J Math Imaging Vis
DOI 10.1007/s10851-009-0174-x

Spatially Varying Mixtures Incorporating Line Processes for Image Segmentation

Giorgos Sfikas · Christophoros Nikou · Nikolaos Galatsanos · Christian Heinrich

© Springer Science+Business Media, LLC 2009

Abstract Spatially varying mixture models are characterized by the dependence of their mixing proportions on location (*contextual mixing proportions*) and they have been widely used in image segmentation. In this work, Gauss-Markov random field (MRF) priors are employed along with spatially varying mixture models to ensure the preservation of region boundaries in image segmentation. To preserve region boundaries, two distinct models for a line process involved in the MRF prior are proposed. The first model considers edge preservation by imposing a Bernoulli prior on the normally distributed local differences of the *contextual mixing proportions*. It is a discrete line process model whose parameters are computed by variational inference. The second model imposes Gamma prior on the Student's-t distributed local differences of the *contextual mixing proportions*. It is a continuous line process whose parameters are also automatically estimated by the Expectation-Maximization (EM) algorithm. The proposed models are numerically evaluated and two important issues in image segmentation by mixture models are also investigated and discussed: the constraints to be imposed on the contextual mixing proportions to be probability vectors and the MRF optimization strategy in the frameworks of the standard and variational EM algorithm.

Keywords Image segmentation · Spatially varying Gaussian mixture model · Gauss-Markov random field · Maximum a posteriori (MAP) estimation · Edge-preservation · Expectation-Maximization (EM) · Variational inference

1 Introduction

Image segmentation methods relying on clustering arrange data into groups having common characteristics [1]. One of the main research directions in the relevant literature is focused on mixture models. Modeling the probability density function (PDF) of pixel attributes (e.g. intensity, texture) with finite mixture models (FMM) [2] is a natural way to cluster data because it automatically provides a grouping. The parameters of the FMM model with Gaussian components can be estimated very efficiently through maximum likelihood (ML) estimation using the Expectation-Maximization (EM) algorithm [3]. Furthermore, it can be shown that Gaussian components allow efficient representation of a large variety of PDFs. Thus, Gaussian mixture models (GMM) are commonly employed in image segmentation tasks.

In the context of image segmentation, a drawback of this approach is the difficulty to capture *spatial coherence* information, due to the over-simplifying, yet useful in terms of model tractability, hypothesis of independent data distribution. While methods to ameliorate this shortcoming have been proposed, for example by incorporating spatial coordinates in the feature vector [4], a more elegant idea is to

Giorgos Sfikas was supported by a grant from Région Alsace (France).

G. Sfikas · C. Nikou (✉)
Department of Computer Science, University of Ioannina,
45110 Ioannina, Greece
e-mail: cnikou@cs.uoi.gr

N. Galatsanos
Department of Electrical and Computer Engineering, University
of Patras, 26500 Rio, Greece

G. Sfikas · C. Heinrich
Laboratoire des Sciences de l'Image de l'Informatique
et de la Télédétection / LSIIT, UMR CNRS-ULP 7005, University
of Strasbourg, BP 67412, Illkirch cedex, France

Published online: 15 October 2009

model the data labels as a Markov random field [5–7]. MRFs are a powerful modeling tool, also employed, for instance, in image restoration [8], image super-resolution [9] and edge-preserving filtering [10].

However, inference of the posterior field distribution is typically intractable and estimation algorithms such as the computationally expensive family of the Markov chain Monte Carlo techniques [2] have to be employed. Other inference methodologies propose convenient approximations for the posterior random field, such as the pseudo-likelihood [7] or the simulated-field approximation [11]. Estimation of discrete class labeling in an MRF mesh has been successfully handled with graph theoretic approaches [12, 13], most notably graph cuts [14, 15].

An alternative to avoid the computational cost of the pixel label MRF estimation is to model the *contextual mixing proportions*, that is probabilities of the pixel labels (or the mixing proportion vector for each distinct pixel), as a Markov random field [16–21]. In such models, MAP estimation of the *contextual mixing proportions* is possible, and the computational cost is transformed from a hard posterior inference problem, as in the discrete *MRF-on-labels* model family, to a difficult constrained optimization problem. In that case, the constraint is that the *contextual mixing proportions* corresponding to a pixel must always sum up to unity as they must be probability vectors. However, as conjectured and experimentally observed in [22], an advantage for the second model would, in general, be a less sharply peaked likelihood function, leading in turn to easier model inference in terms of optimization efficiency and dependency on initialization.

Another drawback of standard MRF priors used in image recovery and segmentation is that, in general, they do not preserve boundaries between image segments as they have the tendency of smoothing neighboring pixels. In the relevant literature, *line processes* [5, 23] have been proposed. They model the presence of a boundary by a binary variable which is accordingly switched on and off. In place of this explicit line process, implemented effectively with a binary variable mesh, dual of the label/mixing proportions MRF mesh, a form of an implicit line process may be considered. Robust clique potentials can be thus used for an edge-preserving effect. The relationship between explicit & implicit line processes has been thoroughly discussed in [24].

In this work, we follow the second family of MRF methods and propose models imposing MRF smoothness priors on the *contextual mixing proportions* of a spatially varying Gaussian mixture model. Moreover, in order to account for the preservation of boundaries between image segments, we choose appropriate priors that take the form of a line process. More specifically, we propose two distinct models.

In the first model, the local differences between the *contextual mixing proportions* are normally distributed and the line processes are considered as *binary* Bernoulli distributed, with Beta conjugate hyperpriors imposed on their parameters. This model is shown to be tractable using variational inference methodology [2]. In the second model, we propose a *continuous* approach to the line process, where we use Student's-t clique functions to model the local differences between *contextual mixing proportions*. The Student's-t distribution is well-known as a robust alternative to the Gaussian distribution [25] and in this context serves as an implicit line process. However, we shall show that this setting is equivalent with an explicit line process with Gaussian-distributed cliques and Gamma-distributed line process variables. A short version of the continuous line process model has been presented in [26] and an application of the binary line process model to brain image segmentation has been presented in [27]. In this study, along with the comparison of the proposed models, we also propose solutions for two important issues in image segmentation by spatially varying mixture models. Firstly, we address the constraint that the *contextual mixing proportions* must be probability vectors. This issue is generally handled by a projection of the estimated *contextual mixing proportions* onto the simplex hyperplane at each step of the EM algorithm [17, 18]. In this paper, we propose a projection method relying on a quadratic approximation of the function involving the unknown *contextual mixing proportions*. The new approach provides more accurate results and higher values for the likelihood of the observations in the EM framework. Secondly, a new strategy for the optimization of the MRF on the image pixels is proposed. The proposed mechanism involves a multiresolution technique with overlapping pixels at each resolution level.

The main contribution of this work is the integration of a line process (continuous or discrete) with spatially varying mixtures for image segmentation and modeling, where the line process parameters are automatically computed from the data.

The remainder of the article is organized as follows. In Sect. 2 we present the background in spatially varying Gaussian mixture models. In Sect. 3 we present the edge preserving MRF priors incorporating a discrete and a continuous line process mechanism. Model inference is also described using Bayesian methodology. The issue of constraining the *contextual mixing proportions* to be probability vectors is addressed in Sect. 4 along with a new projection methodology. The new optimization strategy of the MRF sites is discussed and evaluated in Sect. 5. Experimental results in natural image segmentation are presented in Sect. 6 and conclusions are drawn in Sect. 7.

2 Background on Spatially Varying Gaussian Mixture Models

Let $X=\{x^n\}_{n=1}^N$ be the set of pixel intensities, or in general pixel feature vectors, corresponding to a single image. Viewing the required segmentation as a clustering problem on X, we can assume that the x^n are independent, identically distributed and that they are generated by a finite mixture model [28]:

$$p(x^n)=\sum_{j=1}^{J}\pi_j\phi(x^n;\theta_j)$$

where $\Pi=\{\pi_j\}_{j=1}^J$ are parameters expressing the prior probability of a pixel membership on class j, and evidently being constrained to be positive and summing to unity. The $\{\theta_j\}_{j=1}^J$ is a set of deterministic parameters controlling the shape of the "kernel" functions ϕ. Thus, there is a natural correspondence between pixel class-membership and kernels, and we can classify the pixels according to posterior class memberships (in the sense of being conditioned on the observed data X). A standard and well-known choice of kernel function is the Gaussian distribution [2, 28], with other choices for example including the Student's-t [25] or the Gamma distribution [29]. From now on we make the assumption that our data are generated by a Gaussian mixture model, and subsequently build on this by choosing appropriate prior distributions on π_j.

The J-kernel spatially varying GMM (SVGMM) [16, 18] differs from the standard GMM [2] in the definition of the mixing proportions. More precisely, in the SVGMM, each pixel x^n, $n=1,\dots,N$ has a distinct vector of mixing proportions denoted by π_j^n, $j=1,\dots,J$, with J being the number of Gaussian kernels. We call these parameters *contextual mixing proportions* to distinguish them from the mixing proportions of a standard GMM. Hence, the probability of a distinct pixel is expressed by:

$$f(x^n;\pi,\mu,\Sigma)=\sum_{j=1}^{J}\pi_j^n\mathcal{N}(x^n;\mu_j,\Sigma_j) \tag{1}$$

where $0\le\pi_j^n\le 1$, $\sum_{j=1}^J\pi_j^n=1$ for $j=1,2,\dots,J$ and $n=1,2,\dots,N$, μ_j are the Gaussian kernel means and Σ_j are the Gaussian kernel covariance matrices.

We assume that, conditioned on a hidden variable Z, pixels $X=\{x^1,x^2,\dots,x^N\}$ are independent and Gaussian-distributed:

$$p(X|Z;\mu,\Sigma)=\prod_{j=1}^{J}\prod_{n=1}^{N}\mathcal{N}(x^n;\mu_j,\Sigma_j)^{z_j^n} \tag{2}$$

where the set of $N\times J$ latent variables $Z=\{z_j^n\}_{n=1..N,j=1..J}$ is introduced to make inference tractable for the model. The hidden variables Z are distributed multinomially:

$$p(Z|\Pi)=\prod_{j=1}^{J}\prod_{n=1}^{N}(\pi_j^n)^{z_j^n} \tag{3}$$

where each z^n is a binary vector, with $z_j^n=1$ if datum n is generated by the j-th kernel and $z_j^n=0$ otherwise. It is easy to see that assumptions (2) and (3) combined lead to (1).

Considering the set of *contextual mixing proportions* Π as random variables and assuming a proper prior, we can incorporate the intuitive fact that neighboring pixels are more likely to share the same class label. We assume a Markov random field on Π, which equivalently means that Π is governed by a Gibbs distribution [5], generally expressed by:

$$p(\Pi)\propto\prod_{C}e^{-\psi_c(\Pi)} \tag{4}$$

where ψ_c is a function on clique c, called *clique potential* function in the literature, and the product is over all minimal cliques of the Markov random field.

An appropriate clique distribution choice would be to assume that the local differences of *contextual mixing proportions* follow a Gaussian distribution:

$$\pi_j^n-\pi_j^k\sim\mathcal{N}(0,\beta_{jd}^2),\quad\forall n,j,d,\ \forall k\in\gamma_d(n) \tag{5}$$

and the joint distribution on Π is given by:

$$p(\Pi;\beta)=\prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{n=1}^{N}\prod_{k\in\gamma_d(n)}\mathcal{N}(\pi_j^n;\pi_j^k,\beta_{jd}^2). \tag{6}$$

This distribution[1] treats implicitly the variates in each weight vector $\pi^n=[\pi_1^n\pi_2^n\cdots\pi_J^n]$ as independent to one another, while this is not all the case as sum-to-unity probability constraints have to be always met. In other words, (6) assigns probability mass to Π configurations that are actually impossible; such configurations are suppressed using a constrained optimization step during model learning (see Sect. 4). Similar in spirit choices of modeling the prior of a probability vector set have already been proposed in [16, 17, 19, 20]

Treatment of the constrained set Π in this indirect manner may seem inelegant, however there are reasons that deem such an approach legitimate in the current context.

[1] Note that relations (5) and (6) imply that each clique is counted twice in the product of Gaussians (6), once as a difference between sites n and k and once between k and n, for given j and d. This is equivalent to counting each clique only once. We use the current convention simply for reasons of notation brevity.

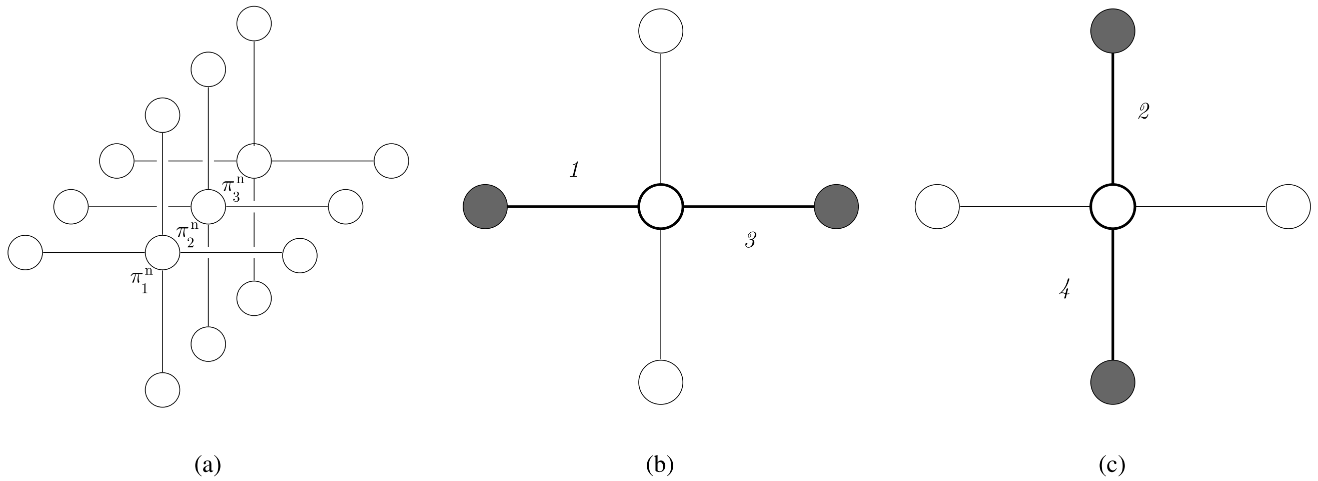

Fig. 1 First-order neighborhood cliques in the Π contextual mixing proportions mesh, used in the present algorithm implementation. (**a**) Each MRF site is associated with a probability scalar value π_j^n, and is dependent on 4 neighbors. The MRF layers for different class values j are independent to one another, reflecting (5) and (6); the sum-to-unity constraint is forced implicitly (see text). (**b**) Set of horizontal neighbors, $\gamma_1(n)$, is highlighted. (**c**) Set of vertical neighbors, $\gamma_2(n)$, is highlighted. Numbers next to links between sites in (**b**) and (**c**) correspond to ϕ function (see Sect. 3.1) values

Firstly, prior (6) rewards MRF configurations with neighboring prior weights close to one other, serving as a smoothing prior. Out of the set of admissible Π realizations, still the smoothest are given the highest probability. Secondly, the simplicity of choosing our prior to be a product of Gaussian distributions is translated later on as simple derivations of the Π-related parameters (namely β and U) on the model training phase (Sect. 3). Thirdly, absence of any straightforward choice of a distribution that would simultaneously impose smoothness of the MRF cliques and rule out inadmissible Π realizations automatically [16, 17, 19, 20].

The $J \times D$ different Gaussian distributions we have introduced in (5) amount to an equal number of parameter sets $\{\beta_{jd}\}_{j=1..J,d=1..D}$. In (5), D stands for the number of a pixel's neighborhood adjacency types and $\gamma_d(n)$ is the set of neighbors of pixel indexed n, with respect to the dth adjacency type. In our model, we assume 4 neighbors for each pixel (*first-order* neighborhood), and partition the corresponding adjacency types into horizontal and vertical, thus, setting $D = 2$ (see Fig. 1 for a detailed illustration). This variability of parameter sets aims to capture the fact that smoothness statistics may vary along clusters and spatial directions [18].

3 Edge-Preserving MRF Priors

In the current work we employ a smoothing prior for the local *contextual mixing proportion* differences. We also assume that the local differences depend on a set of hidden random variables U called in the literature *line process* [5, 23]. This configuration enables to switch on and off the smoothing property of the prior depending on whether there exists an edge or not between neighboring pixels. The general form of the model is presented in Fig. 2. The dependency between U and Π will be described analytically in the descriptions of the proposed models and for the moment it is not explicitly defined.

In any case, our goal is to find *Maximum a posteriori* estimates for Π and the deterministic parameters Ψ (the latter including here μ and Σ) that maximize the model likelihood. Thence, it is straightforward to assign each pixel to one of the J kernels which essentially will yield the desired segmentation.

We shall construct our MAP parameter estimation algorithms by making use of two powerful inference tools, namely *Expectation-Maximization* (EM) [3] and *Variational inference* (see [2]). Both of them are comprised of two analogous steps. On the first step, an estimate of the posterior distribution of the hidden variables (these include sets Z, U of Fig. 2) given the observations and current parameter estimates is computed; on the second step, new parameter estimates (these include sets Π, Ψ of Fig. 2) given the posterior of the hidden variables are computed. Typically these two steps are reiterated until convergence.

In what follows, we discuss two alternatives for defining and incorporating the line process and describe in detail how to infer the model parameters in each case.

3.1 Binary, Bernoulli Distributed Line Process Model

The clique potential functions, set by (5) and (6) for the non-edge preserving model, are now defined to be distributed as

$$\pi_j^n - \pi_j^k | u_j^{nk} = 1 \sim \mathcal{N}(0, \beta_{jd}^2), \quad \forall n, j, d, \ \forall k \in \gamma_d(n) \quad (7)$$

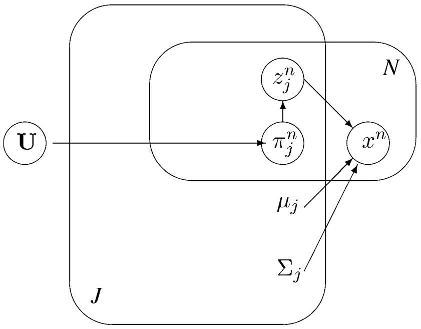

Fig. 2 General form of the graphical model for our edge preserving models. Superscript $n \in [1, N]$ denotes pixel index, subscript $j \in [1, J]$ denotes kernel (segment) index

where we assume a line process set of binary random variables $U = \{u_j^{nk}\}_{k=1..\gamma_d(n),n=1..N,j=1..J,d=1..D}$. Analytically, the distribution, conditioned on the line process, is expressed by:

$$p(\Pi|U;\beta) = \prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{n=1}^{N}\prod_{k\in\gamma_d(n)} \mathcal{N}(\pi_j^n;\pi_j^k,\beta_{jd}^2)^{u_j^{nk}}. \qquad (8)$$

This configuration assigns lower energy (higher probability) on local differences which are close to zero only when there is not an edge between them, that is when $u_j^{nk} = 1$. Otherwise, if $u_j^{nk} = 0$, the corresponding Gaussian is zeroed and therefore makes no contribution to the total MRF energy. Thus, differences are encouraged to be tightened only between pixels not separated by a boundary. We consider the *line process* binary variables u_j^{nk} to be *iid* Bernoulli distributed random variables, governed by a parameter set $\xi = \{\xi^1, \xi^2, \dots, \xi^\Gamma\}$:

$$\begin{aligned} p(U|\xi) &= \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k\in\gamma_d(n)} p(u_j^{nk}|\xi^l) \\ &= \prod_{d=1}^{D}\prod_{n=1}^{N}\prod_{k\in\gamma_d(n)} {\xi^l}^{u_j^{nk}}(1-\xi^l)^{(1-u_j^{nk})}, \qquad (9) \end{aligned}$$

where in the third product with respect to k, we have $l = \phi(n,k)$. Function $\phi(n,k)$ is defined on site indices n and k; necessarily $k \in \gamma_d(n)$ for some $d \in [1, D]$ or ϕ is undefined. Function $\phi(n,k)$ is equal to an index value in the range $[1,\Gamma]$. For fixed n, ϕ defines a one-to-one correspondence between site index k and an index $l \in [1,\Gamma]$. There are thus Γ ξ^l scalar variables, equal to the number of possible neighbors of any given MRF site. Qualitatively, this means that the Bernoulli prior is spatially invariant and only dependent to the direction to the given neighbor.

Aiming at making the line process model fully Bayesian, a Beta distribution, which is the conjugate to the Bernoulli distribution, is imposed on the ξ parameters:

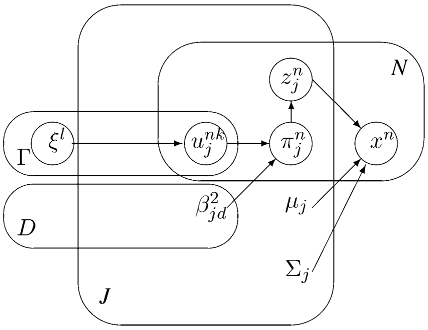

Fig. 3 Graphical model for the binary line process edge preserving model. Superscripts $n, k \in [1, N]$ denotes pixel index, subscript $j \in [1, J]$ denotes kernel (segment) index, $d \in [1, D]$ describes the neighborhood direction type and $l \in [1, \Gamma]$ denotes neighbor index

$$\begin{aligned} &p(\xi;\alpha_{\xi 0},\varpi_{\xi 0}) \\ &\quad = \prod_{l=1}^{\Gamma} \frac{\Gamma(\alpha_{\xi l0}+\varpi_{\xi l0})}{\Gamma(\alpha_{\xi l0})\Gamma(\varpi_{\xi l0})}(\xi^l)^{(\alpha_{\xi l0}-1)}(1-\xi^l)^{(\varpi_{\xi l0}-1)}, \qquad (10) \end{aligned}$$

with $\alpha_{\xi 0} = \{\alpha_{\xi l0}\}_{l=1}^{\Gamma}$, $\varpi_{\xi 0} = \{\varpi_{\xi l0}\}_{l=1}^{\Gamma}$. In order to preserve model clique symmetry, we demand that $\alpha_{\xi l0}$ have the same value for all l corresponding to the same adjacency type d; likewise for $\varpi_{\xi l0}$. In practice, if $\Gamma = 4$, as it is the case in Fig. 1, there are four components in vector ξ but they have two distinct values, one for the horizontal and one for the vertical direction ($\xi_1 = \xi_3$, and $\xi_2 = \xi_4$).

The graphical model showing the dependencies between variables for this model is presented in Fig. 3.

To perform model inference, the likelihood with respect to the model parameters Ψ and the *contextual mixing proportions* Π has to be optimized:

$$\ln p(X|\Pi;\Psi) + \ln p(\Pi;\Psi) = \ln p(X,\Pi;\Psi) \qquad (11)$$

where the deterministic parameters are $\Psi = \{\mu, \Sigma, \beta\}$. The contextual mixing probabilities Π, although being random variables, are treated as parameters and are to be optimized during inference. Thus effectively $p(\Pi)$, defined in (8), acts as a penalty term; in this sense, the proposed inference methods in this section are *Maximum a posteriori* (MAP) algorithms [2].

Calculation of (11) is however intractable and we have to resort to an estimation scheme to perform inference. In our case, the suitable framework is provided by *Variational inference* [2]. This involves finding approximations of the posterior distribution of the hidden variables, denoted by $q(Z)$, $q(U)$, $q(\xi)$, then using them to find Π and Ψ estimates that maximize the *Variational lower bound* (see 33). Details on the computation of the variational lower bound and its connection with the maximization of the model likelihood can be found in the Appendix. As it is shown in [2], optimization of the Variational lower bound $\mathcal{L}(q,\Psi,\Pi)$

boils down to updating each $\ln q(\cdot)$ to the expectation of $\langle \ln p(X,\Pi,Z,U,\xi;\mu,\Sigma,\beta)\rangle$, taken with respect to all latent variables except the one in question. In our case, this means that updates for Z, U, ξ are given by

$$\ln q(Z) = \ln p(X,Z;\mu,\Sigma) + \ln p(Z,\Pi) + \text{const.},$$

$$\ln q(U) = \ln p(\Pi|U;\beta) + \langle \ln p(U|\xi)\rangle_{\xi} + \text{const.},$$

$$\ln q(\xi) = \langle \ln p(U|\xi)\rangle_{U} + \ln p(\xi) + \text{const.}$$

After some manipulation, we obtain the update equations for the model parameters which maximize over $q(Z)$, $q(U)$, $q(\xi)$ and over Π and the deterministic parameters $\Psi = \{\mu, \Sigma, \beta\}$. The form of all q approximating-to-the-posterior functions will remain the same as the corresponding prior, as we have used conjugate priors; namely $q(Z)$, $q(U)$, $q(\xi)$ which approximate $p(Z|X,\Pi;\Psi)$, $p(U|X,\Pi;\Psi)$, $p(\xi|X,\Pi;\Psi)$ will follow the multinomial, Bernoulli and Beta distributions respectively. Also, let us note that for the q functional updates on Z and U we just provide the expected values, which are sufficient to define the distribution. The expectations—updates for $q(Z)$ and $q(U)$ along with the Beta hyperparameters are as follows:

$$\begin{aligned}
\langle z_j^n\rangle^{(t+1)} &= \frac{\pi_j^{n(t)}\mathcal{N}(x^n;\mu_j^{(t)},\Sigma_j^{(t)})}{\sum_{l=1}^{J}\pi_l^{n(t)}\mathcal{N}(x^n;\mu_l^{(t)},\Sigma_l^{(t)})},\\
\langle u_j^{nk}\rangle^{(t+1)} &= \mathrm{sig}\big(\ln\mathcal{N}(\pi_j^{k(t)};\pi_j^{n(t)},\beta_{jd}^{2(t)}) + \langle \ln\xi^l\rangle^{(t)}\\
&\qquad - \langle \ln(1-\xi^l)\rangle^{(t)}\big),\\
\langle \ln\xi^l\rangle^{(t+1)} &= \psi(\alpha_{\xi l}^{(t)}) - \psi(\alpha_{\xi l}^{(t)} + \varpi_{\xi l}^{(t)}),\\
\langle \ln(1-\xi^l)\rangle^{(t+1)} &= \psi(\varpi_{\xi l}^{(t)}) - \psi(\alpha_{\xi l}^{(t)} + \varpi_{\xi l}^{(t)}),\\
\alpha_{\xi l}^{(t)} &= \alpha_{\xi l0} + \sum_{j=1}^{J}\sum_{n=1}^{N}\langle u_j^{nk}\rangle^{(t)},\\
\varpi_{\xi l}^{(t)} &= \varpi_{\xi l0} + \sum_{j=1}^{J}\sum_{n=1}^{N}\langle 1-u_j^{nk}\rangle^{(t)},\\
&\forall n,j,d,\ \forall k\in\gamma_d(n), l=\phi(n,k),
\end{aligned} \tag{12}$$

where $\psi(\cdot)$ is the digamma function and $\mathrm{sig}(x) = (1+e^{-x})^{-1}$.

In order to learn the model for the *contextual mixing proportions* (Π), as we are using a MAP methodology, we optimize the lower bound (33) with respect to Π, always taking account of the prior (8). So setting the derivative of (33), or $\ln p(Z|\Pi) + \ln p(\Pi|U;\beta) + \text{const.}$ (defined in (3) and (8)) with respect to π_j^n to zero, we come up with π_j^n computed as the roots of the quadratic equation

$$a_j^n\left(\pi_j^{n(t+1)}\right)^2 + b_j^n\left(\pi_j^{n(t+1)}\right) + c_j^{n(t+1)} = 0, \tag{13}$$

with coefficients:

$$a_j^n = -\sum_{d=1}^{D}\left\{\beta_{jd}^{-2(t)}\sum_{k\in\gamma_d(n)}\langle u_j^{nk}\rangle^{(t)}\right\},$$

$$b_j^n = \sum_{d=1}^{D}\left\{\beta_{jd}^{-2(t)}\sum_{k\in\gamma_d(n)}\langle u_j^{nk}\rangle^{(t)}\pi_j^{k(t)}\right\},$$

$$c_j^n = \frac{1}{2}\langle z_j^n\rangle^{(t)}.$$

The form of the coefficients guarantees that there is always a non negative solution [30]. However, the solutions of (13) for a given pixel indexed by n, will not, in general, satisfy the constraints $\sum_{j=1}^{J}\pi_j^n = 1$, $\pi_j^n \geq 0, \forall j\in[1..J]$. Hence we have to perform a projection onto the constraints space. We discuss this step in more detail in Sect. 4.

Furthermore, the deterministic parameters of the model are also obtained in closed form:

$$\begin{aligned}
\mu_j^{(t+1)} &= \frac{\sum_{n=1}^{N}\langle z_j^n\rangle^{(t)}x^n}{\sum_{n=1}^{N}\langle z_j^n\rangle^{(t)}},\\
\Sigma_j^{(t+1)} &= \frac{\sum_{n=1}^{N}\langle z_j^n\rangle^{(t)}(x^n-\mu_j^{(t)})(x^n-\mu_j^{(t)})^T}{\sum_{n=1}^{N}\langle z_j^n\rangle^{(t)}}
\end{aligned} \tag{14}$$

$$\beta_{jd}^{2(t+1)} = \frac{\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)}\langle u_j^{nk}\rangle^{(t)}(\pi_j^{n(t)}-\pi_j^{k(t)})^2}{\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)}\langle u_j^{nk}\rangle^{(t)}}. \tag{15}$$

The above updates form an iterative scheme, where we have progressively better estimates $q^{(t)}$, $\Psi^{(t)}$ and $\Pi^{(t)}$ at iteration t, starting from an initial estimate $q^{(0)}$, $\Psi^{(0)}$, $\Pi^{(0)}$ and reiterating until Variational lower bound (33) convergence.

3.2 Continuous, Gamma Distributed Line Process Model

In this model, the local differences of *contextual mixing proportions* are considered to follow a univariate Student's t-distribution (one is referred to the appendix for its definition and other details). The clique potential functions are properly defined in order to impose:

$$\pi_j^n - \pi_j^k \sim \mathcal{S}t(0,\beta_{jd}^2,\nu_{jd}), \quad \forall n,j,d,\ \forall k\in\gamma_d(n), \tag{16}$$

and the joint distribution on Π is given by:

$$p(\Pi;\beta,\nu) = \prod_{d=1}^{D}\prod_{j=1}^{J}\prod_{n=1}^{N}\prod_{k\in\gamma_d(n)}\mathcal{S}t(\pi_j^n;\pi_j^k,\beta_{jd}^2,\nu_{jd}). \tag{17}$$

The distribution of the differences of local *contextual mixing proportions* thus becomes:

$$\begin{aligned}
\pi_j^n - \pi_j^k &\sim \mathcal{N}(0,\beta_{jd}^2/u_j^{nk}),\\
u_j^{nk} &\sim \mathcal{G}(\nu_{jd}/2,\nu_{jd}/2), \quad \forall n,j,d,\ \forall k\in\gamma_d(n).
\end{aligned} \tag{18}$$

This generative model (Fig. 4), apart from being tractable using the EM algorithm, allows better insight in our assumption of Student-t cliques. As a robust-to-outliers distribution, Student's-t cliques exhibit edge-preserving behavior [24]. Following the definition of the t-distribution in (36) and (37) the latent variables $U = \{u_j^{nk}\}_{n=1..N, j=1..J, d=1..D, \forall k \in \gamma_d(n)}$, may be interpreted equivalently as a continuous line process. Since u_j^{nk} depends on datum indexed by n, each weight difference in the MRF can be described by a different instance of a Gaussian distribution. Therefore, as $u_j^{nk} \to +\infty$ the distribution tightens around zero, and enforces neighboring *contextual mixing proportions* to be smooth. On the other hand, when $u_j^{nk} \to 0$ the distribution tends to be uninformative, and enforces no smoothness. Thus, the spatially varying hidden variables $U = \{u_j^{nk}\}_{n=1..N, j=1..J, d=1..D, \forall k \in \gamma_d(n)}$ are continuous line processes and may be considered as the continuous equivalent of the binary line process presented in Sect. 3.1. Consequently, in both models, the variables U provide a very detailed description of the boundary structure of the image.

Model inference is obtained by MAP estimation and under the EM algorithm framework. The incomplete data likelihood is provided by (11) while the complete data log-likelihood is expressed by:

$$\ln p(X, \Pi, Z, U; \Psi). \tag{19}$$

Let us notice that the observed data augmented by the hidden variables Z is still incomplete as the covariance matrices of the t-distributions depend also on the degrees of freedom. Therefore, the complete data vector additionally includes the missing data U. Also, like in Sect. 3.1, quantities Π are maximized in the MAP sense, and are not treated as hidden.

The conditional expectation of the complete data log-likelihood is an important quantity in the EM methodology. In this model, it is defined as:

$$\mathcal{E}_{Z,U|X,\Pi^{(t)}}\left\{\ln p(X, \Pi^{(t)}, Z, U; \Psi^{(t)})\right\}. \tag{20}$$

By optimizing the above expectation with respect to Ψ and Π, given the observed variables and some initial estimate $\Psi^{(0)}$, $\Pi^{(0)}$, we can iteratively update the estimates converging to a local optimum.

The E-step consists in computing the joint expectation of the hidden variables Z and U, with respect to the current parameters $\Pi^{(t)}$,$\Psi^{(t)}$ at iteration t. Observing the graphical model in Fig. 4, it can be seen, that, given X and Π, Z and U are conditionally independent; therefore $\mathcal{E}_{Z,U|X,\Pi}(\cdot) = \mathcal{E}_{Z|X,\Pi}\{\mathcal{E}_{U|X,\Pi}(\cdot)\}$ and we can compute these expectations separately. The updates then become $\forall n, j, d, \forall k \in \gamma_d(n)$:

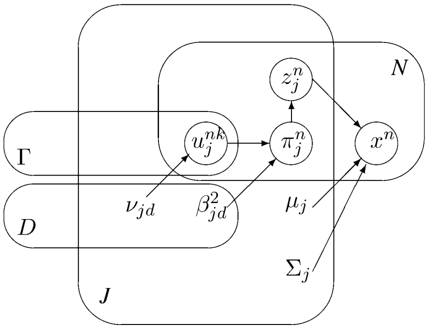

Fig. 4 Graphical model for the continuous line-process edge preserving model. Superscripts $n, k \in [1, N]$ denotes pixel index, subscript $j \in [1, J]$ denotes kernel (segment) index, $d \in [1, D]$ describes the neighborhood direction type. Γ equals the maximum number of possible neighbors

$$\begin{aligned}
\langle z_j^n \rangle^{(t)} &= \frac{\pi_j^{n(t)} \mathcal{N}(x^n; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{l=1}^J \pi_l^{n(t)} \mathcal{N}(x^n; \mu_l^{(t)}, \Sigma_l^{(t)})}, \\
\langle u_j^{nk} \rangle^{(t)} &= \zeta_j^{nk(t)} / \eta_j^{nk(t)}, \\
\langle \ln u_j^{nk} \rangle^{(t)} &= \psi(\zeta_j^{nk(t)}) - \ln \eta_j^{nk(t)},
\end{aligned} \tag{21}$$

where $\psi(\cdot)$ stands for the digamma function, and parameters ζ, η being:

$$\begin{aligned}
\zeta_j^{nk(t)} &= \frac{1}{2}\left(\nu_{jd}^{(t)} + 1\right), \\
\eta_j^{nk(t)} &= \frac{1}{2}\left(\nu_{jd}^{(t)} + \frac{(\pi_j^{n(t)} - \pi_j^{k(t)})^2}{\beta_{jd}^{2(t)}}\right).
\end{aligned}$$

Maximization of the current complete data log-likelihood (20) must be driven with respect to the model parameters Ψ and Π. With some manipulation, (20) may be split into the following terms:

$$\begin{aligned}
&\mathcal{E}_{Z|X,\Pi}\{\ln p(X|Z; \mu, \Sigma)\} + \mathcal{E}_{Z|X,\Pi}\{\ln p(Z|\Pi)\} \\
&\quad + \mathcal{E}_{U|\Pi}\{\ln p(\Pi|U; \beta)\} + \mathcal{E}_{U|\Pi}\{\ln p(U; \nu)\}.
\end{aligned}$$

In this form, parameter optimization is straightforward. The resulting update equation for the class variances is:

$$\beta_{jd}^{2(t+1)} = \frac{\sum_{n=1}^N \sum_{k \in \gamma_d(n)} \langle u_j^{nk} \rangle^{(t)} (\pi_j^{n(t)} - \pi_j^{k(t)})^2}{\sum_{n=1}^N |\gamma_d(n)|}, \tag{22}$$

where $|\gamma_d(n)|$ denotes the cardinality of the set $\gamma_d(n)$, $\forall n$. The updates for the Gaussian mean and covariances remain the same as in (14). The *contextual mixing proportions* π_j^n are also computed as the roots of a quadratic equation (13). Like in the Bernoulli prior model, we also have to perform

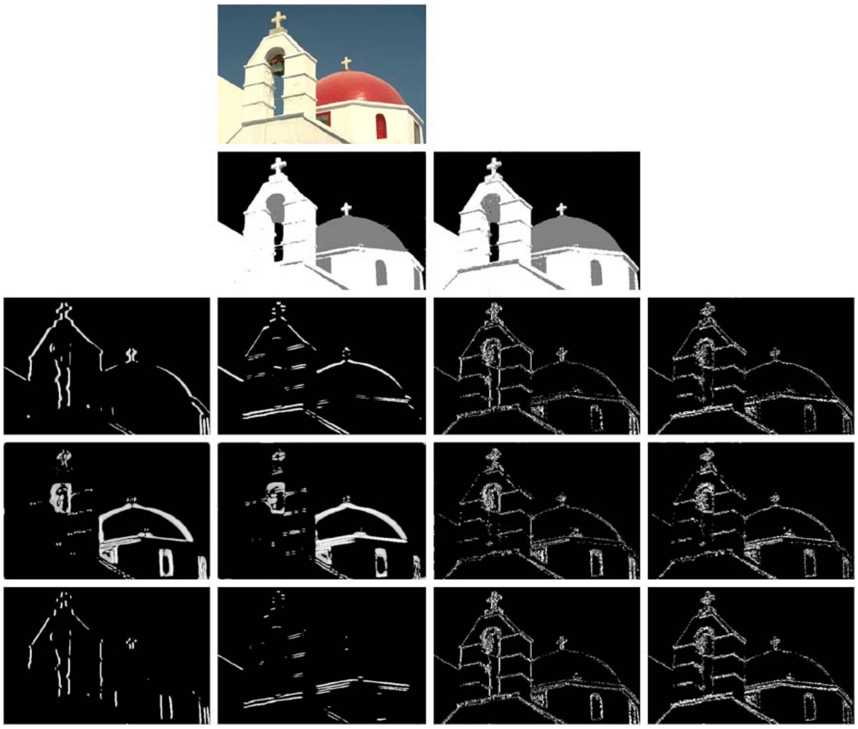

Fig. 5 U-variable maps: The *image on the top* is the original image. The segmented images for $J = 3$ clusters are presented in the *second row*, the continuous line process segmentation is on the left and the binary line process segmentation on the right. The *rows below* show U-variable maps (expected values of u_j^{nk} variables) inferred for both models. The *two columns on the left* correspond to the continuous line process model, and the two columns on the right correspond to the binary line process model. Brighter values represent lower values of u. *In each row*, the U-variable maps for kernel indexed by $j = 1$ *(sky)*, $j = 2$ *(roof and shadows) and* $j = 3$ *(building)*, are shown respectively. In each model, the *left column* corresponds to u values computed for horizontal adjacencies, and the right column for vertical adjacencies

a projection step to constrain the *contextual mixing proportions* to be probability vectors.

Finally, setting the derivative of (20) with respect to the degrees of freedom of the Student's-t distributions equal to zero we obtain $\nu_{jd}^{(t+1)}$ as the solutions of the equation:

$$\ln(\nu_{jd}^{(t+1)}/2) - \psi(\nu_{jd}^{(t+1)}/2) + \left[\frac{\sum_{n=1}^{N} \sum_{k\in\gamma_d(n)} (\langle \ln u_j^{nk}\rangle^{(t)} - \langle u_j^{nk}\rangle^{(t)})}{\sum_{n=1}^{N} |\gamma_d(n)|}\right] + 1 = 0$$

with $\psi(\cdot)$ being again the digamma function.

3.3 Insight

The inference updates computed in this section reveal a certain relation in the behavior of the two models; observe for example the similarity in the updates for β and Π in either case, see (15), (22) and (13). Let us also note that although the model based on the binary line process is solved using variational inference, this is not due to the binary nature of the line process. One could easily omit the Beta hyperprior on the Bernoulli parameters ξ and the model could also be solved by the EM algorithm. The difficulty in this model is introduced by the hyperprior that makes the computation of the expectations with respect to $p(U, \xi|X, \Pi)$ intractable. However, the introduction of the hyperprior permits the model to elegantly adjust the smoothness between pixels. A similar hyperprior is not straightforward to be imposed on the continuous line process model, due to the *degrees of freedom* parameter ν in the Student's-t prior which does not lead to a convenient form for a conjugate prior.

In both edge-preserving models, parameters U play a very important role in the preservation of the boundaries between image regions. The U-variable maps for the jth kernel represent the edges that separate the jth segment of the image from the remaining segments. To demonstrate this point, we show an example in Fig. 5. In this example, a color image is segmented into $J = 3$ segments and therefore there are 6 U-variable maps (all possible pairs of the 3 segments for the horizontal and vertical directions). The first two rows of this figure show the original and the segmented images for the continuous and binary line process priors. Moving from top to bottom, the U-variable maps for the three image segments, namely *sky, roof and shadows, building* are shown, respectively. The left column highlights vertical edges and the right column underpins horizontal edges. Notice that in the second row of the U-maps, where the U-variable maps for segment *sky* are shown, the edges between the segment *sky* and the rest (*roof and shadows, building*) are mainly highlighted. The edges between the other segments, (*roof and shadows* and *building*) are mainly highlighted in the remaining two maps. Similarly, the edges between the segments *sky* and *building* are not highlighted in the third row of images as the U-variable maps for *roof and shadows* are underpinned.

In [18] the segmentation model of [17] is extended using a class-dependent smoothness intensity parameter. This has been proven to capture variations in smoothness along classes. In the same spirit, in this work, we chose to give the line process parameters a higher flexibility. Let us furthermore note that the appearance of an edge between two pixels with a true label of class 1 and class 2 respectively, means that we need π_1 and π_2 to be discontinuous close to these points. For π_j for j other than 1 or 2, we do not need necessarily smoothness or non-smoothness imposed. The current model complies with this situation, while a single line process layer for classes would not. This is a difference with respect to other MRF based models which consider the edge structures in a class-invariant sense [21].

Let us finally note that the U-variable maps carry information about the edge structure, important in itself, that comes as a byproduct of the presented segmentation algorithms. Such information would otherwise be inaccessible if we were to use an implicit approach to edge-preservation [24] as in [17, 20]. The continuous line process model in particular can be seen as an implicit line process model, since we define it by Student's-t robust clique potentials. However due to its inference by the EM algorithm, the line process appears explicitly as a hidden EM variable and computed during model learning.

4 Projection on Constraints Hyperplane Step

The quadratic equation (13), whose non-negative solution are the *contextual mixing proportions* π_j^n is derived by maximizing the objective function:

$$\begin{aligned}&\ln p(Z|\Pi) + \ln p(\Pi|U;\beta) + \text{const.}\\&\quad = \ln \pi_j^n \sum_{n=1}^{N} \langle z_j^n \rangle + \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)} \left\{ -\frac{u_j^{nk}}{\beta_{dj}^2}(\pi_j^n - \pi_j^k)^2 \right\}\\&\qquad + \text{const.}\end{aligned} \tag{23}$$

corresponding to the variational lower bound or the complete data log-likelihood, depending on the model (continuous or binary prior). It can be easily seen, that, for a particular site n, (23) has the form:

$$x^T A x + x^T b + c \ln x + d \tag{24}$$

where we have denoted $[\pi_1^n \pi_2^n \cdots \pi_J^n]$ as x for convenience. Also, note that the above function is concave and the $J \times J$ matrix A is diagonal and negative definite.

We have already discussed that we need a maximizer for (23) also satisfying the constraints:

$$\sum_{j=1}^{J} \pi_j^n = 1, \quad \pi_j^n \geq 0,\ \forall j \in [1..J],\ \forall n \in [1..N].$$

In the general case, the solution of (13) does not satisfy the above constraints, that is, the computed contextual mixing proportion π_j^n, $j = 1, \ldots, J$ for a given pixel n are not the components of a probability vector. However, there is no straightforward way to give an exact solution to the constrained maximization of (24). This is a well-known problem, treated originally in [16] using gradient projection [31]; a projection-based solution was again given in [17], superior to [16].

Here we give a theoretical foundation to our approach—building on and generalizing the solution proposed in [17]—basing our methodology on the hypothesis that one of the terms in (23) is negligible compared to the others. An approximation of the objective function (23) is obtained by dropping the term involving the logarithm:

$$\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)} \left\{ -\frac{u_j^{nk}}{\beta_{dj}^2}(\pi_j^n - \pi_j^k)^2 \right\} + \text{const.} \tag{25}$$

In view of the fact that the objective (23) is a sum of the form *fit-to-data term* + *smoothing term* + const., our hypothesis will be valid in areas were intense smoothing is desirable, for which the smoothing term $\ln p(\Pi|U;\beta)$ will be more important than the fit-to-data term $\ln p(Z|\Pi)$. The reason smoothing priors are used in the first place is based on the exact same assumption that smoothing is desirable for the most part of an image (excluding edges et cetera). Thus we conclude that our hypothesis is reasonable, at least for the vast majority of the input image area.

Let us stress that by hypothesizing that the smoothing term is dominant to the data term we do not mean that we ignore the data term completely. If the data term did not exist, the optimum for (23) would be an homogeneous Π field, and $\pi_j^n = K^{-1}, \forall j, n$. The purpose of the proposed hypothesis is that, *given* the unconstrained solution $a^\star$ for each site in Π (by solving the second-order equation (13) to improve $a^\star$ to be as close as possible to the constrained true optimum.

Let $y^\star$ be the desired constrained maximizer of the objective function (25), and t a point on the constraints plane other than $y^\star$; let $\alpha^\star$ be the unconstrained maximizer as computed by solving (13). It can be shown that $y^\star$ will have to satisfy $(y^\star - \alpha^\star)^T A(t - y^\star) = 0$ for any plane point t. This can be expressed otherwise, as looking for y such that the projection of $\alpha' \equiv A^{\frac{1}{2}}\alpha^{\star}$ on the transformed plane defined by $t' \equiv A^{\frac{1}{2}}t$ will be $y' \equiv A^{\frac{1}{2}}y$. Thus, formally, we have the following quadratic programming problem to solve:

$$\arg\min_{y'} \|\alpha' - y'\|,$$

$$\sum_j y_j = 1, \quad y_j \geq 0, \ j = 1, \dots, J.$$

We now employ an active set type method as suggested in [17], allowing to derive closed form expressions for the Lagrange multipliers. The associated Lagrange function is given by:

$$L(y, \lambda_0, \lambda) = \frac{1}{2}\sum_{j=1}^{J}(b_j y_j - b_j \alpha_j)^2 - \lambda_0\left(\sum_{j=1}^{J} y_j - 1\right) - \sum_{j=1}^{J} b_j^2 \lambda_j y_j$$

where λ_0 is the multiplier for the equality, and λ_j, $j = 1, \dots, J$ are the multipliers for the inequality constraints. We also used the representation of $y^{\star}$, $\alpha^{\star}$ as $[y_1 y_2 \cdots y_J]$ and $[\alpha_1 \alpha_2 \cdots \alpha_J]$ respectively. Parameters b_j are the diagonal elements of the Hessian matrix A:

$$b_j = \sqrt{\sum_{d=1}^{D} \sum_{k \in \gamma_d(n)} u_j^k \beta_{dj}^{-2}}$$

where we have omitted the n data index from b and u for convenience. First-order necessary conditions imply:

$$y_j = \alpha_j + \frac{\lambda_0}{b_j^2} + \lambda_j \tag{26}$$

and injecting it into the equality constraint yields:

$$\lambda_0 = \frac{1}{\sum_j b_j^{-2}} - \frac{\sum_j \alpha_j}{\sum_j b_j^{-2}} - \frac{\sum_j \lambda_j}{\sum_j b_j^{-2}}. \tag{27}$$

Finally, by combining (26) and (27) we obtain:

$$y_j = \alpha_j - c_j + c_j \sum_{l=1}^{J} \alpha_l + c_j \sum_{l=1}^{J} \lambda_l + \lambda_j \tag{28}$$

where $c_j \equiv -\frac{b_j^{-2}}{\sum_{l=1}^{J} b_l^{-2}}$.

Let us notice that the vector $\alpha_j - c_j + c_j \sum_{l=1}^{J} \alpha_l$ is the projection of α on the constraints hyperplane $\sum_{j=1}^{J} y_j = 1$. The set of Lagrange multipliers λ_j, $j = 1, \dots, J$ must satisfy the inequality constraints. Karush-Kuhn-Tucker conditions [31] state that at the minimizer $y^{\star}$ we must have $\lambda_j \geq 0$ and $\lambda_j > 0$ if $y_j^{\star} = 0$ which is the active constraint.

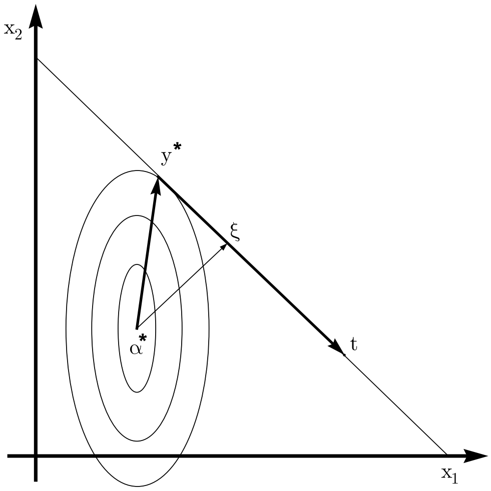

Fig. 6 Example projection to the constraints plane, in the two-dimensional case $J = 2$. *Ellipses* represent contours of the quadratic approximation to the objective function; the *line* joining the x_1 and x_2 axes is the linear constraints plane, here $x_1 + x_2 = 1$, $x_1, x_2 \geq 0$. The unconstrained maximizer is $a^{\star}$, the constrained maximizer is $y^{\star}$ and t is a point on the constraints plane. Point ξ shows the location of the solution proposed in [17]

Comparing our proposed optimizing projection with [17], we can point out that we have constructed our reasoning based on the sole hypothesis that the logarithm in (24) is a negligible quantity with respect to the other terms; this provided, our method will necessarily give the correct constrained optimum. On the contrary the projection in [17] is presented as a rather *ad hoc* solution to the problem, based on no underlying justification for this specific projection choice. Note also that this latter method could be seen as a subcase of our own proposal, for $b_1 = b_2 = \cdots b_J$.

To evaluate the proposed algorithm we have compared it to the algorithm in [17]. We have segmented the color image in Fig. 5 with the proposed model with the continuous line process prior. The resulting comparison revealed that the new algorithm provides consistently higher values for the data likelihood (Fig. 7).

5 MRF Optimization Strategy

In both models considered in this paper, we have to maximize a quantity with respect to the *contextual mixing proportions* Π. In the case of the discrete prior it is the variational lower bound and in the case of the continuous line process prior it is the complete data log-likelihood in the framework of the EM algorithm. A simple and straightforward implementation would be to perform a raster scan for each pixel $n \in [1..N]$ in order to update the sites sequentially; this involves solving J quadratic equations for each site and then projecting the resulting π_j^n vector onto the constraints

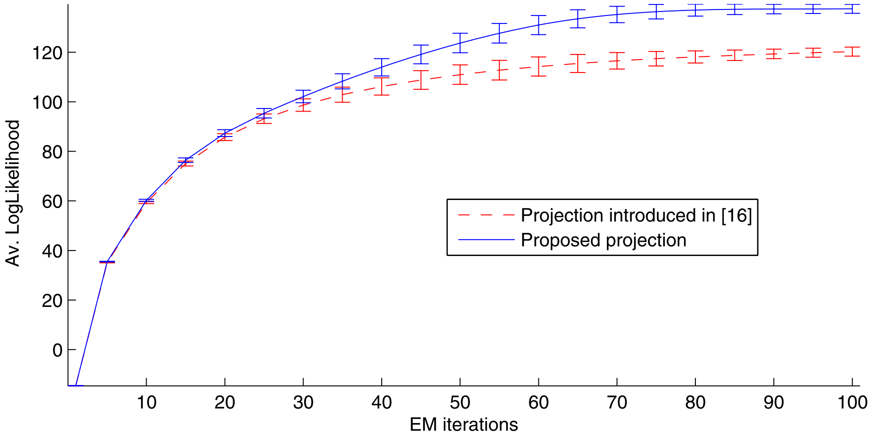

Fig. 7 Comparison of data likelihood values for the projection method in [17] and the algorithm proposed in this section: The test image in Fig. 5 was segmented into three classes using the proposed *continuous line* process prior algorithm as described in Sect. 3.2. The *solid curve* shows our results using the proposed projection against the results using the projection proposed in [17], shown by the *dashed curve*. For each configuration, we ran the segmentation 10 times using k-means initialization perturbed by additive white Gaussian noise of 0.2 units standard deviation. Likelihood values (averaged over number of pixels N and over the 10 different initializations) are shown for the first 100 EM iterations

Algorithm 1 Projection on constraints hyperplane

```
1 Let $y$ denote the vector at the current iteration. Initially,
  we set $y_j \leftarrow b_j, \forall j = 1, 2, \dots, J$. In the general case,
  there exist $m$ negative components $y_j$. The corresponding
  set of indices $S = \{j, \text{ with } y_j < 0\}$ constitutes the active
  set of constraints for the current vector $y$.
2 $\forall j \notin S$, set $\lambda_j \leftarrow 0$.
3 $\forall j \in S$, set $y_j = y_j^\star \leftarrow 0$ and we compute the corre-
  sponding $\lambda_j$ by solving an $m \times m$ linear system that
  forces the inequalities to be satisfied as equalities, namely
  $y_j + \lambda_j + c_j \sum_{l=1}^{J} \lambda_j = 0$, written in matrix form as
  $(I + \mathbf{1}c^T)\lambda = y$. The Sherman-Morisson formula [31]
  gives:
      $\lambda_j \leftarrow y_j + \frac{\sum_{l \in S} c_l y_l}{\sum_{l \notin S} c_l}.$
4 Compute the updated $y_j$ values for $j \notin S$ by (28), using
  the new vector $\lambda$.
5 Return to step 2 until convergence.
```

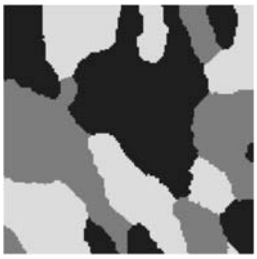

Fig. 8 A synthetic 3-class piecewise constant gray-level image, produced using a Gibbs-sampler [33]. The gray levels for each segment are 30, 125 and 220

$\sum_{j=1}^{J} \pi_j = 1$ and $\pi_j \geq 0$ using the quadratic programmatic method presented Sect. 4. This scheme would typically lead to a local maximum.

However, in practice, this local maximum is often far from the desirable segmentation result both quantitatively and visually (a related work with a detailed discussion on this issue is presented in [32]). This is due to the fact that the values of Π have a direct impact on the segmentation as the hidden variables Z depend on them. These latter variables are updated (see (21) or (12)) by:

$$z_j^n \propto \pi_j^n \mathcal{N}(x^n; \mu_j, \Sigma_j), \quad \forall n \in [1..N], \ \forall j \in [1..J]$$

In order to illustrate the importance of Π and its optimization, we have performed segmentations on a test image (Fig. 8) by applying two different initialization schemes. At first, we have used a standard k-means algorithm which is common in initializing mixture models. The second approach consisted in using as initial condition the ground truth of the image. Although it is impossible to perform the latter initialization in a real segmentation scenario, we applied it in the sense of the best initialization a segmentation method could potentially attain.

A raster scan was applied to both initialization approaches in order to sequentially optimize the parameters Π for each pixel. The results in Table 1 and Fig. 9 validate that the ground truth is indeed a local optimum for our edge-preserving algorithm. However, k-means initialization and standard raster scan MRF optimization lead to a solution that is optimal neither in terms of likelihood nor visually.

Let us consider now the Markov random field example in Fig. 10. Each site represents a vector of *contextual mixing proportions* for a certain pixel location. Consider also a step in the EM update algorithm during which the white sites have mixing proportion vectors equal to $\pi^n = z^n =$

Table 1 The RAND index [34] for the segmentations of the degraded versions of the image in Fig. 8 along different iterations of the EM algorithm are presented. Method names followed by "Π" refer to the hypothetical segmentations computed using Π instead of the hidden variables Z to classify pixels. The average data log-likelihood at the 1000th iteration is also shown

Initialization	Method	2	5	10	20	200	500	1000	Av.Lhood
k-means	[18]	0.70	0.64	0.63	0.63	0.62	0.62	0.62	29.4
	[18] (Π)	0.69	0.74	0.78	0.81	0.68	0.59	0.58	
Ground truth	[18]	0.99	0.97	0.97	0.96	0.89	0.84	0.78	28.4
	[18] (Π)	0.99	0.98	0.98	0.97	0.92	0.86	0.78	
k-means	Continuous LP	0.70	0.64	0.62	0.62	0.62	0.62	0.62	59.0
	Continuous LP (Π)	0.70	0.73	0.75	0.76	0.77	0.77	0.77	
Ground truth	Continuous LP	0.99	0.99	0.99	0.99	0.99	0.99	0.99	129.0
	Continuous LP (Π)	0.99	0.99	0.99	0.99	0.99	0.99	0.99	

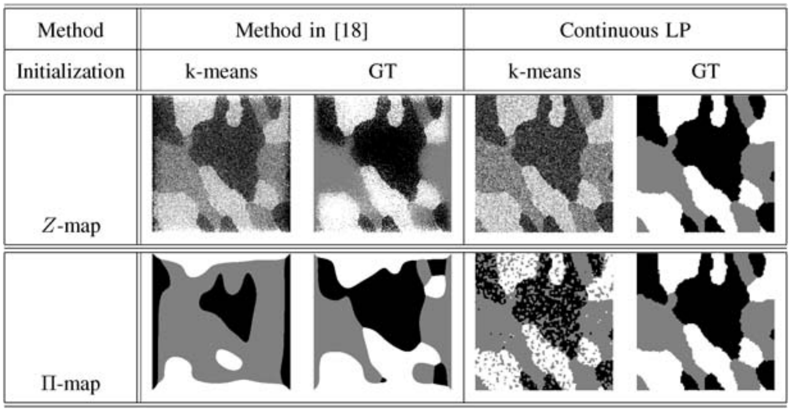

Fig. 9 Segmentation results of the 3-class synthetic image of Fig. 8 degraded by 2 dB additive white Gaussian noise after 1000 iterations. The *top row* shows the segmentations computed using the labels distribution Z to classify the pixels. The *bottom row* shows the hypothetical segmentations computed using the contextual mixing proportions Π instead of Z for classification

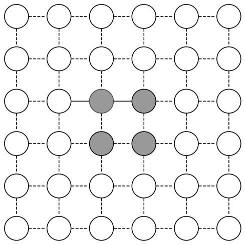

Fig. 10 An example of Markov random field of 6 × 6 sites. The color of each site corresponds to the image class the pixel is more likely to belong to

$[0.5+\epsilon, 0.5-\epsilon]^T$, with $0<\epsilon<0.5$ and the gray sites have $\pi^n = z^n = [0.5-\epsilon, 0.5+\epsilon]^T$. Consider also that we are just before updating these *contextual mixing proportions* using (13). Moreover, we make the hypothesis that the line process u_j^{nk} is constant $\forall j, n, k \in \gamma(n)$, in order not to influence the parameter updates.

Observe, that each gray site is surrounded by exactly two gray and two white neighbors and that all white sites have at most one gray neighbor each. Hence, there is a high probability that given appropriate values for β_1^2, β_2^2 and ϵ the gray sites have their π parameters updated to values closer to the values of the white sites. This will not be the case if β_j^2 are[2] such that the MRF smoothing effect is tight enough. In that case, each individual update for the gray sites will naturally leave their weights unaffected. Therefore, if the gray sites are optimized jointly higher values for the data likelihood could be obtained. Intuitively, this can be achieved by optimizing groups of pixels with the constraint of being all set to the same value.

Having in mind the continuous line process model, we extend the standard raster scan procedure to a new *grid scan* strategy which is described in Algorithm 2.

The update (30) and (31) in step 5 of the proposed algorithm are justified as follows. In each update step of a single grid S, we need to optimize:

$$\ln p(X|Z) + \ln p(Z|\Pi) + \text{const.}$$
$$= \sum_{j=1}^{J} \Bigg\{ \ln \pi_j \sum_{n\in S} (\langle z_j^n \rangle) + \sum_{d=1}^{D} \sum_{n\in S} \sum_{k\in\gamma_d(n), k\notin S} \left(-\frac{u_j^{nk}}{\beta_{dj}^2} (\pi_j - \pi_j^k)^2 \right) \Bigg\} + \text{const.}$$

with respect to $\pi_j, \forall j \in [1..J]$. We can easily conclude that the second-order equation to be solved (13) has coefficients

[2] In this example we omit the 'd' indice for clarity, assuming $D=1$.

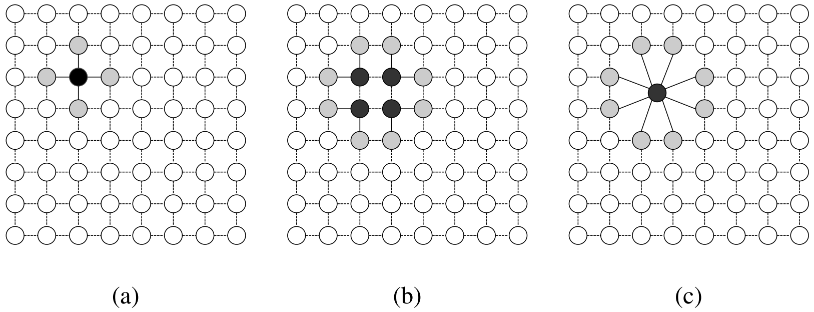

Fig. 11 Grid-scan updates on an example lattice with 8 × 8 elements and 1st order neighborhoods. *Black color* shows the elements whose contextual mixing proportions need to be updated. *Gray color* shows their neighboring pixels. (**a**) Single element to be optimized and its neighbors. (**b**) Elements to be co-optimized by a step of grid scan and their neighbors. (**c**) The same elements to be co-optimized redrawn as one

Algorithm 2 Grid scan

```
1 Calculate the initial grid size, maxLevel. This is empirically set to
  $maxLevel \leftarrow \max(\llcorner \log_2 \max(\dim X, \dim Y)\lrcorner - 3, 3)$ (29)
2 For each $L \leftarrow maxLevel$ to 1 iterate:
3 Let $subsetLength \leftarrow 2^L$. Let $G$ denote the set of sites, with $|G| = \dim X \times \dim Y$.
4 Partition the $\dim X \times \dim Y$ sites into $L$ subsets $\{S^i\}_{i=1}^L$. Also we require $\bigcup_{i=1}^L S_i = G$ and $S_i \cup S_j = \varnothing, \forall i \neq j$.
5 For each site subset $S_i$, $i = 1, \dots, L$, repeat steps 5.1, 5.2.
    5.1 For each neighborhood direction $d = 1, \dots, D$ do
        5.1.1 Define a set of sites $\tilde{\gamma}_d(S_i)$ as
              $\tilde{\gamma}_d(S_i) \triangleq \left\{ \bigcup_{s \in S_i} \gamma_d(s) \right\} \backslash S_i$
    5.2 Optimize the sites in $S_i$ by solving the quadratic equation (13) where $\langle z_j^n \rangle$ and $\gamma_d(n)$ are replaced by
        $\langle \tilde{z}_j \rangle \leftarrow \sum_{n \in S_i} \langle z_j^n \rangle$ (30)
        $\gamma_d \leftarrow \tilde{\gamma}_d(S_i)$ (31)
6 End.
```

given by:

$$a_j^n = -\sum_{d=1}^{D} \left\{ \beta_{jd}^{-2(t)} \sum_{n \in S} \sum_{k \in \gamma_d(n), k \notin S} \langle u_j^{nk} \rangle^{(t)} \right\},$$

$$b_j^n = \sum_{d=1}^{D} \left\{ \beta_{jd}^{-2(t)} \sum_{n \in S} \sum_{k \in \gamma_d(n), k \notin S} \langle u_j^{nk} \rangle^{(t)} \pi_j^{k(t)} \right\},$$

Table 2 Comparison in terms of likelihood and misclassification ratio (MCR) for the continuous line process model, between raster-scan and grid-scan optimization methods

σ	Raster-scan		Grid-scan	
	Av.Likelihood	MCR	Av.Likelihood	MCR
25	43.9	0.1%	51.9	0.13%
28	40.5	0.17%	47.5	0.18%
47	27.8	0.5%	34.6	0.5%
52	28.3	0.8%	33.5	0.6%
95	28.9	3.7%	31.5	3.2%

$$c_j^n = \frac{1}{2} \sum_{n \in S} \langle z_j^n \rangle^{(t)},$$

which makes the derivation of (30) and (31) straightforward.

To evaluate the proposed MRF optimization strategy, we computed a number of segmentations using the grid-scan versus the raster-scan optimization method. All tests were performed on noisy versions of the synthetic 3-class image (Fig. 8) using always the continuous line process prior (Sect. 3.2). In Table 2 we present a comparison of raster-scan and grid-scan algorithms in terms of model likelihood and ratio of misclassified pixels (MCR). Likelihood scores are consistently better for grid-scan for all tested noise levels. Visual result as represented with the segmentation MCR however worsens with grid-scan optimization on low-noise levels. This is justified since as the noise level decreases, the need for smoothing decreases as well and higher probability model states may well be corresponding to undesirable smoothing in the resulting segmentation. However, this is an issue of a MRF prior in general.

6 Natural Image Segmentation Results

In our implementation, we have used a 4-dimensional feature vector to describe the image data. It is comprised by the

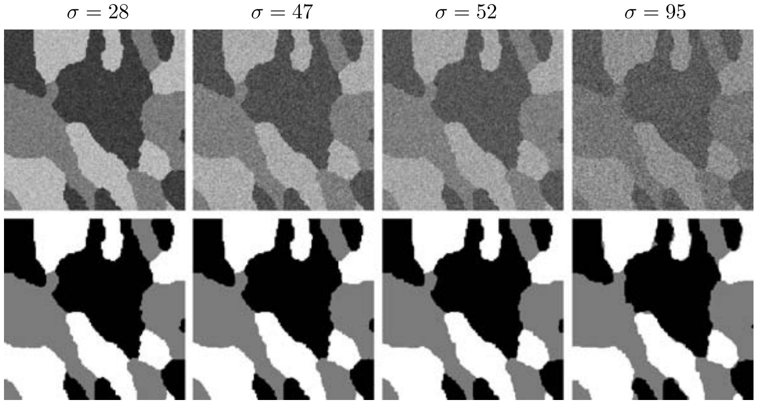

Fig. 12 *Top row*: A synthetic 3-class image degraded by white Gaussian noise, with varying standard deviations $\sigma = \{28, 47, 52, 95\}$. *Bottom row*: Corresponding segmentations using the proposed continuous line process model of Sect. 3.2 and grid-scan optimization as in Sect. 5

Lab color space features and the *Blobworld* contrast texture descriptor as described in [4]. Prior to segmentation, each variate has been separately normalized in order not to have dominating features. We also note that in the binary line process model, we let the hyperparameter values $\alpha_{\xi 0}, \beta_{\xi 0}$ of the Beta prior distribution fixed to $\alpha_{\xi k0} = \beta_{\xi k0} = 1, \forall k$. This value makes the prior effectively uninformative as the data size $N \gg 1$.

Let us also note that our algorithm requires only the determination of the number of segments J as input (which is an open issue in the machine learning community). We consider this issue as an advantage in comparison with state-of-the-art methods like the normalized cut (*ncut*) [14] and the mean-shift [35] algorithms which depend on more parameters to be defined by the user. For instance, the *ncut* algorithm strongly depends on the size of the kernel, the variance of the kernel, involved in the computation of the affinity matrix, and the number of segments. Also the mean-shift algorithm highly depends on the variance of the kernel, the size of the kernel and the termination criterion. For a given image these parameters have to be defined by the user, making straightforward comparison prone to trial-and-error procedure.

We illustrate the above considerations in Fig. 13, where we compare the continuous line process model proposed in this work with the *ncut* algorithm with varying parameter settings. Settings *ncut-1*, *ncut-2* and *ncut-3* correspond to affinity matrix kernels set to influence a progressively larger pixel area, with corresponding areas of influence 5 × 5, 10 × 10, 15 × 15. In the same figure we also compare the two methods on noise-degraded images. The result presents the advantages of our method, which due to its smoothing prior exhibits higher robustness to noise. For both the degraded and non-degraded image cases, the result for the *ncut* clearly depends heavily on the parameters used. In all, for parameter sets *ncut-1* and *ncut-3*, our method visually and numerically outperforms the *ncut* result in the church image case as well as most of the images in the boat image case.

Moreover, in the same figure, we present some cases where although the Rand index is slightly superior for the *ncut* methods, visual examination reveals that there are erroneously merged regions. More specifically, the noisy boat image segmentations for the *ncut-2* and *ncut-3* methods as well as the noise-free boat image for the *ncut-2* method provide better Rand indices with respect to our continuous line process algorithm. However, visual inspection depicts that, for instance, the beach and the sea are merged in the noise free boat image segmentation for *ncut-2*.

The results given by *ncut-1* (small affinity kernel scale) and *ncut-3* (large affinity kernel scale) suggest that the best parameter choice should correspond to a scale between the two. Indeed, *ncut-2* with kernel scale size between that of cases *ncut-1* and *ncut-3* gives the best results for *ncut*. However, choosing *a priori* such a configuration for any image is by no means obvious and the best parameters must be consequently found by trial-and-error determination.

We have evaluated the proposed continuous line-process and binary line-process segmentation schemes on the 300 images of the Berkeley image database [36]. We have applied our algorithm with different values for the number of segments $J = \{3, 5, 7, 10, 15, 20\}$. For comparison purposes, we have also experimented with the standard GMM [2] and the GMM based segmentation with "standard" smoothness constraints [18] with the same number of components.

The obtained segmentations were quantitatively evaluated with two performance measures: the Rand index (RI) [34] and the boundary displacement error (BDE) [37]. The RI measures the consistency between the ground truth and the computed segmentation map while the BDE measures error in terms of boundary displacement with respect to the ground truth. The statistics for these measures are presented in Tables 3 and 4.

Based on the theoretical properties of the edge-preservation models one might have expected that they would introduce erroneous boundaries that did not agree with human segmentation. Therefore that would provide a worse RI as compared to the "classical" non preserving algorithm (SVGMM) [18]. However, as observed in the sta-

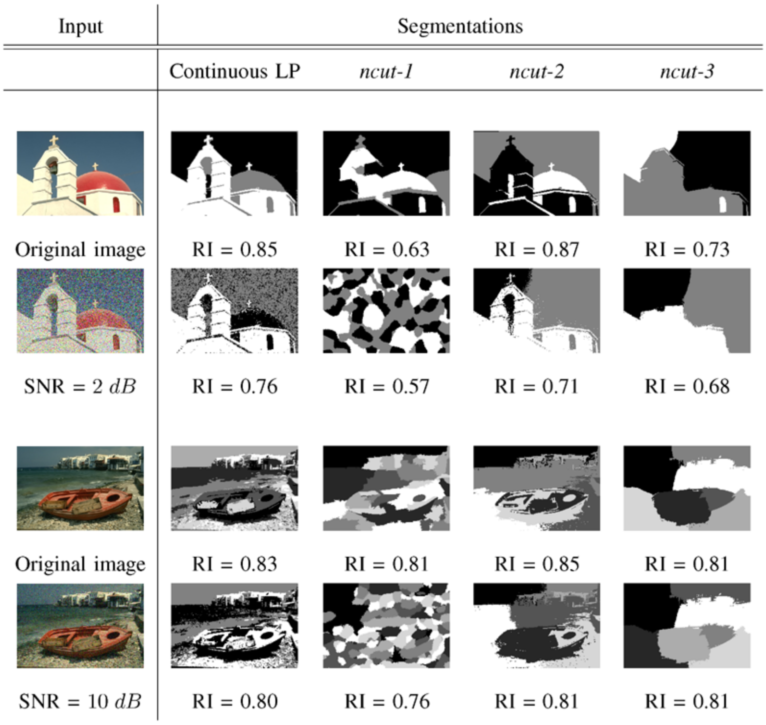

Fig. 13 Comparison of the proposed continuous line process method (3.2) with normalized cuts (*ncut*) [14]. We have tested the two algorithms on two Berkeley database images [36], as well as on noise-degraded versions of the same images. We fixed the number of classes to $J = 3$ for the Church image, and $J = 7$ for the Boat image for both algorithms. *ncut*-1 stands for the normalized cut algorithm with region of affinity kernel support and kernel variance parameters set to influence a 5 × 5 region. *ncut-2* stands for the normalized cut algorithm with region of affinity kernel support and kernel variance parameters set to influence a 10 × 10 region. *ncut*-3 stands for the normalized cut algorithm with region of affinity kernel support and kernel variance parameters set to influence a 15 × 15 region

Table 3 Statistics on the Rand Index (RI) over the 300 images of the Berkeley image data base for the compared methods. Higher values represent better segmentations

J	GMM			SVGMM			Continuous LP			Binary LP		
	Mean	Median	St. dev.	Mean	Median	St. dev.	Mean	Median	St. dev.	Mean	Median	St. dev.
3	0.675	0.680	0.085	0.686	0.690	0.085	**0.690**	**0.693**	0.087	0.690	0.693	0.087
5	0.710	0.735	0.102	0.717	0.745	0.107	**0.720**	0.7462	0.108	0.720	**0.746**	0.107
7	0.717	0.753	0.119	0.723	**0.759**	0.121	**0.724**	0.758	0.121	0.724	0.757	0.121
10	0.717	0.759	0.133	0.721	**0.760**	0.135	**0.721**	0.759	0.136	0.721	0.759	0.136
15	0.712	0.754	0.143	0.716	**0.758**	0.146	0.716	0.757	0.147	**0.717**	0.757	0.146
20	0.709	0.749	0.147	0.706	0.7452	0.153	**0.712**	**0.754**	0.152	**0.712**	0.753	0.152

Table 4 Statistics on boundary displacement error (BDE) over the 300 images of the Berkeley image data base for the compared methods. Lower values represent better segmentations

J	GMM			SVGMM			Continuous LP			Binary LP		
	Mean	Median	St. dev.	Mean	Median	St. dev.	Mean	Median	St. dev.	Mean	Median	St. dev.
3	4.789	4.164	2.386	4.787	4.206	2.397	4.612	4.043	2.302	**4.591**	**4.055**	2.287
5	4.386	3.757	2.173	4.394	3.814	2.174	4.258	3.668	2.147	**4.255**	**3.692**	2.150
7	4.244	3.708	2.095	4.212	3.683	2.055	4.125	3.594	2.053	**4.120**	**3.586**	2.055
10	4.137	3.602	2.009	4.096	3.504	1.986	**4.028**	3.495	1.999	4.040	**3.492**	2.011
15	4.010	3.635	1.976	4.034	3.504	1.940	**3.959**	**3.431**	1.954	3.955	3.408	1.967
20	4.128	3.678	2.011	4.191	3.655	1.908	**3.923**	**3.393**	1.924	3.921	3.425	1.934

tistics of the RI (Table 3), both edge preservation schemes outperform the standard GMM in all cases and the SVGMM in the overwhelming majority of the different number of components.

Also, in terms of correct region boundary estimation, expressed by the BDE (Table 4), the edge-preservation models outperform the SVGMM, as theoretically expected. However, they also outperform standard GMM and the difference in performance increases with the number of segments. The explanation for this behavior is that since the standard GMM does not integrate a smoothing step it generally computes correctly the boundaries between segments (it also outperforms the SVGMM in the same median values). However, as the number of segments increases, the complexity of the image cannot be captured by a simple GMM and smoothness constraints that model the image edge structure become increasingly beneficial.

Comparing the proposed edge-preserving priors, their performance scores are in general close. The continuous line process prior seems to give better results for the RI, while the binary line process prior gives better results for the BDE. The difference in performance is however too slight to draw a safe conclusion about the behavior of the one prior compared to the other. To illustrate this, one can observe that on RI and BDE the mean scores differ respectively by 8×10^{-5} and 8×10^{-3} (on average over the number of kernels J) between the two models. This is only a fraction of the improvement the proposed schemes exhibit over the non-edgepreserving scheme SVGMM, namely 4% and 6% for each case. Overall, the proposed schemes not only preserve region boundaries but also improve the correct classification rates with respect to the standard methods. Some representative segmentation examples for the two proposed models are shown in Figs. 14 (continuous line process model), and 15 (binary line process model).

7 Conclusion and Future Work

In this paper we have presented an image segmentation algorithm having the property of taking into account spatial relationships to classify image pixels. We have explored two alternative ways to make the model edge-preserving, which is the main contribution of the paper. We have also noted the importance of properly optimizing the Markov random field energy in the current model, and we have proposed improvements over the field optimization methods used for similar models like [16, 18]. The corresponding edge-preserving prior choices, the *binary* and the *continuous* line process priors, lead to model solutions feasible with variational inference and Expectation-Maximization respectively. We have seen that the binary line process model includes a set of fixed hyperparameters $(\alpha_{\xi 0}, \varpi_{\xi 0})$ that can affect the model's sensitivity to what is regarded as an edge; the continuous line process model is, on the other hand, computationally and conceptually simpler. The automatic estimation of model parameters from the data is crucial, as many state-of-the-art segmentation algorithms rely on empirical parameter selection. An important perspective of this study is to automatically estimate the number of components. To this end, criteria appropriate to constrained mixtures could be conceived.

Appendix A: Variational Lower Bound Derivation

The model likelihood (11) may be written as

$$\sum_{Z,U}\int_{\xi} q(Z,U,\xi)\log\frac{p(X,\Pi,Z,U,\xi;\Psi)}{q(Z,U,\xi)}d\xi$$
$$-\sum_{Z,U}\int_{\xi} q(Z,U,\xi)\log\frac{p(Z,U,\xi|X,\Pi;\Psi)}{q(Z,U,\xi)}d\xi. \tag{32}$$

The first term is called *variational lower bound* in the related literature [2], while the second is the Kullback-Leibler divergence between the posterior distribution of the latent variables conditioned on the observations and Π, and a distribution $q(\cdot)$ which represents an estimate of the posterior. It is well-known that any Kullback-Leibler divergence has a minimum at zero, and that minimum is achieved when the comparing distributions are identical. This means that (a) the first term in (32) is a lower bound of the likelihood, and (b) this bound is maximized with respect to q if and only if $q(Z,U,\xi) = p(Z,U,\xi|X,\Pi;\Psi)$. So instead of working with the likelihood, which here involves an intractable marginalization over Z, U, ξ, in variational inference, the idea is to find estimates that maximize the variational lower bound: the variational lower bound $\mathcal{L}$ is given by (33):

$$\begin{aligned}
&\mathcal{L}(q,\Psi,\Pi)\\
&\quad\triangleq \sum_{Z,U}\int_{\xi} q(Z,U,\xi)\log\frac{p(X,\Pi,Z,U,\xi;\Psi)}{q(Z,U,\xi)}d\xi\\
&\quad=\mathcal{E}_{q(Z,U,\xi)}\left(\ln\frac{p(X,\Pi,Z,U,\xi;\Psi)}{q(Z,U,\xi)}\right)\\
&\quad=\langle \ln p(X,\Pi,Z,U,\xi;\mu,\Sigma,\beta)\rangle - \langle \ln q(Z,U,\xi)\rangle.
\end{aligned} \tag{33}$$

To proceed with the computation of optimal q on $\mathcal{L}$, we must introduce here the *mean field approximation* which stems from statistical physics [2]:

$$q(Z,U,\xi) = q(Z)q(U)q(\xi). \tag{34}$$

Note that in the proposed model, we only need to assume $q(U,\xi) = q(U)q(\xi)$, as $q(Z,U,\xi) = q(Z)q(U,\xi)$ is induced from the model structure. We can thus rewrite (33) as

$$\begin{aligned}
\mathcal{L}(q,\Psi,\Pi) &= \langle \ln p(X|Z;\mu,\Sigma)\rangle + \langle \ln p(Z|\Pi)\rangle\\
&\quad + \langle \ln p(\Pi|U;\beta)\rangle + \langle \ln p(U|\xi)\rangle + \langle \ln p(\xi)\rangle\\
&\quad - \langle \ln q(Z)\rangle - \langle \ln q(U)\rangle - \langle \ln q(\xi)\rangle.
\end{aligned}$$

The expectations in (33), over the estimate posterior distribution q are given by:

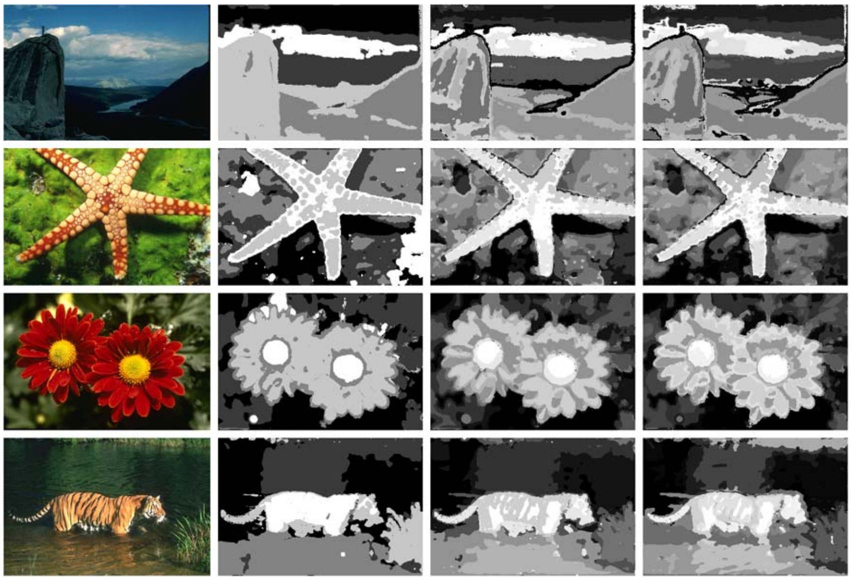

Fig. 14 Segmentation examples using the proposed continuous line process spatially variant mixture for varying number of classes J. *From left to right*, the *columns* show: the original image, segmentation with $J = 5$, $J = 10$ and $J = 15$

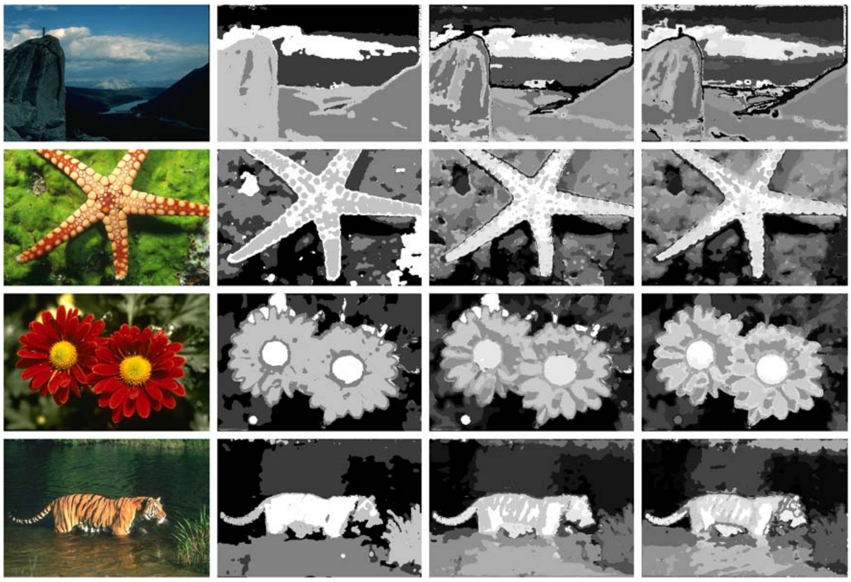

Fig. 15 Segmentation examples using the proposed binary line process spatially variant mixture for varying number of classes J. *From left to right*, the *columns* show: the original image, segmentation with $J = 5$, $J = 10$ and $J = 15$

$$\langle \ln p(X|Z;\mu,\Sigma)\rangle = \sum_{n=1}^{N}\sum_{j=1}^{J}\langle z_j^n\rangle \mathcal{N}(x^n;\mu_j,\Sigma_j),$$

$$\langle \ln p(Z|\Pi)\rangle = \sum_{n=1}^{N}\sum_{j=1}^{J}\langle z_j^n\rangle \ln \pi_j^n,$$

$$\langle \ln p(\Pi|U;\beta)\rangle$$
$$= \sum_{n=1}^{N}\sum_{j=1}^{J}\sum_{d=1}^{D}\sum_{k\in\gamma_d(n)} \langle u_j^{nk}\rangle \ln \mathcal{N}(\pi_j^n - \pi_j^k|\beta_{jd}^2),$$

$$\langle \ln p(U|\xi)\rangle$$
$$= \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)} \big(\langle u_j^{nk}\rangle\langle \ln \xi^k\rangle + (1-\langle u_j^{nk}\rangle)\langle \ln(1-\xi^k)\rangle\big),$$

$$\langle \ln p(\xi;\alpha_{\xi 0},\beta_{\xi 0})\rangle$$
$$= \sum_{k=1}^{\Gamma}\Bigg(\frac{\Gamma(\alpha_{\xi k0}+\beta_{\xi k0})}{\Gamma(\alpha_{\xi k0})\Gamma(\beta_{\xi k0})} + (\alpha_{\xi k0}-1)\langle \ln \xi^k\rangle$$
$$+ (\beta_{\xi k0}-1)\langle \ln(1-\xi^k)\rangle\Bigg),$$

$$\langle \ln q(Z)\rangle = \sum_{n=1}^{N}\sum_{j=1}^{J}\langle z_j^n\rangle \ln\langle z_j^n\rangle,$$

$$\langle \ln q(U)\rangle = \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{k\in\gamma_d(n)}\Bigg(\langle u_j^{nk}\rangle \ln\langle u_j^{nk}\rangle$$
$$+ (1-\langle u_j^{nk}\rangle)\ln(1-\langle u_j^{nk}\rangle)\Bigg),$$

$$\langle \ln q(\xi) \rangle = \sum_{k=1}^{\Gamma} \Bigg(\frac{\Gamma(\alpha_{\xi k} + \beta_{\xi k})}{\Gamma(\alpha_{\xi k})\Gamma(\beta_{\xi k})} + (\alpha_{\xi k} - 1)\langle \ln \xi^k \rangle$$
$$+ (\beta_{\xi k} - 1)\langle \ln(1 - \xi^k) \rangle \Bigg).$$

The quantities $\langle u_j^{nk} \rangle$, $\langle \ln \xi^l \rangle$, $\langle \ln(1 - \xi^l) \rangle$, $\langle z_j^n \rangle$ and the hyperparameters $\alpha_{\xi l}$, $\beta_{\xi l}$ are given in (12) .

In order to maximize $\mathcal{L}$ over Ψ and Π, after dropping constant terms we can observe that we only need to maximize the expectation:

$$\mathcal{L}(q, \Psi, \Pi) = \mathcal{E}_{q(Z,U,\xi)}\{p(X, \Pi, Z, U, \xi; \Psi)\} + \text{const.} \quad (35)$$

where the index denotes that the expectation is computed over q. This optimization is made tractable due to the approximation (34).

Appendix B: The Student's-t Distribution

A d-dimensional random variable X follows a multivariate t-distribution, $X \sim \mathcal{S}t(\mu, \Sigma, \nu)$, with mean μ, positive definite, symmetric and real $d \times d$ covariance matrix Σ and has $\nu \in [0, \infty)$ degrees of freedom when [2], given the weight u, the variable X has the multivariate normal distribution with mean μ and covariance Σ/u:

$$X|\mu, \Sigma, u \sim \mathcal{N}(\mu, \Sigma/u), \quad (36)$$

and the weight u follows a Gamma distribution parameterized by ν:

$$u \sim \mathcal{G}(\nu/2, \nu/2). \quad (37)$$

Integrating out the weights from the joint density leads to the density function of the marginal distribution:

$$p(x; \mu, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+d}{2})|\Sigma|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{d}{2}}\Gamma(\frac{\nu}{2})[1 + \nu^{-1}\delta(x, \mu; \Sigma)]^{\frac{\nu+d}{2}}} \quad (38)$$

where $\delta(x, \mu; \Sigma) = (x - \mu)^T \Sigma^{-1} (x - \mu)$ is the Mahalanobis squared distance and Γ is the Gamma function [2]. It can be shown that for $\nu \to \infty$ the Student's t-distribution tends to a Gaussian distribution with covariance Σ. Also, if $\nu > 1$, μ is the mean of X and if $\nu > 2$, $\nu(\nu - 2)^{-1}\Sigma$ is the covariance matrix of X. Therefore, the family of t-distributions provides a heavy-tailed alternative to the normal family with mean μ and covariance matrix that is equal to a scalar multiple of Σ, if $\nu > 2$ (Fig. 16) [2]. The Student's-t has been used successfully as a robust alternative to the Gaussian distribution in maximum likelihood fitting to data that contain outliers [25, 38, 39]. In the context of the edge-preservation prior, the differences between pixels at an edge can be perceived as outliers, which in effect means that the fitting process will not take them into account when estimating the model parameters. Consequently, the fitting process will not smooth out such mixing proportion differences.

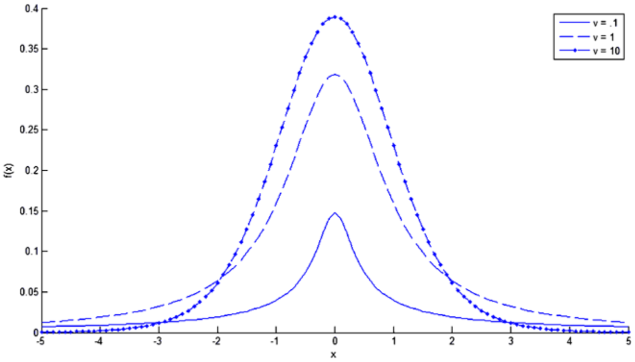

Fig. 16 The Student's t-distribution for various degrees of freedom. As $\nu \to \infty$ the distribution tends to a Gaussian. For small values of ν the distribution has heavier tails than a Gaussian

References

1. Xu, R., Wunsch II, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
3. Dempster, P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
4. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
5. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 721–741 (1984)
6. Besag, J.: On the statistical analysis of dirty pictures. J. R. Stat. Soc. **48**(3), 259–302 (1986)
7. Besag, J.: Statistical analysis of non-lattice data. Statistician **24**, 179–195 (1975)
8. Molina, R., Mateos, J., Katsaggelos, A.K., Vega, M.: Bayesian multichannel image restoration using compound Gauss-Markov random fields. IEEE Trans. Image Process. **12**, 1642–1654 (2003)
9. Kanemura, A., Maeda, S., Ishii, S.: Edge-preserving Bayesian image superresolution based on compound Markov random fields. In: Proceedings of the 17th International conference on Artificial Neural Networks, Porto, Portugal (2007)
10. Winkler, G.: Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. Springer, Berlin (2006)
11. Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognit. **36**, 131–144 (2003)
12. Kohli, P., Torr, P.H.S.: Dynamic graph cuts for efficient inference in Markov random fields. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2079–2088 (2007)

13. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1768–1783 (2006)
14. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
15. Zabih, R., Kolmogorov, V.: Spatially coherent clustering using graph cuts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, pp. 437–444 (2004)
16. Sanjay-Gopal, S., Hebert, T.: Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. IEEE Trans. Image Process. **7**(7), 1014–1028 (1998)
17. Blekas, K., Likas, A., Galatsanos, N., Lagaris, I.: A spatially constrained mixture model for image segmentation. IEEE Trans. Neural Netw. **16**(2), 494–498 (2005)
18. Nikou, C., Galatsanos, N., Likas, A.: A class-adaptive spatially variant mixture model for image segmentation. IEEE Trans. Image Process. **16**(4), 1121–1130 (2007)
19. Marroquin, J., Arce, E., Botello, S.: Hidden Markov measure field models for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(11), 1380–1387 (2003)
20. Rivera, M., Ocegueda, O., Marroquin, J.L.: Entropy-controlled quadratic Markov measure field models for efficient image segmentation. IEEE Trans. Image Process. **16**(12), 3047–3057 (2007)
21. Rivera, M., Dalmau, O., Tago, J.: Image segmentation by convex quadratic programming. In: 19th International Conference on Pattern Recognition, Tampa, FL (2008)
22. Diplaros, A., Vlassis, N., Gevers, T.: A spatially constrained generative model and an EM algorithm for image segmentation. IEEE Trans. Neural Netw. **18**(3), 798–808 (2007)
23. Blake, A., Zisserman, A.: Visual Reconstruction. MIT Press, Cambridge (1987)
24. Black, M.J., Rangarajan, A.: On the unification of line processes, outliers rejection and robust statistics in early vision. Int. J. Comput. Vis. **19**(1), 57–91 (1996)
25. Peel, D., McLachlan, G.J.: Robust mixture modeling using the t-distribution. Stat. Comput. **10**, 339–348 (2000)
26. Sfikas, G., Nikou, C., Galatsanos, N.: Edge preserving spatially varying mixtures for image segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
27. Sfikas, G., Nikou, C., Galatsanos, N., Heinrich, C.: MR brain tissue classification using an edge-preserving spatially variant Bayesian mixture model. In: Proceedings of the 11th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'08), vol. 1, pp. 43–50, New York, USA (2008)
28. McLachlan, G.: Finite Mixture Models. Wiley-Interscience, New York (2000)
29. Xia, G.S., He, C., Sun, H.: A rapid and automatic MRF-based clustering method for SAR images. IEEE Geosci. Remote Sens. Lett. **4**(4), 596–600 (2007)
30. Birkhoff, G., MacLane, S.: A Survey of Modern Algebra. McMillan, New York (1953)
31. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Berlin (1999)
32. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields. In: Proceedings of the European Conference on Computer Vision (ECCV'06), Graz, Austria (2006)
33. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging **20**(1), 45–57 (2001)
34. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 929–944 (2007)
35. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
36. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th International Conference on Computer Vision (ICCV '01), vol. 2, pp. 416–423, July 2001
37. Freixenet, J., Munoz, X., Raba, D., Marti, J., Cuff, X.: Yet another survey on image segmentation: region and boundary information integration. In: Lecture Notes in Computer Science. Proceedings of the European Conference on Computer Vision (ECCV'02), pp. 408–422 (2002)
38. Bishop, C., Svensen, M.: Robust Bayesian mixture modeling. In: Proceedings of the 12th European Symposium on Artificial Neural Networks (ESANN'04), Bruges, Belgium (2004)
39. Sfikas, G., Nikou, C., Galatsanos, N.: Robust image segmentation with mixtures of Student's t-distributions. In: Proceedings of the 14th International Conference on Image Processing (ICIP'07), San Antonio, TX, USA (2007)

Giorgos Sfikas received the B.Sc. and M.Sc. degrees in computer science from the University of Ioannina, Greece, in 2005 and 2007 respectively. He is currently pursuing his PhD studies at the University of Ioannina in cotutelle with the University of Strasbourg, France. His research interests include statistical image processing, machine learning and computer vision.

Christophoros Nikou received the Diploma degree in Electrical Engineering from the Aristotle University of Thessaloniki, Greece, in 1994 and the DEA and PhD degrees in image processing and computer vision from Louis Pasteur University, Strasbourg, France, in 1995 and 1999, respectively. During 2001, he was a Senior Researcher with the Department of Informatics, Aristotle University of Thessaloniki. From 2002 to 2004, he was with Compucon S.A., Thessaloniki. Since 2004, he is with the Department of Computer Science, University of Ioannina, Greece where he was a Lecturer (2004–2009) and he is now an Assistant Professor. His research interests mainly include statistical image processing and computer vision and their application to medical imaging.

Nikolaos Galatsanos received the Diploma of Electrical Engineering from the National Technical University of Athens-Greece in 1982. He received the MSEE and Ph. D. degrees from the Electrical and Computer Engineering Department of the University of Wisconsin-Madison in 1984 and 1989, respectively. During the period 1989–2002 he was on the faculty of the Electrical and Computer Engineering Department at the Illinois Institute of Technology. From 2002–2007, he was with the Department of Computer Science at the University of Ioannina-Greece. Currently, he is with the Department of Electrical and Computer Engineering of the University of Patras-Greece. His research interests center on Bayesian methods for image processing, medical imaging, bioinformatics, and visual communications applications. He has served as an Associate Editor for the IEEE Transactions on Image Processing and the IEEE Signal Processing Magazine and the Journal of Electronic Imaging. Dr. Galatsanos has co-edited a book titled: Image Recovery Techniques for Image and Video Compression and Transmission, Kluwer Academic, October 1998.

Christian Heinrich received the engineer degree in Electrical Engineering from Supelec, Paris, France, in 1990 and the Ph.D. degree from the University of Paris-Sud, Orsay, France, in 1997. He is now an Associate Professor at the Ecole Nationale Superieure de Physique de Strasbourg (University of Strasbourg, France). He is also with the Laboratoire des Sciences de l’Image, de l’Informatique et de la Teledetection (LSIIT, UMR CNRS-UDS 7005). His research interests are medical imaging, inverse problems and statistical image modeling.