed data, estimating the underlying intensity as described, and shifting back to the original position. This is repeated for all possible shifts. The final estimate is produced by averaging the results over all shifts.

IV. EM-Based Model Parameter Estimation

The posterior mean estimates obtained in Section III-B require knowledge of the model parameters $\boldsymbol{\mu} = \{\boldsymbol{\pi}_j, \boldsymbol{\alpha}_j\}$, i.e., the mixture weights and the beta/Dirichlet mixture parameters governing the splitting factors $\Theta_j(k,\ell)$. One could consider these parameters random, assign a hyper-prior $p(\boldsymbol{\mu})$ on them, and proceed with the resulting hierarchical Bayesian model [31]. However, such an approach can be computationally very challenging. Herein we opt for an empirical Bayes treatment of the problem, amounting to estimating the parameters $\boldsymbol{\mu}$ for each noisy image and considering them fixed during the estimation task.

Previous works following the empirical Bayes treatment of the problem have not addressed satisfactorily the parameter estimation problem. The authors in [13] model the random variables $\Theta_j(k,\ell)$ with mixtures of 3 symmetric betas where the beta parameters are heuristically set for every scale of analysis. They also assume that one of the mixture weights is known and they estimate the remaining two by utilizing the method of moments. Actually, since $\pi_{j,1}$ is assumed to be known and $\pi_{j,3}$ is constrained to be $\pi_{j,3} = 1 - \pi_{j,1} - \pi_{j,2}$, only $\pi_{j,2}$ needs to be estimated. This method is quite restrictive, being appropriate only for a setup with at most three mixture components. It also has the drawback that it often produces inconsistent results, since one of the mixture weights may take a negative value as has also been noticed in [16]. A different approach was followed in [16] where the authors utilize an auxiliary wavelet thresholding denoising method to obtain an estimate of the splitting factors, on which $\boldsymbol{\mu}$ is finally fitted with EM. A drawback of this approach is that since it relies on a denoising method, which does not take into consideration the Poisson statistics of the image, it is prone to a potential failure of the auxiliary method. In addition, the performance of the estimation task using the parameters obtained by this approach is limited, since the resulting estimates of the splitting factors, by the auxiliary method, differ from the best possible ones.

To account for these limitations, we pursue an ML estimator to infer model parameters $\boldsymbol{\mu}$ *directly* from the observed Poisson data. Since the multiscale approach allows handling each scale independently, we drop index j for clarity. Also, instead of (k,ℓ), we use the index n to denote position, assuming that we have raster-scanned the observed image at each scale into a vector $\mathbf{x}$ of length N. The key idea is to integrate-out the unobserved splitting factors $\boldsymbol{\theta}_n$ and, thus, relate the model parameters $\boldsymbol{\mu}$ directly to the observations $\mathbf{x}$. Using the identity (22), this integration can be computed in closed form, yielding the factorized likelihood function

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{\mu}) d\boldsymbol{\theta} = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\mu}) d\boldsymbol{\theta} \\ &= \prod_{n=1}^{N} \int p\left(\mathbf{x}_n^c|x_n, \boldsymbol{\theta}_n\right) p(\boldsymbol{\theta}_n|\boldsymbol{\mu}) d\boldsymbol{\theta}_n \\ &= \prod_{n=1}^{N} \left(\sum_{m=1}^{M} \pi_m \int p\left(\mathbf{x}_n^c|x_n, \boldsymbol{\theta}_n\right) p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}_m) d\boldsymbol{\theta}_n\right) \\ &= \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m \mathrm{Polya}\left(\mathbf{x}_n^c|x_n, \boldsymbol{\alpha}_m\right) \end{aligned} \tag{28}$$

where $\mathbf{x}_n^c$ are the Poisson counts of x_n's children. For the separable and the proposed multiscale Poisson-Haar image partitioning, $p(\mathbf{x}_n^c|x_n, \boldsymbol{\theta}_n)$ and $p(\boldsymbol{\theta}_n|\boldsymbol{\alpha}_m)$ are the binomial and mix-

ture of beta distributions with $\boldsymbol{\alpha}_m = [\alpha_m, \beta_m]$, while for the quad model they are the multinomial and Dirichlet mixture distributions, respectively.

The ML parameters maximize the log-likelihood $L(\boldsymbol{\mu}) = \log p(\mathbf{x}|\boldsymbol{\mu})$. Optimizing $L(\boldsymbol{\mu})$ directly is hard, due to the multiple mixture components. We, thus, invoke the EM algorithm [28] and work with the complete log-likelihood. According to the EM algorithm, a latent M-dimensional binary random variable $\mathbf{z}_n$ is assigned to each observation sample $\mathbf{x}_n^c$. In any instance of $\mathbf{z}_n$ only a particular element z_{nm} can equal one, while all the others must be zero. The element z_{nm} is equal to one only if the mth mixture component of the Polya distribution is responsible for generating the observations $\mathbf{x}_n^c$, given the parent observation x_n. With $\mathbf{z}$ we denote the set of the discrete mixture component assignments for all the observation samples, i.e., $\mathbf{z} = \{\mathbf{z}_1 \dots \mathbf{z}_n\}$. The marginal distribution over each $\mathbf{z}_n$ is specified in terms of the mixture weights, such that $p(z_{nm} = 1) = \pi_m$. Based on this notion $L(\boldsymbol{\mu})$ is considered as the incomplete log-likelihood and $L_c(\boldsymbol{\mu}) = \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\mu})$ as the complete one. Now if we suppose that, apart from the observation samples, we are also given the values of the corresponding latent variables in $\mathbf{z}$, we can express the complete log-likelihood in the form

$$\begin{aligned} L_c(\boldsymbol{\mu}) &= \log \prod_{n=1}^{N} \prod_{m=1}^{M} \left[\pi_m \mathrm{Polya}\left(\mathbf{x}_n^c | x_n, \boldsymbol{\alpha}_m\right)\right]^{z_{nm}} \\ &= \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log(\pi_m) \\ &\quad + \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log \mathrm{Polya}\left(\mathbf{x}_n^c | x_n, \boldsymbol{\alpha}_m\right). \end{aligned} \quad (29)$$

This form is much easier to work with since the logarithm does not act any more on the sums of the mixture components but directly on the Polya distribution. However, the problem now is that the true values for the latent variables are unknown. For this reason, instead of considering the complete log-likelihood, we compute its expected value under the posterior distribution of the latent variables. This computation takes place in the expectation step (E-step) of the algorithm, where the conditional expectation of the complete likelihood, given the observed data and the current estimates of the unknown parameters, is of the form

$$\begin{aligned} E\left[L_c(\boldsymbol{\mu})|\mathbf{x}, \boldsymbol{\mu}^{(l)}\right] &= \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(z_n) \log(\pi_m) \\ &\quad + \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_m(z_n) \log \mathrm{Polya}\left(\mathbf{x}_n^c | x_n, \boldsymbol{\alpha}_m\right) \end{aligned} \quad (30)$$

where $\gamma_m(z_n) = E[z_{nm} = 1|\mathbf{x}]$ is defined as in (24) and (27) according to the selected partitioning scheme. Note that the expected value of the complete log-likelihood (30) consists of two separate terms. Thus, to estimate the updated model parameters $\boldsymbol{\mu}^{(l+1)} = \{\boldsymbol{\pi}^{(l+1)}, \boldsymbol{\alpha}^{(l+1)}\}$ we can maximize the first term with respect to $\boldsymbol{\pi}$, and the second with respect to $\boldsymbol{\alpha}$. This is the maximization step of the algorithm (M-step). Maximizing (30) with respect to $\boldsymbol{\pi}$, under the constraint $\sum_{m=1}^{M} \pi_m = 1$, yields the updated mixture weights

$$\pi_m^{(l+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_m(z_n). \quad (31)$$

In order to obtain the updated parameters $\boldsymbol{\alpha}^{(l+1)}$ we first have to compute the partial derivative of the log-likelihood (30) with respect to α_m^t

$$\begin{aligned} \frac{\partial E[L_c]}{\partial \alpha_m^t} = \sum_{n=1}^{N} \gamma_m(z_n) \Big[&\psi\left(x_n^{c,t} + \alpha_m^t\right) - \psi\left(\alpha_m^t\right) \\ &+ \psi\left(\sum_{t=1}^{D} \alpha_m^t\right) - \psi\left(x_n + \sum_{t=1}^{D} \alpha_m^t\right) \Big] \end{aligned} \quad (32)$$

where $\psi(\cdot)$ is the digamma function [32] defined as $\psi(x) = d \log \Gamma(x)/dx$ and the superscript $t = 1, \dots, D$, with $D = 4$ for the quad model and $D = 2$ for the other two partitioning models, denotes the tth element of the corresponding vector. The updated parameters can be found by setting (32) equal to zero and solving for α_m^t. Unfortunately this equation is nonlinear so we cannot obtain a closed form solution for the parameters. Further, the Polya distribution does not belong to the exponential family of distributions; thus, it does not have sufficient statistics [33]. Therefore, in order to find a solution we have to resort to an iterative method. At this point, note that a similar equation to (32) also arises in other research areas such as text and language modeling [34], [35] and DNA sequence modeling [36]. In these cases the authors in order to find a solution either employ a gradient descent method or the fixed-point iteration method proposed in [37].

For finding the root of (32) we propose a novel technique which employs the Newton-Raphson method [38]. A problem that conventional techniques encounter is that for images with flat cartoon-like content, the histogram of splitting factors $\boldsymbol{\theta}$ is strongly peaked (at 0.25 for the quad model and 0.5 for the separable and our Poisson-Haar partitioning scheme), resulting in very large α_m^t parameters. The ML criterion then can lead to over-fitting, with the largest of the α_m^t parameters unboundedly increasing at every EM iteration (the corresponding beta mixture is increasingly peaked at 0.5). Numerically finding the very large root of the log-likelihood derivative (32) also becomes unstable and time consuming. To address this issue we have added a regularization term $-\epsilon \sum_{m=1}^{M} \sum_{t=1}^{D} \alpha_m^t$ in (30), where ϵ is a small positive constant. With this approach, even for large values of α_m^t, we succeed in finding a root in the regularized version of (32) and consequently a maximum for the penalized version of (30). Furthermore, for small and medium values of α_m^t, the root of the regularized version of (32) would only negligibly differ from the original one, as illustrated in Fig. 7. This regularization term can also be interpreted as specifying a conjugate prior for $\boldsymbol{\alpha}_m$ and leads to a MAP instead of the standard ML estimation in the M-step of the algorithm [28] (for details see Appendix C). The resulting penalized EM algorithm is extremely robust in practice.

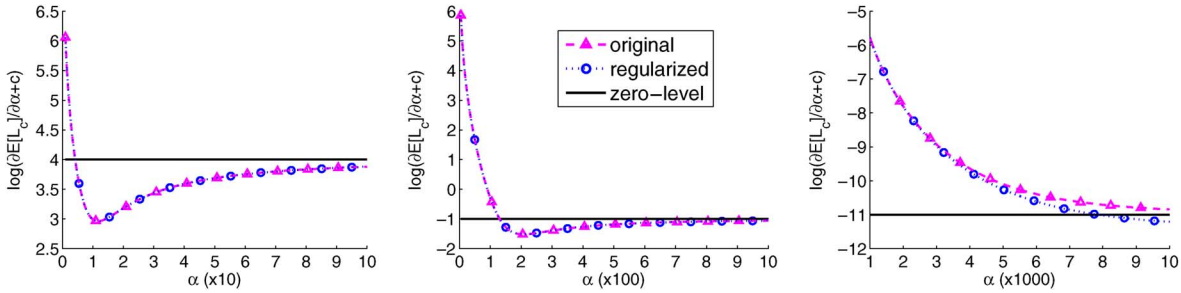

Fig. 7. Log-likelihood partial derivative with respect to the symmetric beta parameter α_m and its regularized version. The three plots correspond to the components with small (left), medium (center), and large (right) α_m parameters. Regularization effectively bounds the root of the largest parameter component, while essentially leaving the other two roots intact.

The EM algorithm needs to be properly initialized. To render our parameter estimation algorithm fully automatic, we have developed a bootstrapping technique. Specifically, we start with a single-component density whose parameters are fitted by ML. Then the number of mixtures is incremented up to the desired number. This is achieved by repeatedly splitting the mixture with the largest total assignment weight $\gamma_m = \sum_{n=1}^{N} \gamma_m(z_n)$. Each split yields two mixture components whose initial parameters are perturbed versions of their generating component. In case that reasonable initial parameters from training on similar images are available, one can use them instead as initial condition to accelerate training.

V. Modeling Interscale Dependencies With Hidden Markov Trees

In Section II we modeled the splitting factors, Θ, as independent mixtures of beta/Dirichlet random variables across scales, yielding the factorized prior $p(\boldsymbol{\lambda}_e)$ in the form of (6), (14), and (18). Often this independence assumption across scales may be too simplistic, limiting the performance of the models adopting it. To address this issue, we can integrate into the multiscale Poisson decomposition framework the HMT model, first introduced in the context of signal denoising in [19], which better models interscale dependencies between mixture assignments and can, thus, provide additional benefits in the intensity estimation process. Adopting the HMT allows us to capture the intrinsic hierarchical structure of the data and at the same time exploit efficient scale-recursive tree-structured algorithms. While using the HMT model in conjunction with photon-limited imaging has been previously suggested by [20], its wider adoption in this context has been hindered so far by lack of a satisfactory solution to the model parameter estimation problem. We address this shortcoming by extending our EM-based technique of Section IV to the HMT case.

The HMT model is similar to the hidden Markov model (HMM) which is widely adopted in speech analysis [39]. The HMT models the interscale dependencies between the splitting factors Θ indirectly, by imposing a Markov tree dependence structure between the mixture assignments represented by the discrete latent variables $\mathbf{z}_n$, defined in Section IV. More specifically, let $\boldsymbol{\mu} = \{\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\alpha}\}$ be the vector of HMT parameters corresponding to prior root node state probabilities, interscale state transition probabilities, and parameters for each mixture, respectively. Also let n index the $N+1$ nodes of the quad-tree, where $n = 0$ is the root node, $n = N$ is the last node of the finest scale, $a(n) = \lfloor (n+1)/4 \rfloor$ is n's parent and $\mathbf{x}_n^c$ is the vector of x_n's children. The adopted model satisfies the key conditional independence property that $\mathbf{z}_n$ is independent of $\mathbf{z}_{a(a(n))}$ given $\mathbf{z}_{a(n)}$. It is also assumed that given the latent variables all the observation samples are mutually independent. Based on these assumptions the *a priori* probability of the hidden state path $p(\mathbf{z}) = p(\mathbf{z}_0, \ldots, \mathbf{z}_N)$ can be expressed as

$$p(\mathbf{z}) = \prod_{m=1}^{M} \pi_m^{z_{0m}} \prod_{n=1}^{N} \prod_{m'=1}^{M} \prod_{m=1}^{M} A_{m,m',j(n)}^{z_{nm} z_{a(n)m'}} \tag{33}$$

while the conditional likelihood $p(\mathbf{x}|\mathbf{z})$ as

$$p(\mathbf{x}|\mathbf{z}) = \prod_{n=0}^{N} \prod_{m=1}^{M} \mathrm{Polya}\left(\mathbf{x}_n^c | x_n, \boldsymbol{\alpha}_{j(n),m}\right)^{z_{nm}} \tag{34}$$

where $P(z_{0m} = 1) = \pi_m$, $P(z_{nm} = 1 | z_{a(n)m'} = 1) = A_{m,m',j(n)}$ and $j(n)$ denotes the level of the tree for the corresponding node. Equation (34) relates the model parameters directly to the noisy observations in vector $\mathbf{x}$ resulting in a Polya-mixture distribution, exactly as in Section IV.

Having at hand the expressions for the prior probability of the latent variables and the conditional probabilities of the observation samples given the latent variables, we can find an expression for the complete log-likelihood $L_c(\boldsymbol{\mu})$ useful for EM-based training

$$\begin{aligned} L_c(\boldsymbol{\mu}) &= \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\mu}) = \log p(\mathbf{z}|\boldsymbol{\mu}) + \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) \\ &= \sum_{m=1}^{M} z_{0m} \log \pi_m \\ &\quad + \sum_{n=1}^{N} \sum_{m'=1}^{M} \sum_{m=1}^{M} z_{nm} z_{a(n)m'} \log A_{m,m',j(n)} \\ &\quad + \sum_{n=0}^{N} \sum_{m=1}^{M} z_{nm} \log \mathrm{Polya}\left(\mathbf{x}_n^c | x_n, \boldsymbol{\alpha}_{j(n),m}\right). \end{aligned} \tag{35}$$

As we can observe the complete log-likelihood is in a separable form with respect to the model parameters; thus, we can find each parameter by simply maximizing the corresponding term. However, as in the previous section, we do not know the values of the latent variables. Once again the standard approach is to maximize the expected value of $L_c(\boldsymbol{\mu})$ with respect to the posterior distribution of the latent variables. So in the E-step we use

Fig. 8. Set of test images used for numerical comparisons among algorithms. The images are, from left to right, Lena, Barbara, Boat, Fingerprint, and Cameraman.

the current estimate of the model parameters $\boldsymbol{\mu}^{(l)}$ to find the posterior distribution and use this value to compute the expectation of the complete data likelihood, as a function of the parameters $\boldsymbol{\mu}$. The conditional expectation of the complete log-likelihood estimated in the E-step, will be of the form

$$\begin{aligned} E\left[L_c(\boldsymbol{\mu})|\mathbf{x},\boldsymbol{\mu}^{(l)}\right] &= \sum_{m=1}^{M}\gamma_m(z_0)\log\pi_m \\ &+ \sum_{n=1}^{N}\sum_{m'=1}^{M}\sum_{m=1}^{M}\xi_{m,m'}\left(z_n,z_{a(n)}\right)\log A_{m,m',j(n)} \\ &+ \sum_{n=0}^{N}\sum_{m=1}^{M}\gamma_m(z_n)\log\mathrm{Polya}\left(\mathbf{x}_n^c|x_n,\boldsymbol{\alpha}_{j(n),m}\right). \end{aligned} \quad (36)$$

Note that in the HMT case observations from all scales are processed as a whole, in contrary to the independent case where only the observations from a single scale are considered each time. Utilizing the upward-downward algorithm [19] we can efficiently compute the conditional probability $\gamma_m(z_n) = P(z_{nm} = 1|\mathbf{x})$ which is also required in the posterior mean estimates of (25) and (26), as well as the joint state probability $\xi_{m,m'}(z_n, z_{a(n)}) = P(z_{nm} = 1, z_{a(n)m'} = 1|\mathbf{x})$. Then in the M-step new estimates for the parameters of the model are obtained. Treating $\boldsymbol{\pi}$ and $\mathbf{A}$ is done as in [19], while maximizing (36) with respect to $\boldsymbol{\alpha}_{j(n),m}$ is done exactly as in the independent case we discussed in Section IV. Regularizing the solution is similarly important to achieve robustness.

At this point we note that in the case of our novel multiscale Poisson-Haar image partitioning presented in Section II-D, we employ three independent HMTs to model the interscale dependencies of horizontal, vertical and diagonal rate-ratio coefficients. Specifically, each HMT is responsible for modeling the dependencies of one of the Θ^h, Θ^v and Θ^d variables across the scales. In addition, the children vector appearing in the Polya distribution will consist of the terms from the intermediate scales, i.e., $\mathbf{x}_n^{c,o} = [x_{o_1}, x_{o_2}]$ with o denoting a particular orientation and taking one of the values (h, v, d). These terms are obtained as discussed in Section II-D and their sum results in the parent observation sample x_n. Analogously, for the separable scheme we employ two HMTs, one modeling the interscale dependencies of the $\Theta^{(1)}$ variables in the horizontal orientation and one for modeling the interscale dependencies of the $\{\Theta^{(2)}, \Theta^{(3)}\}$ variables in the vertical orientation. Finally, for the quad scheme we use only one HMT which models the dependencies of the random vector $\boldsymbol{\Theta}$. Employing independent HMTs as done for the first two schemes is also common in the 2-D wavelet transform case where an HMT is used for the wavelet coefficients at each subband orientation [40].

VI. Experiments and Applications

In order to validate the effectiveness of our proposed intensity estimation methods, we provide experimental comparisons with other competing image denoising methods. The methods under comparison belong to two different classes. In Section VI-A we compare denoising methods which are designed explicitly for dealing with Poisson data, as it also holds for our proposed schemes, while in Section VI-B we provide comparisons with methods which initially preprocess the Poisson data by variance stabilizing transforms (VST) in order to transform the noise statistics, and then apply denoising methods designed for handling homogeneous Gaussian additive noise.

All the comparisons have been performed on a set of five 8-bit gray scale standard test images, shown in Fig. 8. Their size is 512×512, apart from the 'Cameraman' image whose size is 256×256 pixels. The performance of the algorithms under comparison was examined at seven different intensity levels, corresponding to varying SNR values of Poisson noise. Specifically each image was scaled to have a maximum intensity of (1, 2, 3, 4, 5, 10, 20). Then the realizations of photon counts were simulated by using a Poisson random number generator. Since Poisson noise is signal dependent with local $SNR = \sqrt{\lambda_j(k,\ell)}$, the noise level increases as the intensity of the image decreases. The mean intensity for each selected maximum intensity, for all test images, varies in the range of (0.47–0.56, 0.94–1.11, 1.40–1.67, 1.88–2.22, 2.35–2.78, 4.70–5.55, 9.40–11.10), respectively, covering a wide range of noisy conditions.

A. Comparisons With Bayesian Methods Specifically Tailored for Poisson Data

In this section we compare the performance of the proposed intensity estimation algorithms with the Poisson denoising methods of Timmerman and Nowak (TN) [13] and of Lu, Kim and Anderson (LKA) [16]. Both methods adopt a multiscale Bayesian framework for photon-limited imaging (see Section II-B), differing in the parameter estimation method as described in Section IV.

All reported experiments were obtained following a multiscale analysis until reaching a 16×16 pixel image at the coarsest scale and refer to the shift-invariant versions of the respective methods. These shift-invariant versions were obtained by averaging 16 in the case of the 256×256 Cameraman image and 32 for all other 512×512 images out of all possible circularly shifted image estimates, except for the LKA method,

TABLE II
PHOTON-LIMITED INTENSITY ESTIMATION USING 3-MIXTURE DISTRIBUTIONS. THE METHODS UNDER COMPARISON ARE THE TN [13] AND LKA [16] METHODS AND ALL THE VARIANTS OF OUR PROPOSED METHODS. THE RESULTS ARE PRESENTED IN TERMS OF PSNR (dB) FOR VARIOUS PEAK INTENSITIES CORRESPONDING TO DIFFERENT POISSON NOISE LEVELS

	PSNR (dB) / Methods								
Image/ Peak Int.	noisy	TN	LKA	Sep IND	Quad IND	PH. IND	Sep HMT	Quad HMT	PH HMT
Cam./1	3.29	19.04	19.23	19.48	19.84	19.68	19.88	20.11	20.03
Cam./2	6.29	20.04	20.39	20.79	21.00	20.82	21.22	21.48	21.41
Cam./3	8.06	20.66	21.05	21.38	21.62	21.52	22.10	22.24	22.31
Cam./4	9.30	21.16	21.52	21.77	22.39	22.19	22.74	22.74	22.90
Cam./5	10.26	21.59	21.91	22.22	22.84	22.65	23.22	23.18	23.37
Cam./10	13.28	22.86	23.18	23.85	24.08	23.93	24.65	24.64	24.97
Cam./20	16.29	24.35	24.67	24.99	25.90	25.77	26.39	26.23	26.61
Lena/1	2.97	21.32	21.84	22.16	22.32	22.30	22.47	22.58	22.66
Lena/2	5.96	22.48	23.05	23.25	23.56	23.48	23.77	23.86	23.91
Lena/3	7.73	23.23	23.76	23.99	24.30	24.17	24.55	24.60	24.69
Lena/4	8.97	23.79	24.29	24.55	24.81	24.68	25.08	25.18	25.29
Lena/5	9.95	24.19	24.72	25.01	25.24	25.11	25.59	25.68	25.78
Lena/10	12.96	25.53	25.98	26.23	26.60	26.48	27.03	27.09	27.21
Lena/20	15.97	26.97	27.31	27.69	28.15	27.90	28.45	28.41	28.66
Boat/1	2.94	21.06	21.17	21.37	21.43	21.44	21.62	21.66	21.76
Boat/2	5.95	21.90	22.14	22.21	22.36	22.37	22.58	22.60	22.77
Boat/3	7.71	22.43	22.71	22.79	22.97	23.02	23.22	23.20	23.45
Boat/4	8.96	22.82	23.11	23.22	23.36	23.41	23.63	23.65	23.90
Boat/5	9.93	23.17	23.46	23.59	23.70	23.77	24.08	24.04	24.31
Boat/10	12.94	24.23	24.51	24.65	25.00	25.05	25.32	25.21	25.57
Boat/20	15.95	25.43	25.73	26.10	26.23	26.34	26.59	26.48	26.96
Barb./1	3.22	19.66	20.00	20.22	20.29	20.29	20.37	20.40	20.48
Barb./2	6.22	20.51	20.85	20.97	21.07	21.05	21.20	21.18	21.27
Barb./3	7.99	21.01	21.31	21.43	21.52	21.53	21.65	21.62	21.72
Barb./4	9.24	21.35	21.64	21.78	21.84	21.87	21.98	21.96	22.07
Barb./5	10.20	21.62	21.87	22.01	22.06	22.08	22.22	22.19	22.33
Barb./10	13.21	22.47	22.58	22.79	22.84	22.94	23.14	23.05	23.45
Barb./20	16.22	23.34	23.30	23.74	23.81	24.05	24.65	24.36	24.92
Fgrpt./1	2.56	16.42	16.70	17.24	17.19	17.28	17.34	17.23	17.39
Fgrpt./2	5.56	17.09	17.24	18.28	18.23	18.33	18.46	18.33	18.55
Fgrpt./3	7.33	17.70	17.68	19.02	18.97	19.08	19.27	19.09	19.36
Fgrpt./4	8.57	18.16	18.03	19.60	19.55	19.66	19.86	19.66	19.94
Fgrpt./5	9.55	18.58	18.36	20.05	20.01	20.12	20.32	20.12	20.42
Fgrpt./10	12.56	20.19	19.80	21.56	21.54	21.66	21.79	21.62	21.91
Fgrpt./20	15.58	22.02	21.79	23.13	23.15	23.26	23.32	23.19	23.46

where we used 64 circular shifts exactly as described in [16]. From our experience, using more shifts results to negligible improvements compared to the extra computational time needed. Further, since the method in [13] can only handle parameter estimation of at most three symmetric-mixture components, we also use in the following experiments three component symmetric mixture densities for direct comparisons among the models. However, our methods can equally well be applied to estimate more flexible nonsymmetric densities and an arbitrary number of mixture components, only limited by computational cost considerations and the amount of training data.

The quality of the resulting images from all methods is in terms of Peak Signal to Noise Ratio (PSNR) measured in dB and defined as $10\log_{10}(\mathrm{peak}^2/\mathrm{MSE})$ where by "peak" we denote the maximum intensity of the original clean image and MSE is the mean squared error between the restored image and the original one. In Table II the results for all test images, noisy conditions and all considered methods are presented. The PSNR is estimated using 10 independent trials in each case, in order for the performance comparisons to be less biased. As we can clearly see from the results, using the model parameters obtained fully automatically by our EM methods (with bootstrapping) with the proposed algorithms consistently yields quantitatively better results than the considered competing techniques. Specifically, our independent mixture models Sep-IND (separable), Quad-IND (quad), and PH-IND (Poisson-Haar), give on average for each image and all noisy conditions roughly 0.5–1 dB improvement over the best results of TN [13] and LKA [16]. Modeling scale dependencies with our Sep-HMT, Quad-HMT, and PH-HMT HMT-based models gives a further 0.5 dB improvement. Regarding the comparison between our corresponding separable, quad, and Poisson-Haar variants, in the case where we infer the parameters at each scale independently, the quad and Poisson-Haar model perform about the same and slightly better than the separable one. However, when we examine the corresponding HMT-models, the PH-HMT, is superior giving an average improvement of approximately 0.25 dB over the other HMT variants. The efficacy of our methods relative to the alternative techniques can be visually appreciated from the representative Cameraman and Boat denoising examples shown in Figs. 9 and 10. In the first figure, we present the denoised Cameraman images produced by the competing methods TN and LKA and our Quad-HMT variant while in the second one we present the TN, LKA, and

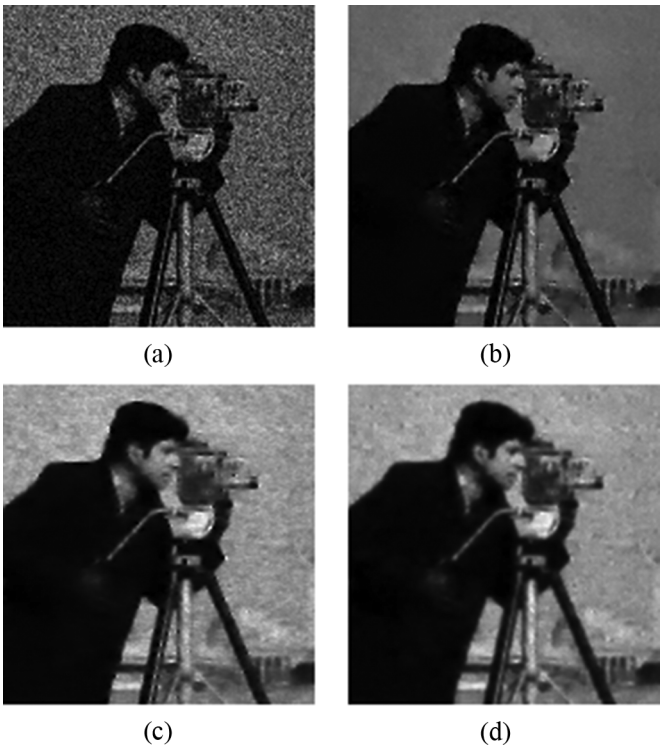

Fig. 9. Results on Cameraman image with peak intensity 20 and simulated Poisson noise. The mean number of counts per pixel in the noisy image is 9.4. Close-up of (a) Noisy image ($\mathrm{PSNR} = 16.29$), (b) our Quad-HMT result ($\mathrm{PSNR} = 26.23$), (c) TN result [13] ($\mathrm{PSNR} = 24.35$), (d) LKA result [16] ($\mathrm{PSNR} = 24.67$).

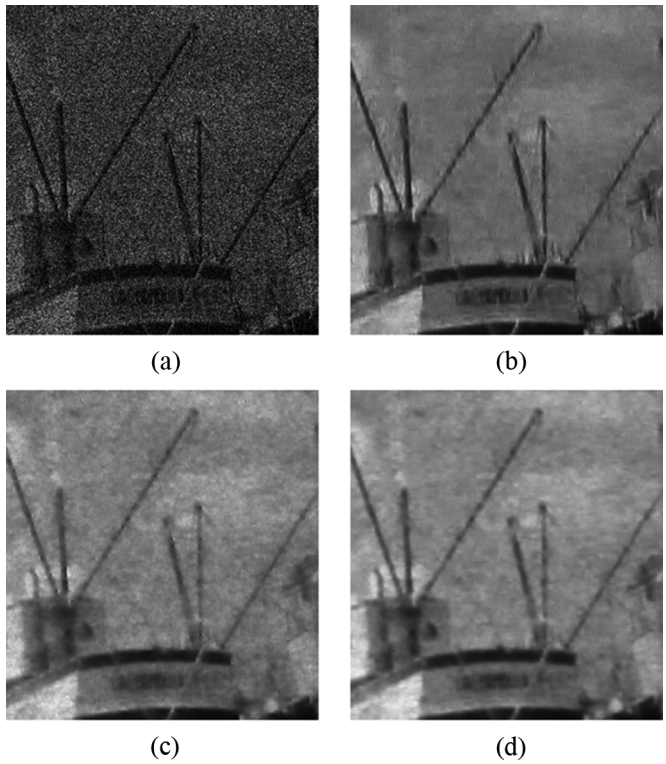

Fig. 10. Results on Boat image with peak intensity 10 and simulated Poisson noise. The mean number of counts per pixel in the noisy image is 5. Close-up of (a) Noisy image ($\mathrm{PSNR} = 12.94$), (b) our PH-HMT result ($\mathrm{PSNR} = 25.57$), (c) TN result [13] ($\mathrm{PSNR} = 24.23$), (d) LKA result [16] ($\mathrm{PSNR} = 24.51$).

our PH-HMT results for the Boat image. From these figures, we can clearly see that the proposed methods remove the noise in a more efficient manner and at the same time minimize blurring artifacts relative to the other methods. In addition, from Fig. 10, we can also verify the ability of the proposed Poisson-Haar model to better retain the edge structure of the image, due to its improved orientation selectivity. Concerning the number of used mixture-components, repeating the same experiments (results not fully reported here) using a single mixture component results in a degradation of the performance of about 1.5–2 dB, while using two mixtures presents a reduction of just 0.15–0.2 dB relative to the three mixture case. Further, using four mixtures only slightly increases performance.

Regarding the computational cost, we have measured for the Cameraman image at peak intensity 20 the run time for all tested methods. The fastest is the TN method with 8 s, while LKA needs about 60 s. Concerning the proposed methods the computational time for fully automatic training using unoptimized MATLAB code is 58 s and 45 s for the Quad-IND and Quad-HMT, respectively, 89 s and 80 s for the Sep-IND and Sep-HMT and 115 s and 92 s for the PH-IND and PH-HMT. We note that for the HMT variants no bootstrapping is used, instead the estimated parameters of the corresponding independent models serve as initial EM parameters. A significant speed-up for just a small performance loss can be achieved using two mixtures. Computational cost can be also reduced by setting directly initial EM parameters, based on previous knowledge from training on similar images.

B. Comparisons With VST-Based Methods

In this section we provide comparative results of our best-performing method, Poisson-Haar HMT model, with VST methods employing current state of the art denoising algorithms. Following the VST approach, we first stabilize the noisy data, $\mathbf{x}$, with the Anscombe transform [4] and obtain the approximately Gaussian data, $\mathbf{y} = 2\sqrt{\mathbf{x} + 3/8}$. The stabilized image $\mathbf{y}$ is denoised with a method designed for handling Gaussian data, and then we obtain the final image estimate through the inverse Anscombe transform, $\hat{\mathbf{x}} = (\hat{\mathbf{y}}/2)^2 - 1/8$. Note that the mismatch between the inverse Anscombe transform and the algebraic inversion formula $\hat{\mathbf{x}}_{alg} = (\hat{\mathbf{y}}/2)^2 - 3/8$, compensates for the bias in the mean introduced by the forward transform [4]. As Gaussian denoising techniques within the VST framework we employ two alternative state-of-the-art denoising algorithms. The first is the Bayes least squares Gaussian scale mixture (BLS-GSM) method [41] which is a parametric technique and the second is the block-matching and 3D filtering (BM3D) algorithm [42], belonging to the class of recent nonlocal, nonparametric techniques. We additionally provide comparisons of our Poisson-Haar HMT method with the multiscale VST method of [8]. As before, all results have been obtained using 10 independent trials.

From the results presented in Table III, we note that our method in the low-level counts case (low SNR) almost always performs better than BLS-GSM and BM3D. This is expected since the Anscombe VST Gaussian noise approximation is quite poor in the low SNR regime. For the case of mid-level

TABLE III
DENOISING PERFORMANCE COMPARISON OF OUR PH-HMT METHOD WITH TWO VST-BASED METHODS WHICH USE THE BLS-GSM [41] AND BM3D [42] TECHNIQUES AS GAUSSIAN DENOISING SUB-ROUTINES. THE RESULTS ARE PRESENTED IN TERMS OF PSNR (dB) FOR VARIOUS PEAK INTENSITIES CORRESPONDING TO DIFFERENT POISSON NOISE LEVELS

Methods	Peak Intensities						
	1	2	3	4	5	10	20
Cameraman							
noisy	3.29	6.29	8.06	9.30	10.26	13.28	16.29
BM3D	14.90	20.13	22.18	23.41	24.16	26.13	27.69
BLS-GSM	14.37	18.99	20.72	21.64	22.43	24.53	26.54
PH-HMT	20.03	21.41	22.31	22.90	23.37	24.97	26.61
Lena							
noisy	2.97	5.96	7.73	8.97	9.95	12.96	15.97
BM3D	16.14	22.72	24.93	25.91	26.52	28.30	29.94
BLS-GSM	15.84	21.99	24.24	25.36	26.16	28.02	29.56
PH-HMT	22.66	23.91	24.69	25.29	25.78	27.21	28.66
Boat							
noisy	2.94	5.95	7.71	8.96	9.93	12.94	15.95
BM3D	15.83	21.62	23.33	24.15	24.70	26.27	27.84
BLS-GSM	15.57	20.86	22.68	23.65	24.32	25.94	27.40
PH-HMT	21.76	22.77	23.45	23.90	24.31	25.57	26.96
Barbara							
noisy	3.22	6.22	7.99	9.24	10.20	13.21	16.22
BM3D	15.31	20.79	22.76	23.74	24.41	26.34	28.20
BLS-GSM	14.96	19.87	21.33	21.92	22.57	24.66	26.50
PH-HMT	20.48	21.27	21.72	22.07	22.33	23.45	24.92
Fingerprint							
noisy	2.56	5.56	7.33	8.57	9.55	12.56	15.58
BM3D	14.56	19.39	20.93	21.67	22.20	23.81	25.39
BLS-GSM	13.92	18.87	20.53	21.34	21.88	23.39	24.96
PH-HMT	17.39	18.55	19.36	19.94	20.42	21.91	23.46

counts for nontexture images, our method is still comparable providing similar results or even better than BLS-GSM. A difference of performance appears for the last two images, Barbara and Fingerprint, and can be attributed to the rich texture information of their content which our algorithm does not fully exploit; on the other hand, nonlocal denoising algorithms such as the BM3D are particularly effective in making good use of periodic patterns dominating texture images. Finally, at higher intensity levels (high SNR) the stabilizing transform leads to data that more closely follow Gaussian statistics, and, thus, BLS-GSM and BM3D perform better than the proposed technique.

Beyond the PSNR comparison, in order to give a sense of the typical qualitative behavior of the proposed method relative to the two VST-based techniques, we present in Fig. 11 comparative denoising results on the Lena image degraded by simulated Poisson noise with peak intensity 5, focusing on the person's face. One can see in this example that, despite the fact that our proposed method gives a lower PSNR score, it introduces fewer visual artifacts than the other two techniques. This property of our method is important in applications like astronomical and medical imaging where it is crucial that the denoising process is as faithful as possible to the visual content of the original image.

Regarding the computational cost of the VST-based techniques using either the BM3D or BLS-GSM methods as Gaussian denoising sub-routines, we have measured as in Section VI-A the execution time for the Cameraman image at peak intensity 20. The fastest execution was achieved by BM3D, 3s with optimized MATLAB/MEX code, while the run time for BLS-GSM was 18s.

Finally, we also provide comparisons of our PH-HMT method with the best of the MS-VST results on four different images reported in [8]. Since software implementing their technique is not yet available, we have applied our method at only the intensity level reported for each figure in [8]. Moreover, to allow direct comparisons, we use the same *normalized mean integrated square error* (NMISE) as quality metric, $\mathrm{NMISE} = E[(\sum_i^N (\hat{\lambda}_i - \lambda_i)^2/\lambda_i)/N]$, where $\hat{\lambda}_i$ is the estimated intensity. In the results shown in Table IV, our PH-HMT method performed better than MS-VST in 3 out of the 4 images, indicating that it is competitive with state-of-the-art VST-based methods.

C. Application to Astronomical and Medical Imaging

Our interest in intensity estimation of inhomogeneous Poisson processes is motivated by the problem of photon-limited imaging. Data acquired in low light conditions are commonly met in astronomical and medical imaging. These images often suffer from quantum or shot noise due to the variability of the detected photon counts during the image acquisition. In such cases the detected counts can be well modeled as arising from a temporally homogeneous and spatially inhomogeneous Poisson process. In this section, in order to assess the potential of our proposed methods in real-world applications, we are

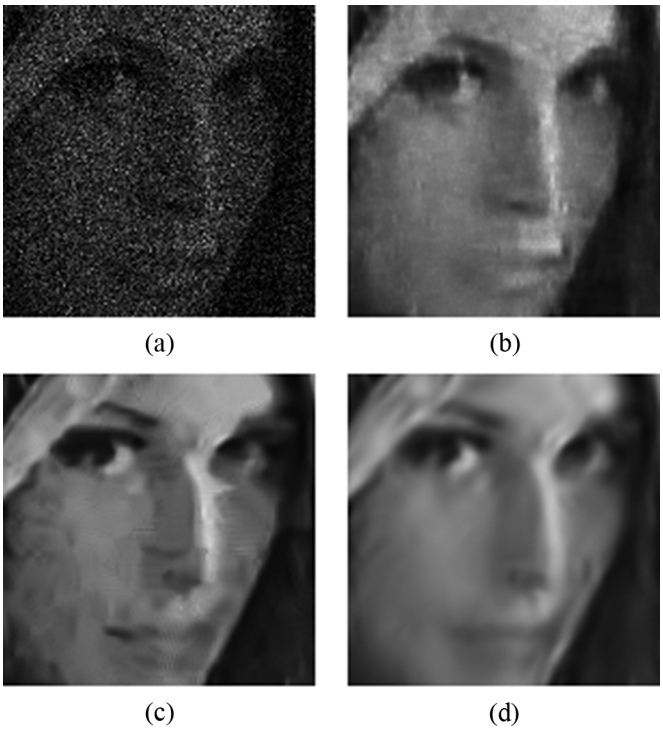

Fig. 11. Results on Lena image with peak intensity 5 and simulated Poisson noise. The mean number of counts per pixel in the noisy image is 2.5. Close-up of (a) Noisy image ($\mathrm{PSNR} = 9.95$), (b) our PH-HMT result ($\mathrm{PSNR} = 25.78$), (c) BM3D result [42] ($\mathrm{PSNR} = 26.52$), (d) BLS-GSM result [41] ($\mathrm{PSNR} = 26.16$). Note that, despite the fact that the proposed method gives a lower PSNR score, it introduces fewer visual artifacts that the two VST-based techniques.

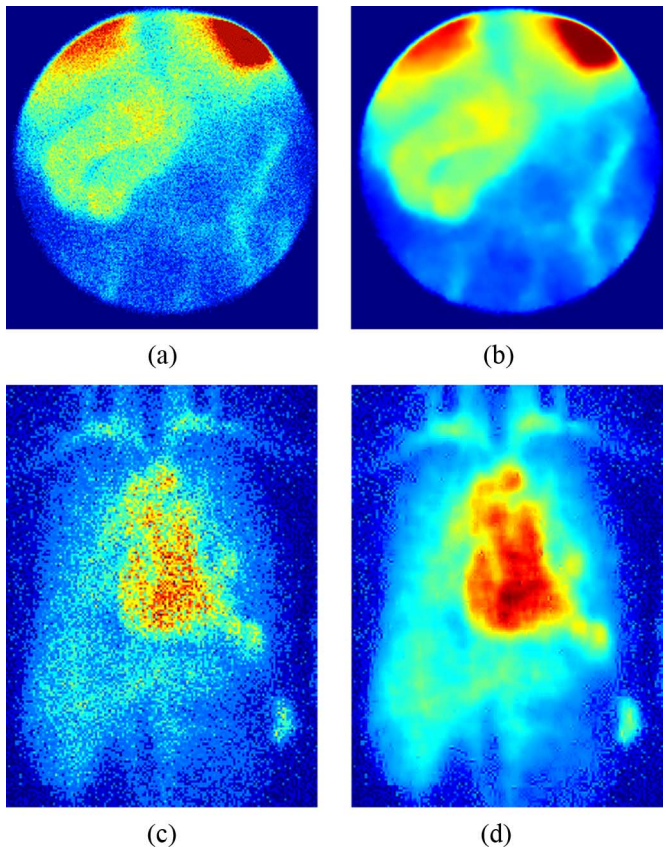

Fig. 12. Nuclear medicine image estimation. (a) Abdomen raw data counts, (b) image intensity estimation by the proposed Quad-HMT model, (c) chest raw data counts, (d) image intensity estimation by the proposed PH-HMT model.

presenting intensity estimation results from photon-limited astronomical and nuclear medical images. For each image we illustrate the indicative estimate obtained by only one of our proposed methods even though all of them produce comparable results.

In Fig. 12, the raw data counts along with the image estimates for two nuclear medicine images obtained from [43], are presented. Fig. 12(a) depicts the raw counts of an abdomen image while 12(b) presents the image estimate derived by applying our Quad-HMT model. The mean number of counts in the raw image is 3.5. This implies that the degradation of the image is significant, a fact that can be also verified visually. Fig. 12(c) shows the raw data counts from a chest image. In this case the mean number of counts is 5.4. Fig. 12(d) depicts the intensity image estimate obtained by applying the proposed PH-HMT model. From the presented figures we see that the proposed models succeed in improving the quality of the degraded images and at the same time preserve the image structures. In Fig. 13, two astronomical images with low level counts obtained from [44], are presented with their corresponding intensity estimates. Fig. 13(a) and (b) shows the photon limited version and the image estimate by the Sep-HMT model of the Messier 100 spiral galaxy, also known as NGC 4321. In Figs. 13(c) and (d), the photon limited version of the galaxy M51A along with the image estimate by the PH-HMT model is depicted. In both cases we can conclude that our models perform satisfactorily and considerably improve the visual quality of the original raw ones without affecting important structural details.

TABLE IV
DENOISING PERFORMANCE IN TERMS OF NMISE (SMALLER VALUE INDICATES BETTER PERFORMANCE) ON FIGS. 3 (SPOTS), 4 (GALAXY), 6 (BARBARA), AND 7 (CELLS) FROM [8]. WE COMPARE OUR PH-HMT RESULT WITH THE BEST OF THE MS-VST RESULTS REPORTED IN [8]

Methods	Images/Intensity range			
	Spots $[0.03\,,\,5.02]$	Galaxy $[0\,,\,5]$	Barbara $[0.93\,,\,15.73]$	Cells $[0.53\,,\,16.93]$
noisy	1.002	1.002	0.999	1.002
MS-VST	0.069	0.035	0.170	0.078
PH-HMT	0.048	0.030	0.159	0.082

VII. DISCUSSION AND CONCLUSIONS

In this paper we have presented an improved statistical model for intensity estimation of Poisson processes, with applications to photon-limited imaging. We have built on previous work, adopting a multiscale representation of the Poisson process which significantly simplifies the intensity estimation problem. Extending this framework, we have provided an efficient and robust algorithm to infer the necessary model parameters *directly* from the observed noisy data. These parameters are crucial for the efficacy of the model as we demonstrated experimentally. We have further proposed a novel multiscale representation which better models the image edge structure at different orientations, thus yielding further improvements. We also considered refined versions of these multiscale schemes that take into account dependencies across scales, instead of considering each scale of analysis as independent to the others.

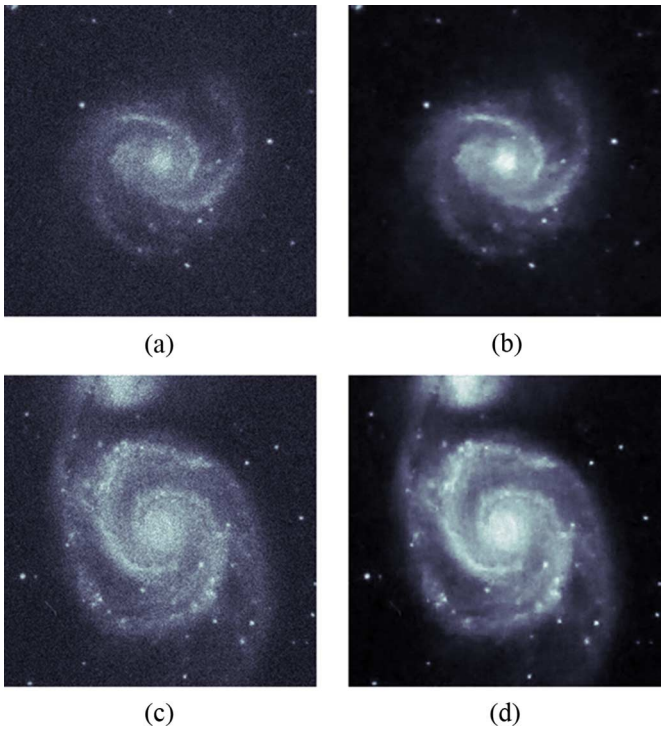

Fig. 13. Astronomical image estimation. (a) Photon limited image of Messier 100 spiral galaxy, (b) image intensity estimation by the proposed Sep-HMT model, (c) photon limited image of M51A galaxy, (d) image intensity estimation by the proposed PH-HMT model.

We extended our parameter estimation methods to also work with this HMT case.

The performance and practical relevance of the proposed inference methods and the Poisson-Haar multiscale representation scheme has been illustrated through comparisons with state-of-the art methods, where our proposed techniques have shown to produce competitive results, especially in the case of low level counts. Finally, we have applied our proposed methods to photon limited imaging problems and in particular to astronomical and nuclear medical imaging. The results we have obtained are quite satisfactory. In our future work we intend to investigate whether the discussed models and methods can be applicable to related problems such as image feature detection and segmentation under low light conditions.

Appendix A
Derivation of Poisson Likelihood Factorization on the Quad-Tree

Let us assume that the image $\mathbf{x}$ at the finer scale of analysis is raster-scanned into a vector of size $N_1 \times N_2$, that is $\mathbf{x}_0 = [x_0(0,0),\ldots,x_0(0,N_2-1),\ldots x_0(N_1-1,N_2-1)]$. Since all the random variables $X_0(k,\ell)$ are assumed conditionally independent, the likelihood function $p(\mathbf{x}|\boldsymbol{\lambda})$ can be written as

$$p(\mathbf{x}|\boldsymbol{\lambda}) = \prod_{k=0}^{N_1-1}\prod_{\ell=0}^{N_2-1} p\left(x_0(k,\ell)|\lambda_0(k,\ell)\right). \tag{37}$$

In addition, since at every scale of analysis the produced observations $x_j(k,\ell)$ correspond to the observation samples of conditionally independent Poisson random variables $X_j(k,\ell)$, it will also hold that

$$p\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right) = \frac{\prod_{(k',\ell')\in C_{k,\ell}} p\left(x_{j-1}(k',\ell')|\lambda_{j-1}(k',\ell')\right)}{p\left(x_j(k,\ell)|\lambda_j(k,\ell)\right)} \tag{38}$$

which results in

$$p\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right) p\left(x_j(k,\ell)|\lambda_j(k,\ell)\right) = \prod_{(k',\ell')\in C_{k,\ell}} p\left(x_{j-1}(k',\ell')|\lambda_{j-1}(k',\ell')\right). \tag{39}$$

The Poisson likelihood, with the help of (38) and (39) can be re-written as

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\lambda}) &= \prod_{k=0}^{(N_1/2)-1}\prod_{\ell=0}^{(N_2/2)-1}\prod_{(k',\ell')\in C_{k,\ell}} p\left(x_0(k',\ell')|\lambda_0(k',\ell')\right) \\ &\overset{(39)}{=} \prod_{k=0}^{(N_1/2)-1}\prod_{\ell=0}^{(N_2/2)-1} p\left(\mathbf{x}_1^c(k,\ell)|x_1(k,\ell)\right) \\ &\quad\cdot \prod_{k=0}^{(N_1/2)-1}\prod_{\ell=0}^{(N_2/2)-1} p\left(x_1(k,\ell)|\lambda_1(k,\ell)\right). \end{aligned} \tag{40}$$

Following the same procedure iteratively on the second product term of the right hand-side, we finally obtain

$$p(\mathbf{x}|\boldsymbol{\lambda}) = p\left(x_J(0,0)|\lambda_J(0,0)\right) \cdot \prod_{j=1}^{J}\prod_{k=0}^{\frac{N_1}{2^j}-1}\prod_{\ell=0}^{\frac{N_2}{2^j}-1} p\left(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)\right) \tag{41}$$

with

$$p(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)) = \mathrm{Mult}(\mathbf{x}_j^c(k,\ell)|x_j(k,\ell),\boldsymbol{\theta}_j(k,\ell)).$$

We have, thus, proved the factorization of the likelihood over multiple scales.

Appendix B
Equivalence of MMSE Estimation in the Image and the Transformation Domain

In this section we prove the equivalence between the MMSE estimation in the image and the transformation domain. The proof concerns the 1-D estimator, but extends analogously to the 2-D estimator.

Let us indicate by $\boldsymbol{\lambda} = \lambda_J(0)\cdot\boldsymbol{\Phi}(\boldsymbol{\theta})$ the one-to-one mapping of the N-sized intensity vector $\boldsymbol{\lambda}$ in the image domain with the re-parameterized vector $\boldsymbol{\lambda}_e = [\lambda_J(0),\boldsymbol{\theta}]$ in the transformation domain, where $\boldsymbol{\Phi}(\boldsymbol{\theta}) = [\Phi_1(\boldsymbol{\theta})\ldots\Phi_N(\boldsymbol{\theta})]^T$, $\Phi_i(\boldsymbol{\theta}) = \prod_{j=1}^{J} g_{i,j}(\theta_{i,j})$, and $g_{i,j}(\theta_{i,j})$ equals to either $\theta_{i,j}$ or $1-\theta_{i,j}$ with $\theta_{i,j}$ being the ancestor of the λ_i pixel in scale j. Based on

this notation and the factorization property of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ in (20) we can express the posterior estimator for each component of $\boldsymbol{\lambda}$, λ_i as

$$\begin{aligned}\hat{\lambda}_i &= \int_{\boldsymbol{\Lambda}} \lambda_i p_{\boldsymbol{\Lambda}|\mathbf{X}}(\boldsymbol{\lambda}|\mathbf{x}) d\boldsymbol{\lambda} = \hat{\lambda}_J(0) \int_{\boldsymbol{\Theta}} \Phi_i(\boldsymbol{\theta}) p_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \\ &= \hat{\lambda}_J(0) \prod_{j=1}^{J} \int_0^1 g_{i,j}(\theta_{i,j}) p(\theta_{i,j}|\mathbf{x}) d\theta_{i,j} \\ &= \hat{\lambda}_J(0) \prod_{j=1}^{J} g_{i,j}(\hat{\theta}_{i,j}) = \hat{\lambda}_J(0) \Phi_i(\hat{\boldsymbol{\theta}}). \end{aligned} \tag{42}$$

We have, thus, proved the equivalence on the MMSE estimation for a specific element λ_i of $\boldsymbol{\lambda}$, in the image and the transformation domain. This implies the general equivalence $\hat{\boldsymbol{\lambda}} = \hat{\lambda}_j(0) \cdot \boldsymbol{\Phi}(\hat{\boldsymbol{\theta}})$.

Appendix C
Penalized EM Estimation

In this section we assume that the splitting factors $\boldsymbol{\Theta}$ obey a Dirichlet-mixture distribution, which also covers the beta distribution as a special case.

The Dirichlet distribution belongs to the exponential family and, thus, can be written in the standard form [26]

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = f(\boldsymbol{\theta}) g(\boldsymbol{\alpha}) e^{\phi(\boldsymbol{\alpha})^T u(\boldsymbol{\theta})} \tag{43}$$

with

$$f(\boldsymbol{\theta}) = 1, g(\boldsymbol{\alpha}) = 1/B(\boldsymbol{\alpha})$$

where

$$B(\boldsymbol{\alpha}) = \Pi_{t=1}^{D} \Gamma(\alpha_t) / \Gamma(\sum_{t=1}^{D} \alpha_t), \phi(\boldsymbol{\alpha}) = \begin{pmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_D - 1 \end{pmatrix}$$

and

$$u(\boldsymbol{\theta}) = \begin{pmatrix} \ln \theta_1 \\ \vdots \\ \ln \theta_D \end{pmatrix}.$$

Since for any member of the exponential family there exists a conjugate prior that can be written in the form

$$p(\boldsymbol{\alpha}|\mathbf{v}, \eta) \propto g(\boldsymbol{\alpha})^{\eta} e^{\phi(\boldsymbol{\alpha})^T \mathbf{v}} \tag{44}$$

a suitable conjugate prior distribution for the parameters $\boldsymbol{\alpha}$ of the Dirichlet is

$$p(\boldsymbol{\alpha}|\mathbf{v}, \eta) \propto \frac{1}{B(\boldsymbol{\alpha})^{\eta}} e^{-\sum_{t=1}^{D} v_t \alpha_t}. \tag{45}$$

Adding the resulting log-prior term $\log p(\boldsymbol{\alpha}|\mathbf{v}, \eta)$ to the complete log-likelihood (29) yields the MAP (penalized ML) instead of the standard ML solution [28]. This amounts to adding to the right hand side of (30) the penalty

$$-\eta \log B(\boldsymbol{\alpha}) - \sum_{m=1}^{M} \sum_{t=1}^{D} \alpha_m^t v_m^t + (\mathrm{const}). \tag{46}$$

By selecting $\eta = 0$ and $v_m^t = \epsilon$, we obtain the simplified regularization term $-\epsilon \sum_{m=1}^{M} \sum_{t=1}^{D} \alpha_m^t$ which provides us with a robust solution in the optimization problem at the M-step of the EM algorithm.

Acknowledgment

The authors would like to thank B. Zhang for providing the images used for comparisons in Table IV. They would also like to thank the anonymous reviewers for their comments and suggestions which have considerably improved the paper.

References

[1] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. 1, pp. 113–122, 1982.
[2] D. L. Snyder and A. M. Hammoud, "Image recovery from data acquired with a charge-coupled-device camera," *J. Opt. Soc. Amer. A*, vol. 10, no. 5, pp. 1014–1023, 1993.
[3] H. Barett, "Objective assessment of image quality: Effects of quantum noise and object variability," *J. Opt. Soc. Amer. A*, vol. 7, no. 7, pp. 1266–1278, 1990.
[4] F. J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3, pp. 246–254, 1948.
[5] M. Fisz, "The limiting distribution of a function of two independent random variables and its statistical application," *Colloq. Math.*, vol. 3, pp. 138–146, 1955.
[6] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," in *Proc. Symp. Applied Mathematics*, 1993, vol. 47, pp. 173–205.
[7] P. Fryźlewicz and G. P. Nason, "A Haar-Fisz algorithm for Poisson intensity estimation," *J. Comput. Graph. Statist.*, vol. 13, pp. 621–638, 2004.
[8] B. Zhang, M. Fadili, and J.-L. Starck, "Wavelets, ridgelets and curvelets for Poisson noise removal," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1093–1108, Jul. 2008.
[9] J.-L. Starck, E. Candès, and D. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 131–141, Jun. 2002.
[10] E. Candès and D. Donoho, "Ridgelets: The key to high dimensional intermittency?," *Philosoph. Trans. Roy. Soc. Lond. A*, vol. 357, pp. 2495–2509, 1999.
[11] E. D. Kolaczyk, "Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds," *Statist. Sin.*, vol. 9, pp. 119–135, 1999.
[12] R. D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *IEEE Trans. Image Process.*, vol. 8, no. 5, pp. 666–678, May 1999.
[13] K. Timmerman and R. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 846–862, Mar. 1999.
[14] E. Kolaczyk, "Bayesian multiscale models for Poisson processes," *J. Amer. Statist. Assoc.*, vol. 94, no. 447, pp. 920–933, 1999.
[15] R. D. Nowak and E. D. Kolaczyk, "A statistical multiscale framework for Poisson inverse problems," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1811–1825, May 2000.
[16] H. Lu, Y. Kim, and J. Anderson, "Improved Poisson intensity estimation: Denoising application using Poisson data," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1128–1135, Aug. 2004.
[17] D. N. Esch, A. Connors, M. Karovska, and D. A. Van Dyk, "An image restoration technique with error estimates," *The Astrophys. J.*, vol. 610, pp. 1213–1227, 2004.
[18] S. Lefkimmiatis, G. Papandreou, and P. Maragos, "Photon-limited image denoising by inference on multiscale models," in *Proc. Int. Conf. Image Processing*, 2008.
[19] M. Crouse, R. Nowak, and G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[20] R. Nowak, "Multiscale hidden Markov models for Bayesian image analysis," in *Bayesian Inference in Wavelet Based Models*, B. Vidacovic and P. Muller, Eds. New York: Springer Verlag, 1999.
[21] J.-M. Lafertè, P. Pèrez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 390–404, Mar. 2000.

[22] A. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.
[23] S. M. Kay, *Fundamentals of Statistical Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[24] D. Karlis and E. Xekalaki, "Mixed Poisson distributions," *Int. Statist. Rev.*, vol. 73, no. 1, pp. 35–58, 2005.
[25] F. Chatelain, S. Lambert-Lacroix, and J.-Y. Tourneret, "Pairwise likelihood estimation for multivariate mixed Poisson models generated by gamma intensities," *Statist. Comput*.
[26] J. Bernardo and A. Smith, *Bayesian Theory*. New York: Wiley, 2000.
[27] N. Balahrishnan and V. B. Nevzorov, *A Primer on Statistical Distributions*. New York: Wiley, 2003.
[28] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[29] N. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*. New York: Wiley, 1997.
[30] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Lecture Notes in Statistics*. New York: Springer Verlag, 1995, pp. 125–150.
[31] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, 2nd ed. London, U.K.: Chapman & Hall, 2003.
[32] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1972.
[33] A. Gelfand and M. Ghosh, "Generalized linear models: A Bayesian view," in *Generalized Linear Models: A Bayesian Perspective*, D. Dey, S. K. Ghosh, and B. K. Mallick, Eds. New York: Marcel Dekker, 2000, ch. 1, pp. 3–22.
[34] D. J. C. MacKay and L. Peto, "An hierarchical Dirichlet language model," *Nat. Lang. Eng.*, vol. 1, no. 3, pp. 1–19, 1994.
[35] K. Sadamitsu, T. Mishina, and M. Yamamoto, "Topic-based language models using Dirichlet mixtures," *Syst. Comput. Jpn.*, vol. 38, no. 12, pp. 1771–1779, 2007.
[36] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology," *Comput. Appl. Biosci.*, vol. 12, no. 4, pp. 327–345, 1996.
[37] T. Minka, Estimating a Dirichlet Distribution Microsoft Research, 2003 [Online]. Available: http://research.microsoft.com/~minka/papers/dirichlet, Tech. Rep.
[38] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
[40] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1056–1068, Jul. 2001.
[41] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.
[42] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3d transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
[43] Division of Nuclear Medicine, Washington Univ. School of Medicine. St. Louis, MO [Online]. Available: http://gamma.wustl.edu/
[44] B. J. McLean, D. A. Golombek, Jeffrey, J. E. Hayes, and H. E. Payne, "New horizons from multi-wavelength sky surveys," in *Proc. 179th Symp. Int. Astr. Un.*, 1996, pp. 465–466, Kluwer Academic.

Stamatios Lefkimmiatis (S'08) received the Diploma degree in computer engineering and informatics (with highest honors) from the University of Patras, Patras, Greece, in 2004. He is currently pursuing the Ph.D. degree at the National Technical University of Athens (NTUA), Athens, Greece.

Since 2004, he has been a Graduate Research Assistant at the NTUA, participating in national and European research projects in the areas of image and speech analysis and microphone array processing. His research interests lie in the general areas of image analysis, statistical modeling, and nonlinear signal processing.

Petros Maragos (S'81–M'85–SM'91–F'96) received the Diploma in electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1980 and the M.Sc.E.E. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1982 and 1985.

In 1985, he joined the faculty of the Division of Applied Sciences at Harvard University, Cambridge, MA, where he worked for eight years as Professor of electrical engineering. In 1993, he joined the faculty of the ECE School, Georgia Tech. During parts of 1996–1998, he was on sabbatical and academic leave working as Director of Research at the Institute for Language and Speech Processing, Athens. Since 1998, he has been a Professor at the NTUA School of ECE. His research and teaching interests include signal processing, systems theory, pattern recognition, informatics, and their applications to image processing and computer vision, speech and language processing, and multimedia. He recently coedited a book on multimodal processing and interaction.

Dr. Maragos has served as Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING and the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and as an editorial board member for the journals *Signal Processing* and *Visual Communications and Image Representation*; general Chairman or Co-Chair of conferences or workshops (VCIP'92, ISMM'96, VLBV'01, MMSP'07); and member of IEEE SPS committees. His research has received several awards, including: a 1987 NSF Presidential Young Investigator Award; the 1988 IEEE SP Society's Young Author Paper Award for the paper "Morphological Filters"; the 1994 IEEE SP Senior Award and the 1995 IEEE W.R.G. Baker Prize Award for the paper "Energy Separation in Signal Modulations with Application to Speech Analysis"; the 1996 Pattern Recognition Society's Honorable Mention Award for the paper "Min-Max Classifiers"; 1996 election to IEEE Fellow; and the 2007 EURASIP Technical Achievements Award for contributions to nonlinear signal processing and systems theory, image processing, and speech processing.

George Papandreou (S'03) received the Diploma in electrical and computer engineering (with highest honors) in 2003 from the National Technical University of Athens (NTUA), Greece, where he is currently pursuing the Ph.D. degree.

Since 2003, he has been a Graduate Research Assistant at NTUA, participating in national and European research projects in the areas of computer vision and audiovisual speech analysis. During the summer of 2006, he visited Trinity College, Dublin, Ireland, where he worked on image restoration. From 2001 to 2003, he was with the "Demokritos" Greek National Center for Scientific Research, participating in projects on wireless Internet technologies. His research interests are in image analysis, computer vision, and multimodal processing.