Fusion of Hidden Markov Random Field Models and Its Bayesian Estimation

François Destrempes, Jean-François Angers, and Max Mignotte

***Abstract*—In this paper, we present a Hidden Markov Random Field (HMRF) data-fusion model. The proposed model is applied to the segmentation of natural images based on the fusion of colors and textons into Julesz ensembles. The corresponding Exploration/Selection/Estimation (ESE) procedure for the estimation of the parameters is presented. This method achieves the estimation of the parameters of the Gaussian kernels, the mixture proportions, the region labels, the number of regions, and the Markov hyper-parameter. Meanwhile, we present a new proof of the asymptotic convergence of the ESE procedure, based on original finite time bounds for the rate of convergence.**

***Index Terms*—Bayesian estimation, color and texture segmentation, Exploration/Selection algorithm, Exploration/Selection/Estimation procedure, fusion of hidden Markov random field models, Julesz ensembles, Markov Chain Monte Carlo (MCMC) algorithm.**

I. Introduction

DATA fusion of image channels provided by various sensors is an important problem in image processing, with applications to image segmentation of natural images, or in areas such as geophysical imaging, medical imaging, and radio-astronomy (see [1]). In this paper, we focus on image segmentation based on data fusion.

One of the goals of image segmentation is to decompose an image into meaningful regions, such as consistent parts of objects or of the background, based on the fusion of various types of features. Our point of view is to characterize each meaningful region in the image by the distribution of the features on the region. This characterization of a region is called a Julesz ensemble [2].

When working with Julesz ensembles, it is customary to combine the various features assuming the independence property conditional to the region process. An example can be found in [3], in the case of univariate Gaussian kernels and uniform priors on the estimated Gaussian parameters. In [4], a Gaussian and an inverse gamma priors are set on the mean and the variance, respectively, of each Gaussian kernel. One can also assume a correlation between the channels conditional to the region process. For instance, the channels can be combined into a single multi-channel vector [5], and the joint likelihood is then defined directly on that vector. In [6], a correlation is also set on the channels. In this paper, we set a correlation between the features of a same type, but consider features of different types as independent conditional to the region process.

In order to perform data fusion, a stationary distribution can also be put directly on the image features knowing the marginals of the features, based on the maximum entropy principle [1], [7], [8]. The solution belongs to the generalized exponential family (cf. the filters, random fields, and maximum entropy (FRAME) model [9]). A fundamental result states the equivalence [10] between FRAME models and Julesz ensembles. In this paper, we consider directly Julesz ensembles as a mean of data fusion, due to the computational load of estimating FRAME models. In [11], possible different generic (FRAME) models are set on the various regions. But since the likelihoods of distinct models might not be comparable, appropriate weights need to be assigned on each type of generic model [11]. In contrast, we adopt the same model for all the regions, that is flexible enough to represent *any* type of region.

In the models mentioned above, the fusion process is based on the data itself. One can also base the fusion decision on the individual channel decisions. In [12], the fusion of the decisions is based on an *ad hoc* Markov model. In [13] and [14], the various channel decisions are combined together according to Dempster–Shafer theory.

As it stands, the image itself could form a single region, or on the contrary, each region could be formed of only a few pixels. Thus, in order to obtain a meaningful decomposition of the image, one needs a prior probability on the region process that sets a constraint on the spatial organization of the region labels. In this paper, the spatial prior is defined by a Markov model of order 2, as well as a global constraint based on the size of connected regions.

Once a model is established, an equally important problem is the estimation of the model parameters. We adopt the Bayesian paradigm for the estimation of the model. We propose to estimate not only the Gaussian parameters of the kernels and the region labels, but also the mixture proportions, the number of regions, and the Markov hyper-parameter. Various Bayesian priors are set on the parameters. We choose to compute the MAP of the proposed model, weighted by a global constraint on the region process.

The algorithm presented in this paper in order to compute the MAP is an extension of the Exploration/Selection/Estimation (ESE) procedure [15] to the proposed fusion model. It is a

Manuscript received August 3, 2005; revised February 10, 2006. This work was supported by the former Fonds Formation Chercheurs and Aide Recherche (FCAR), QC, Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joachim M. Buhmann.

F. Destrempes and M. Mignotte are with the Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, Pavillon André-Aisenstadt, Montréal, QC H3C 3J7 Canada (e-mail: destremp@iro.umontreal.ca; mignotte@iro.umontreal.ca).

J.-F. Angers is with the Département de Mathématiques et de Statistique, Université de Montréal, Montréal, QC H3C 3J7 Canada (e-mail: angers@dms.umontreal.ca).

Color versions of Figs. 1, 2, 3, 6, 7, and 8 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2006.877522

variant of the Exploration/Selection (ES) algorithm [16], that integrates an (approximate) Markov Chain Monte Carlo (MCMC) transition kernel into the exploration scheme. Meanwhile, we present a new proof of the asymptotic convergence of the ES algorithm, based on original finite time bounds for the rate of convergence. The ES algorithm can be viewed as a mix between genetic algorithms and simulated annealing. See [17] and [18] for closely related algorithms.

Among the main algorithms for simulating the posterior distribution of models of variable dimension, are the Reversible Jump Markov Chain Monte Carlo (RJMCMC) [19]–[21], the Data Driven MCMC (DDMCMC) [11], the Birth-and-Death MCMC (BDMCMC) [22], the Delayed Rejection MCMC (DRMCMC) [23], the general Continuous Time MCMC (CTMCMC) [24], and a generalization [25] of the Swendsen–Wang algorithm [26]. In our case, we avoid the (major) difficulty of engineering a Metropolis–Hastings dynamics with a sufficiently high rate of acceptance, upon using the ES algorithm. The point is that in order to compute the MAP of the model, it is not required to simulate precisely the posterior distribution of its parameters.

In this paper, we illustrate the proposed data fusion model and the estimation method with color and texture features. The color features are the *Luv* components, and the texture features are the responses to Gabor and Gaussian filters. As mentioned in [27], textons refer to micro-structures in natural images. In [28], a texton is modeled by a Voronoi cell in the space of texture features. In this paper, we model a texton by a unimodal (Gaussian) distribution on the space of texture features. The distribution of the texture features on the region class is then a mixture of the unimodal kernels, each one appearing according to a certain proportion within the region. That distribution describes a Julesz texture.

This work develops on our previous paper [15] in the following aspects: 1) the ESE procedure is applied to a triplet of Markov random fields, rather than a pair of random fields, in the context of data fusion; 2) the local likelihoods are mixtures of distributions rather than unimodal distributions; 3) the hyper-parameter of the Markov prior on the region process is estimated rather than being fixed; 4) the proposed model is shown to be identifiable; and 5) finite time bounds for the proposed algorithm are given rather than just a proof of asymptotic convergence.

The remaining part of this paper is organized as follows. In Section II, we present the hidden Markov random field (HMRF) color and texture models considered in this paper, as well as their fusion model. In Section III, we present the Bayesian estimator and its algorithmic computation. Experimental results are briefly presented in Section IV.

II. Fusion of Colors and Textures

A. Random Fields Considered

The lattice of the image pixels is viewed as a graph G, with set of nodes V. We consider a hidden discrete random field $X = (X_s)$ on G with random variable X_s taking its values in a finite set of labels $\Lambda = \{e_1, e_2, \ldots, e_K\}$. Our intention is to consider Λ as the set of region classes, and we call X the region process.

We consider M levels of analysis of the image, such as colors and textures. At each level of analysis $n = 1, \ldots, M$, an observable random field Y_n is defined on the graph G. Accordingly, for each level of analysis n, an observable (continuous) random variable $Y_{s,n}$ is defined at each site $s \in V$. The variables $Y_{s,n}$ take their values in a space of image features Υ_n of dimension d_n, depending only on the level n. Our intention is to consider each set Υ_n as a space of image features, such as color or texture features. We collect the various levels of analysis together, upon considering the random field $\hat{Y} = Y_1 \times \cdots \times Y_M$ on the graph G.

Next, for each level of analysis $n = 1, \ldots, M$, we consider discrete random variables $C_{s,n}$ that take their values in a finite set of K_n cue labels $\Omega_n = \{g_{1,n}, \ldots, g_{K_n,n}\}$. Each label represents an equivalence class of similar image features. The discrete random field $C_n = (C_{s,n})$ is called a cue process. Examples of cue processes are the color process and the texton process (Section II-D). We collect the various cue processes together, upon considering the random field $\hat{C} = C_1 \times \cdots \times C_M$ on the graph G. See Fig. 1 for an illustration of the region process X, and the cue processes C_1 and C_2 in the case of color and texture features.

B. Likelihood

We now present a model for the likelihood of the observable image features $\hat{y} = (y_1, \ldots, y_M)$ conditional to the hidden field of region labels x, and the hidden field of cue labels $\hat{c} = (c_1, \ldots, c_M)$ (such as color labels and texton labels). The main point is to use a unimodal distribution for the local likelihood of the observable image features $y_{s,n}$ conditional to a cue label, and to use a mixture of these unimodal distributions for the local likelihood of the observable image features conditional to a region label.

For each site s at level n and each cue class $g_{i,n} \in \Omega_n$, the likelihood of the observable features $y_{s,n}$ conditional to the cue label $c_{s,n}$ is modeled essentially by a Gaussian kernel. More precisely, we consider the diffeomorphism $h : (0,1)^{d_n} \to \mathbb{R}^{d_n}$ defined by $\tanh^{-1}(2x - 1)$ on each component $x \in (0,1)$, where d_n is the dimension of the feature space Υ_n. We define $P(y_{s,n} \mid c_{s,n} = g_{i,n}, \mu_{i,n}, \Sigma_{i,n})$ by

$$\mathcal{N}(h(y_{s,n}); \mu_{i,n}, \Sigma_{i,n}) J(h)(y_{s,n}) \tag{1}$$

where $J(h)$ denotes the Jacobian of the map h.

We collect the local likelihoods together by setting

$$P(y_n \mid c_n, \mu_n, \Sigma_n) = \prod_s P(y_{s,n} \mid c_{s,n}, \mu_n, \Sigma_n) \tag{2}$$

where $\mu_n = (\mu_{i,n})$ and $\Sigma_n = (\Sigma_{i,n})$. We view the M levels as independent, conditional to $\hat{C}$; more precisely, the joint distribution of $\hat{Y}$ conditional to $\hat{C}$ is modeled by

$$P(\hat{y} \mid \hat{c}, \mu, \Sigma) = \prod_{n=1}^{M} P(y_n \mid c_n, \mu_n, \Sigma_n) \tag{3}$$

where $\mu = (\mu_n)$ and $\Sigma = (\Sigma_n)$.

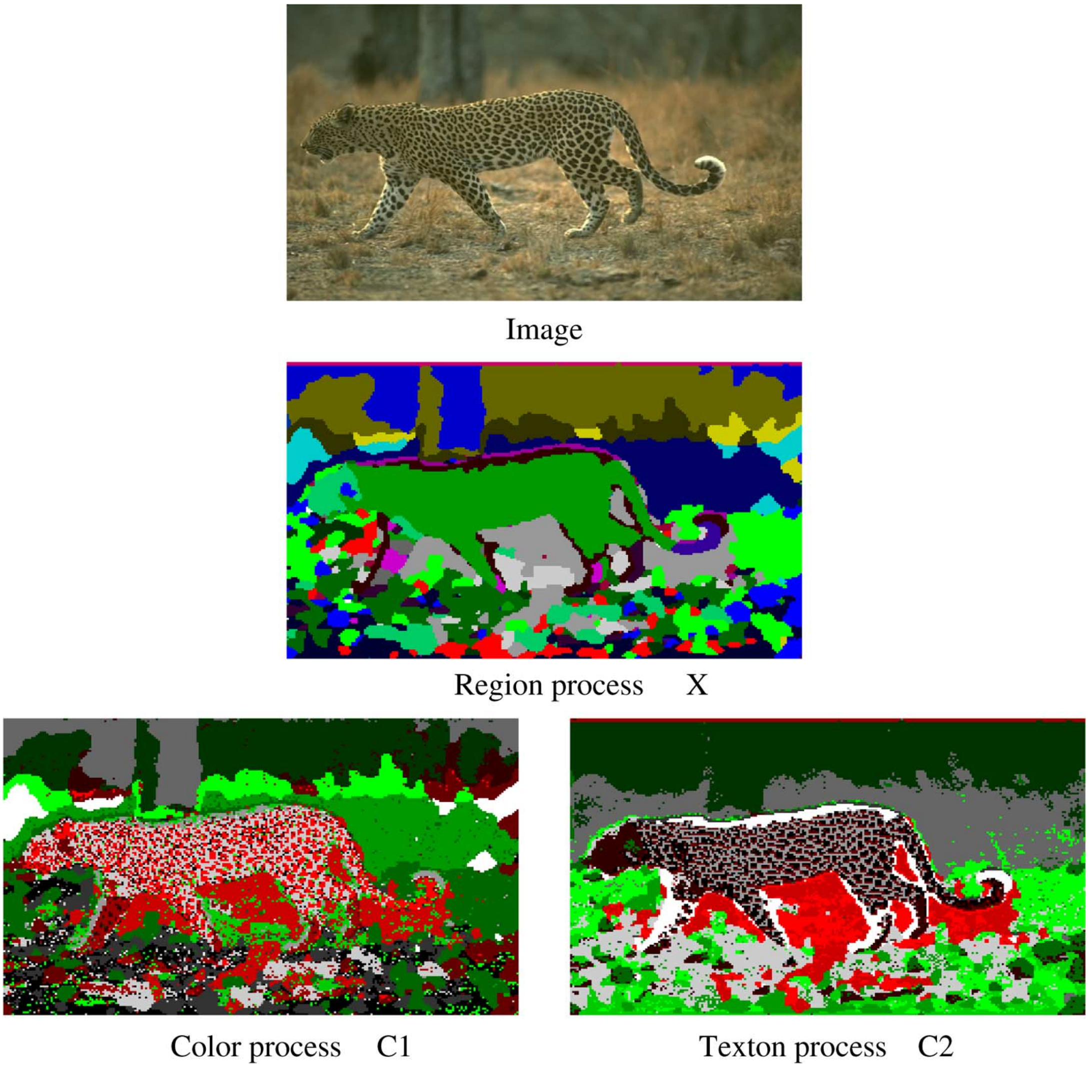

Fig. 1. Natural image, the estimated region process X, the simulated cue processes: the color process C^1 and the texton process C^2 (cf. Section II-A). See Section II-D for a description of the color features Y^1 and the texture features Y^2.

Next, at each level of analysis $n = 1, \ldots, M$, we consider each cue label $g_{i,n}$ to appear in some proportion within the region class e_k. Namely, let $\pi_n = (\pi_{(i,k),n})$ satisfy

$$\pi_{(i,k),n} \geq 0, \quad \sum_{i=1}^{K_n} \pi_{(i,k),n} = 1, \ \ 1 \leq i \leq K_n, \ \ 1 \leq k \leq K. \tag{4}$$

Then, we model the probability of $c_{s,n}$ conditional to x_s by

$$P(c_{s,n} = g_{i,n} \mid x_s = e_k, \pi_n) = \pi_{(i,k),n}. \tag{5}$$

We collect these local likelihoods together by setting

$$P(c_n \mid x, \pi_n) = \prod_s P(c_{s,n} \mid x_s, \pi_n). \tag{6}$$

Again, we consider the M level of analysis to be independent. So, we set

$$P(\hat{c} \mid x, \pi) = \prod_{n=1}^{M} P(c_n \mid x, \pi_n) \tag{7}$$

where $\pi = (\pi_n)$.

The joint distribution of $(\hat{y}, \hat{c})$ conditional to the region process x is expressed as

$$\begin{aligned} &P(\hat{y}, \hat{c} \mid x, \mu, \Sigma, \pi) \\ &\quad = P(\hat{y} \mid \hat{c}, \mu, \Sigma) P(\hat{c} \mid x, \pi) \\ &\quad = \prod_{n=1}^{M} P(y_n \mid c_n, \mu_n, \Sigma_n) P(c_n \mid x, \pi_n) \\ &\quad = \prod_{n=1}^{M} \prod_s P(y_{s,n} \mid c_{s,n}, \mu_n, \Sigma_n) P(c_{s,n} \mid x_s, \pi_n) \end{aligned} \tag{8}$$

using (2), (3), (6), and (7). We deduce that the marginal of $\hat{y}$ conditional to x is equal to

$$\begin{aligned} P(\hat{y} \mid x, \mu, \Sigma, \pi) &= \sum_{\hat{c}} P(\hat{y}, \hat{c} \mid x, \mu, \Sigma, \pi) \\ &= \sum_{\hat{c}} \prod_{n=1}^{M} \prod_{s} P(y_{s,n} \mid c_{s,n}, \mu_n, \Sigma_n) P(c_{s,n} \mid x_s, \pi_n) \\ &= \prod_{n=1}^{M} \prod_{s} \left\{ \sum_{c_{s,n}} P(y_{s,n} \mid c_{s,n}, \mu_n, \Sigma_n) P(c_{s,n} \mid x_s, \pi_n) \right\} \\ &= \prod_{n=1}^{M} \prod_{s} P(y_{s,n} \mid x_s, \mu_n, \Sigma_n, \pi_n) \qquad (9) \end{aligned}$$

where each factor $P(y_{s,n} \mid x_s = e_k, \mu_n, \Sigma_n, \pi_n)$ is equal to

$$\sum_{i=1}^{K_n} \pi_{(i,k),n} \mathcal{N}(h(y_{s,n}) \mid \mu_{i,n}, \Sigma_{i,n}) J(h(y_{s,n})). \qquad (10)$$

Thus, for each region label x_s, the likelihood $P(y_{s,n} \mid x_s, \mu_n, \Sigma_n, \pi_n)$ is a mixture of the K_n distributions $P(y_{s,n} \mid c_{s,n} = g_{i,n}, \mu_{i,n}, \Sigma_{i,n})$, and only the mixture proportions $\pi_{(i,k),n}$ vary from one region class to another. In particular, the Gaussian kernels $P(y_{s,n} \mid c_{s,n} = g_{i,n}, \mu_{i,n}, \Sigma_{i,n})$ are independent of the region label e_k. The proposed family of distributions is quite flexible, since any continuous distribution can be approximated by a mixture of a sufficiently large number of Gaussian kernels. See Fig. 2.

The marginal distributions of the features $P(y_{s,n} \mid x_s = e_k, \mu_n, \Sigma_n, \pi_n)$, $n = 1, \ldots, M$, define uniquely a Julesz ensemble [2] for each region e_k; namely, the set of stationary fields with the distributions $P(y_{s,n} \mid x_s = e_k, \mu_n, \Sigma_n, \pi_n)$, $n = 1, \ldots, M$, as marginals. In the case of texture features, the Julesz ensemble is referred to as a Julesz texture. Furthermore, we then call the micro-texture corresponding to a single Gaussian kernel $P(y_{s,n} \mid c_{s,n} = g_{i,n}, \mu_{i,n}, \Sigma_{i,n})$ a texton. Thus, a texture is a mixture of textons.

C. Prior on the Region Process

In this paper, we consider a Potts model of order 2 on G. Namely, we consider the family of potential functions

$$\begin{aligned} U_{\langle s,t \rangle}(x) &= \sqrt{2}\delta(x_s \neq x_t) \\ &\qquad \text{if } \langle s,t \rangle \text{ is horizontal or vertical} \\ U_{\langle s,t \rangle}(x) &= \delta(x_s \neq x_t) \\ &\qquad \text{if } \langle s,t \rangle \text{ is diagonal} \\ U(x) &= \sum_{\langle s,t \rangle} U_{\langle s,t \rangle}(x) \qquad (11) \end{aligned}$$

where $\langle s,t \rangle$ ranges over the set of all binary cliques. The diagonal cliques have less weight than horizontal or vertical cliques as in [29].

Now, as in [15], K is considered as the maximal number of regions labels allowed in the region process X. In order to handle the case of possibly less classes than K, we consider a vector v of K bits, with the constraint that $|v| = \sum_{k=1}^{K} v_k \geq 1$ (cf. [15]). The vector v indicates which regions labels are allocated (i.e., e_k is allocated if $v_k = 1$).

Let $\beta > 0$ be a hyper-parameter. The prior distribution $P(x \mid \beta, v)$ is then defined by

$$\frac{1}{Z(\beta, v)} \chi(x, v) e^{-\beta U(x)} \qquad (12)$$

where $\chi(x, v) = 1$ if the labels appearing in x are precisely the ones allocated by the vector v (i.e., $v_k = 1$ if and only if $x_s = e_k$ for some pixel s), and $\chi(x, v) = 0$, otherwise. Here, $Z(\beta, v)$ is a normalizing constant called the *partition function*

$$Z(\beta, v) = \sum_{x : \chi(x,v) = 1} e^{-\beta U(x)}. \qquad (13)$$

An important uniqueness property of this model will be discussed in Section III-A11.

D. Image Features

We now present the color features at level $n = 1$. At each pixel of the image, the raw color RGB components yields the CIE XYZ components under the hypothesis that the NTSC phosphors standard [30] was used. The features $y_{s,1}$ are then the Luv color components [31] at the pixel s, computed from its XYZ components. In particular, $d_1 = 3$ is the dimension of the space of color features Υ_1. The purpose of the Luv components is to provide a perceptually uniform color space. Note, however, that we could have used any color system since they all differ by a (possibly non-linear) change of variables.

Next, we present the texture features at level $n = 2$. A fundamental result [9] states the perfect reconstruction of the luminance density (a stationary field) from the marginals of *all linear* filter responses (one-dimensional random variables). In practice, only a few filters suffice to distinguish textures within a given image. In that case, the joint distribution of the chosen filter responses defines uniquely a Julesz ensemble [2] that describes the texture.

Let $\mathcal{M}$ be a given filter bank. We consider the observable random vector $y_{s,2}$ defined by the filter responses $(f * Y)(s)$, where Y_s is the image luminance Y of the XYZ components at pixel s (and *not* the L component of the Luv coordinates). In particular, $d_2 = |\mathcal{M}|$. An important issue is the design of a filter bank [8], [32]. In this paper, we choose a filter bank as follows.

Recall that the linear 2-D-Gabor filters [33] *are optimal* for a joint spatial and spectral resolution (cf. the *uncertainty principle* of [33]). Such a filter $f(x, y; \phi, \mu_x, \mu_y, \sigma_x, \sigma_y)$ is defined by

$$\exp\left\{-2\pi^2\left((\sigma_x x')^2 + (\sigma_y y')^2\right) - 2\pi j(\mu_x x + \mu_y y)\right\} \qquad (14)$$

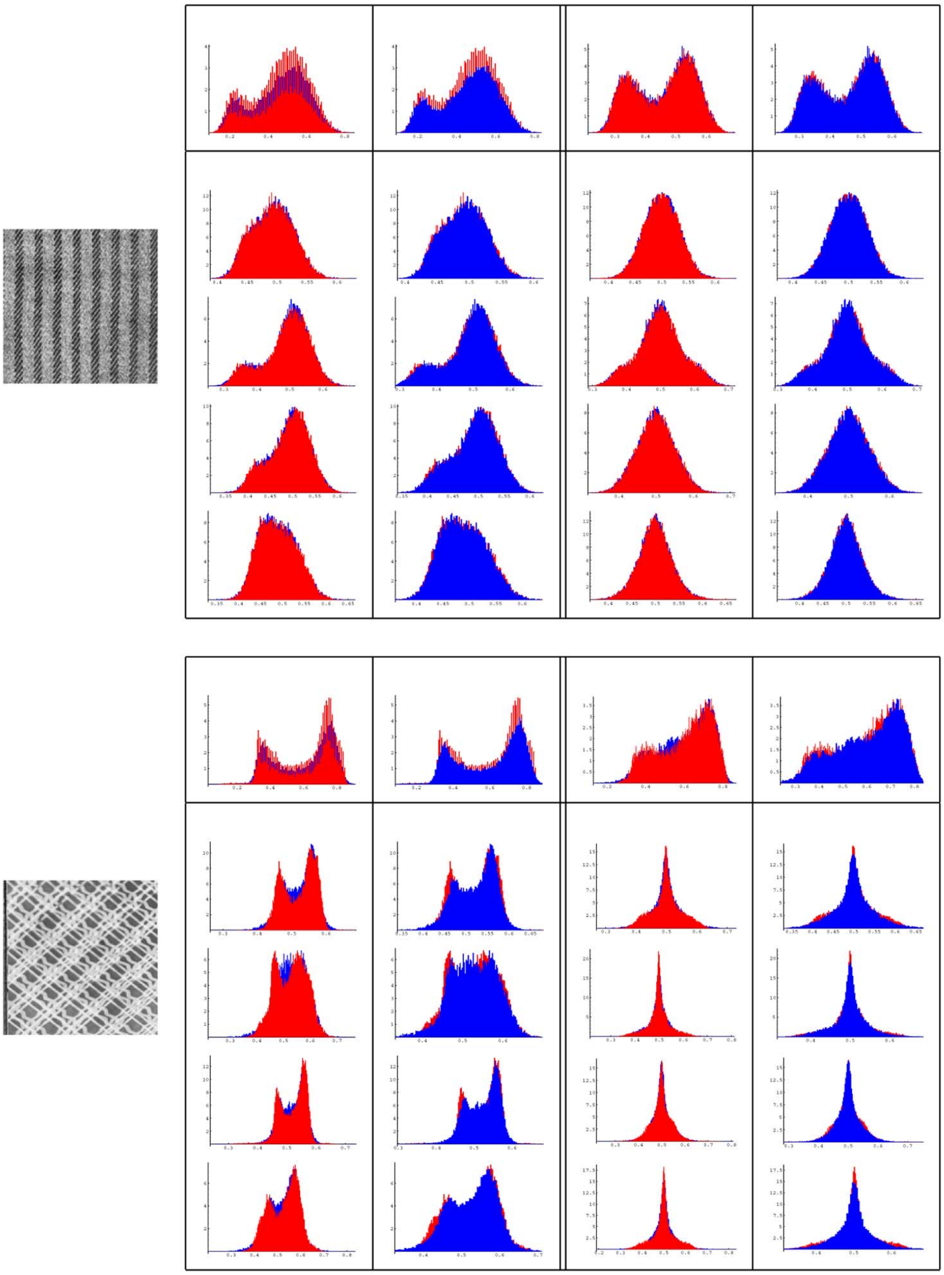

Fig. 2. Examples of empirical and simulated distributions for the gray-level and the Gabor filter responses, based on the parameters estimated by the ESE procedure. Mixtures of Gaussian kernels are quite flexible in modeling various continuous distributions.

where $x' = \cos(\phi)x + \sin(\phi)y$ and $y' = -\sin(\phi)x + \cos(\phi)y$, and $j = \sqrt{-1}$. The first term defines a Gaussian kernel with mean (μ_x, μ_y) (of dimension 2) and covariance matrix $\Sigma_{\phi,\sigma_x,\sigma_y}$

$$\begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}^t . \tag{15}$$

The second term produces a harmonic modulation with mean frequency (μ_x, μ_y). The angle ϕ is called the orientation, and σ_x, σ_y the standard-deviations. Accordingly, its Fourier transform F is a Gaussian kernel of mean (μ_x, μ_y) and covariance matrix $\Sigma_{\phi,\sigma_x,\sigma_y}$.

Note that for a Gabor filter $f = f(\phi, \mu_x, \mu_y, \sigma_x, \sigma_y)$, the spatial and the spectral uncertainties [33] are, respectively, equal to

$$\Delta(f) = \frac{1}{8\pi^2}\sigma_x^{-1}\sigma_y^{-1}, \quad \Delta(F) = \frac{1}{2}\sigma_x\sigma_y. \tag{16}$$

We establish the architecture of the filter bank as follows. We set in what follows $\sigma_x = \sigma_y = \sigma$, and $\mu_y = 0$. We choose a *bandwidth* of two octaves; i.e., $\left(\left(\mu_x + \sqrt{2\log(2)}\sigma\right) / \left(\mu_x - \sqrt{2\log(2)}\sigma\right)\right) = 4$. Thus, the mean frequency is equal to $\mu_x = (5/3)\sqrt{2\log(2)}\sigma$. We take four equally-spaced rotations $\phi = 0$, $(\pi/4)$, $(\pi/2)$, $(3\pi/4)$. Taking the real and imaginary parts of each Gabor

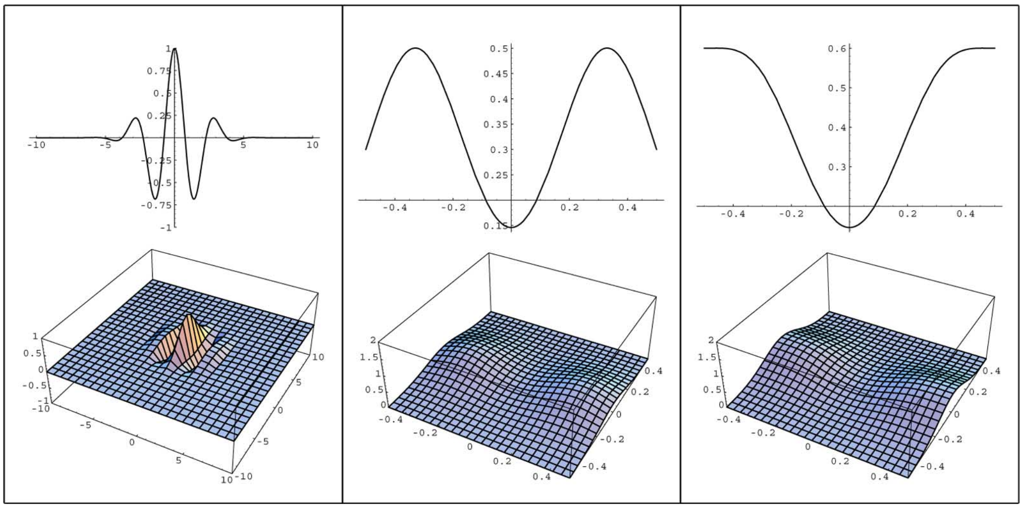

Fig. 3. Left: real part of a Gabor filter. Center: its Fourier transform. Right: spectral aliasing due to spatial digitization. Parameters: $\mu_x = 0.33, \mu_y = 0, \phi = 0, \sigma_x = \sigma_y = 0.168166$

filter, we obtain 8 high-pass filters (see Fig. 3). We also consider a low-pass Gaussian filter with *same spectral resolution* as the Gabor filters, i.e., $\left(1/\sqrt{2}\right)\sigma$, and hence with same variance. The corresponding spatial resolution of those nine filters is $\left(1/\left(\sqrt{8}\pi\sigma\right)\right)$. In our tests, we chose $\mu_x = 0.33$, which yields $\sigma \approx 0.168166$, $3\sqrt{\Delta(f)} \approx 2.00765$ and $3\sqrt{\Delta(F)} \approx 0.356733$.

III. Estimation of the Model

A. Bayesian Estimator Considered

The fusion model $(X, \hat{Y})$ presented in Section II is completely described by the vectors of parameters

$$\begin{aligned}\mu &= (\mu_{i,n}), \quad \Sigma = (\Sigma_{i,n})\\ \pi &= (\pi_{(i,k),n}); \quad \beta, v\end{aligned} \tag{17}$$

where $1 \leq k \leq K$, $1 \leq i \leq K_n$, and $1 \leq n \leq M$, as presented in (1), (4), and (12). We estimate the vector of parameters $\theta = (x, \mu, \Sigma, \pi, \beta, v)$ in a Bayesian framework. Under that paradigm, it is essential to specify carefully the prior on the parameters to be estimated.

1) Prior on the Mixture Proportions: It is now a standard practice in Bayesian statistics to set a Dirichlet process prior [34] on the mixture proportions. Namely, the prior on the mixture proportions π (conditional to the allocation vector v) is defined independently at each level $n = 1, \ldots, M$ and for each allowed region class $k = 1, \ldots, K$ (i.e., such that $v_k = 1$) by a Dirichlet distribution $\mathcal{D}(\pi_{(1,k),n}, \ldots, \pi_{(K_n,k),n}; A_0, \alpha_1, \ldots, \alpha_{K_n})$ equal to

$$\begin{aligned}&\frac{\Gamma(A_0)}{\prod_i \Gamma(A_0\alpha_i)} \prod_{i=1}^{K_n} (\pi_{(i,k),n})^{A_0\alpha_i - 1}\\ &\sum_{i=1}^{K_n} \alpha_i = 1, \quad \alpha_i > 0\end{aligned} \tag{18}$$

where Γ is the Euler gamma function. The constant $A_0 > 0$ is called the dispersion parameter, and the constants $\alpha_1, \ldots, \alpha_{K_n}$ represent the prior information of the latent variable $C_{s,n}$.

In our case, the initial guess on the mixture proportions π_n is that the cue label $c_{s,n}$ conditional to each region label $x_s = e_k$ is distributed uniformly on the set Ω_n of cue labels. So, we set the prior proportions $\alpha_1, \ldots, \alpha_{K_n}$ equal to $(1/K_n)$. This is called a Dirichlet process prior with base measure the uniform distribution on Ω_n, and with dispersion parameter $A_0 > 0$. In our setting, we want a non-informative uniform distribution for the mixture proportions, so that we take $A_0 = K_n$. Thus, we obtain the uniform prior $P(\pi \mid v)$

$$\prod_{k:v_k=1} \prod_{n=1}^{M} \mathcal{D}\left(\pi_{(1,k),n}, \ldots, \pi_{(K_n,k),n}; K_n, \frac{1}{K_n}, \ldots, \frac{1}{K_n}\right). \tag{19}$$

In order to sample $(\pi_{(1,k),n}, \ldots, \pi_{(K_n,k),n})$ from a Dirichlet distribution $\mathcal{D}(A_0, \alpha_1, \ldots, \alpha_{K_n})$, we use the algorithm of Table I. An interesting advantage of the Dirichlet prior on the mixture parameters $(\pi_{(1,k),n}, \ldots, \pi_{(K_n,k),n})$ is that the posterior distribution of the parameters conditional to the cue process c_n and the region process x is also a Dirichlet distribution as detailed in Table II. A prior with that property is called a conjugate prior.

2) Prior on the Cue Likelihood Parameters: For the cue Gaussian likelihood parameters $(\mu_{i,n})$ and $(\Sigma_{i,n})$ of (1), we consider independently for each level of analysis $n = 1, \ldots, M$ and for each cue class $i = 1, \ldots, K_n$, the usual conjugate prior for multivariate Gaussian distributions defined by [35, Th. 7.7.3]

$$\begin{aligned}\mu_{i,n} \mid \Sigma_{i,n} &\sim \mathcal{N}\left(\mu_0, \frac{1}{k_0}\Sigma_{i,n}\right)\\ \Sigma_{i,n} &\sim \mathcal{IW}(\Lambda_0, \nu_0)\end{aligned} \tag{20}$$

where $\mathcal{N}$ is the Normal distribution and $\mathcal{IW}$ is the *inverted Wishart* distribution with ν_0 degrees of freedom. Here, μ_0 is a vector of dimension d_n, the dimension of the feature space Υ_n, k_0 is a positive constant, and Λ_0 is a positive-definite symmetric

TABLE I
SIMULATION OF A DIRICHLET DISTRIBUTION

```
Simulation of $(\pi_{(1,k),n}, ..., \pi_{(K_n,k),n})$ according to a Dirichlet distribution $\mathcal{D}(A_0, \alpha_1, ..., \alpha_{K_n})$:
  for $i = 1, ..., K_n$ do
    Sample $Z_i$ from a Gamma distribution $\mathcal{G}(A_0\alpha_i, 1)$.
  end for
  for $i = 1, ..., K_n$ do
    Set $\pi_{(i,k),n} = \frac{Z_i}{\sum_{i=1}^{K_n} Z_i}$.
  end for
```

TABLE II
EXPRESSION OF THE POSTERIOR DIRICHLET DISTRIBUTION

```
Let $n$ be a fixed level of analysis. Let $c_n$ be a realization of the cue process at level $n$, and $x$ be a realization of the region process. Let the prior on $(\pi_{(1,k),n}, ..., \pi_{(K_n,k),n})$ be as in equation (18).
  Compute $N_{(i,k)} = |\{s : c_{s,n} = g_{i,n}, x_s = e_k\}|$ and $N_k = |\{s : x_s = e_k\}|$, for $i = 1, ..., K_n$, and $k = 1, ..., K$.
Then, the posterior distribution of $(\pi_{(1,k),n}, ..., \pi_{(K_n,k),n})$ given $c_n$ and $x$ is the Dirichlet distribution $\mathcal{D}(N_k + A_0, \frac{N_{(i,k)}+A_0\alpha_1}{N_k+A_0}, ..., \frac{N_{(K_n,k)}+A_0\alpha_{K_n}}{N_k+A_0})$.
```

matrix of dimension $d_n \times d_n$. The inverted Wishart distribution is defined by (21), shown at the bottom of the page, where $|A|$ denotes the determinant of a square matrix $A = (a_{jk})$, $\mathrm{tr}(A) = \sum_{j=1}^{d_n} a_{jj}$ denotes its trace, and

$$\Gamma_{d_n}\left(\frac{1}{2}\nu_0\right) = \pi^{d_n(d_n-1)/4} \prod_{j=1}^{d_n} \Gamma\left(\frac{1}{2}\nu_0 - \frac{1}{2}(j-1)\right) \tag{22}$$

with Γ the Euler gamma function. The expectation of $\Sigma_{i,n}$ is $(1/(\nu_0 - (d_n+1)))\Lambda_0$, for $\nu_0 > d_n + 1$. See [35, Lemma 7.7.1].

In our tests, we fix $k_0 = 0.01$ and $\nu_0 = d_n + 2$, and the values of Λ_0 and μ_0 are estimated, once and for all on a given image, at each level of analysis according to the method presented in Table III. Thus, we obtain the prior defined by $P(\mu, \Sigma)$ equal to

$$\prod_{n=1}^{M}\prod_{i=1}^{K_n} \mathcal{N}\left(\mu_{i,n}; \bar{\mu}_n, \frac{1}{k_0}\Sigma_{i,n}\right) \times \mathcal{IW}(\Sigma_{i,n}; (d_n+2)S_n, d_n+2) \tag{23}$$

where $\bar{\mu}_n$ and S_n are as in Table III.

TABLE III
COMPUTATION OF THE PRIOR GAUSSIAN/INVERTED WISHART PARAMETERS

```
Let $n$ be a fixed level of analysis.
  Compute the empirical mean $\bar{\mu}_n$ of the set $\mathcal{E}_n = \{h(y_{s,n}) : s \in V\}$. Compute its ML covariance matrix $S_n$.
  Set $k_0 = 0.01$, $\nu_0 = d_n + 2$, $\mu_0 = \bar{\mu}_n$ and $\Lambda_0 = \nu_0 S_n$ in equation (20).
```

TABLE IV
SIMULATION OF AN INVERTED WISHART DISTRIBUTION

```
Simulation of $\Sigma_{i,n}$ according to an inverted Wishart distribution $\mathcal{IW}(\tilde{\Lambda}, \tilde{\nu})$, with $\nu > d_n + 1$:
  for $j = 1, ...d_n$ do
    sample $R_{jj}$ from a chi distribution with $\tilde{\nu} - j + 1$ degrees of freedom.
  end for
  for $1 \leq j < k \leq d_n$ do
    sample $R_{jk}$ from a Gaussian distribution $\mathcal{N}(0, 1)$.
  end for
  for $1 \leq k < j \leq d_n$ do
    set $R_{jk} = 0$.
  end for
  Decompose $\tilde{\Lambda} + I$ in the form $WDW^t$, where $W$ is an orthogonal matrix and $D$ is a diagonal matrix.
  Set $A = R^{-1}(D - I)^{1/2}W^t$.
  Set $\Sigma_{i,n} = A^t A$.
```

In order to simulate $\Sigma_{i,n}$ according to an inverted Wishart distribution $\mathcal{IW}(\tilde{\Lambda}, \tilde{\nu})$, we use the algorithm of Table IV, which is a variant of Jones' algorithm [36]. In order to simulate $\mu_{i,n}$ according to a Gaussian distribution $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$, we use the algorithm of Table V. Again, an interesting advantage of the Gaussian/Inverted Wishart prior on the likelihood parameters $\mu_{i,n}$ and $\Sigma_{i,n}$ is that the posterior distribution of the parameters conditional to the cue process c_n is also a Gaussian/Inverted Wishart distribution as described in Table VI. See [35, Th. 7.7.3]. Furthermore, Table VI shows that the matrix $\tilde{\Lambda}_{i,n}$ will be positive-definite even if the empirical matrix $S_{i,n}$ is singular, as long as the prior matrix Λ_0 is non-singular. This property is useful when analyzing an image in which some regions have a constant feature vector (because then, $S_{i,n} = 0$).

3) Prior on the Region Process: The prior on the region process x is given by the model $P(x \mid \beta, v)$ of (12).

4) Prior on the Spatial Hyper-Parameter: We adopt the following non-informative improper prior distribution on the

$$\mathcal{IW}(\Sigma_{i,n}; \Lambda_0, \nu_0) = \frac{|\Lambda_0|^{(1/2)\nu_0} |\Sigma_{i,n}|^{-(1/2)(\nu_0+d_n+1)} e^{-(1/2)\mathrm{tr}\Lambda_0(\Sigma_{i,n})^{-1}}}{2^{(1/2)\nu_0 d_n}\Gamma_{d_n}\left(\frac{1}{2}\nu_0\right)} \tag{21}$$

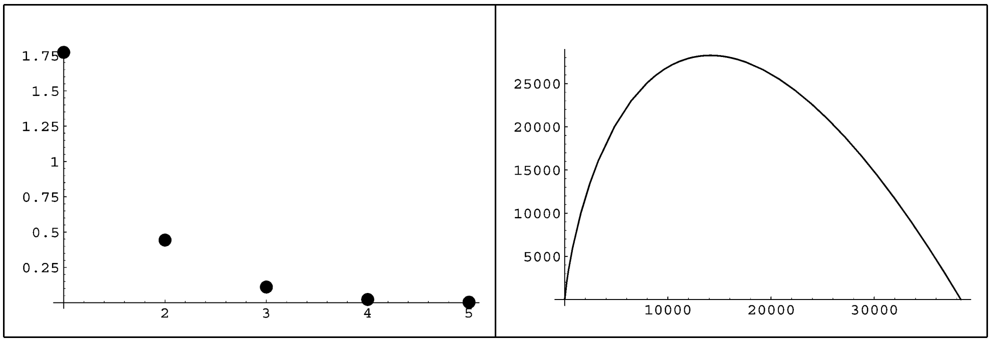

Fig. 4. Left: Prior distribution on the number $|v|$ of allocated region labels (see Section III-A5). Right: Global constraint on the number of connected regions in the case of 38 400 pixels and equal size components (see Section III-A9).

TABLE V
SIMULATION OF A GAUSSIAN DISTRIBUTION

```
Simulation of $\mu_{i,n}$ according to a Gaussian distribution $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$:
  **for** $j = 1, ...d_n$ **do**
    sample $Z_j$ from a Gaussian distribution $\mathcal{N}(0, 1)$.
  **end for**
  Set $Z = (Z_j)$.
  Decompose $\tilde{\Sigma} + I$ in the form $WDW^t$, where $W$ is an orthogonal matrix and $D$ is a diagonal matrix.
  Set $\mu_{i,n} = \tilde{\mu} + W(D - I)^{1/2} Z$.
```

TABLE VI
EXPRESSION OF THE POSTERIOR GAUSSIAN/INVERTED WISHART DISTRIBUTION

Let n be a fixed level of analysis. Let c_n be a realization of the cue process at level n. Let $1 \leq i \leq K_n$ be fixed. Let $\mu_{i,n}, \Sigma_{i,n}$ have the prior of equation (20).

Compute $N_{i,n} = |\{s \,:\, c_{s,n} = g_{i,n}\}|$.

Compute the empirical mean $\bar{\mu}_{i,n}$ of the set $\mathcal{E}_{i,n} = \{h(y_{s,n}) \,:\, c_{s,n} = g_{i,n}\}$.

Compute its ML covariance matrix $S_{i,n}$.

Set

$$\begin{aligned}
\tilde{\mu}_{i,n} &= \bar{\mu}_{i,n} - \frac{k_0}{N_{i,n} + k_0}(\bar{\mu}_{i,n} - \mu_0), \\
\tilde{\Lambda}_{i,n} &= \Lambda_0 + N_{i,n} S_{i,n} \\
&\quad + \frac{N_{i,n} k_0}{N_{i,n} + k_0}(\bar{\mu}_{i,n} - \mu_0)(\bar{\mu}_{i,n} - \mu_0)^t.
\end{aligned}$$

Then, the posterior distribution of $(\mu_{i,n}, \Sigma_{i,n})$ given c_n is the Gaussian/Inverted-Wishart distribution $\mathcal{N}(\mu_{i,n}; \tilde{\mu}_{i,n}, \frac{1}{N_{i,n}+k_0}\Sigma_{i,n})\mathcal{IW}(\Sigma_{i,n}; \tilde{\Lambda}_{i,n}, N_{i,n} + \nu_0)$.

parameter β of the Markov prior model $P(x \mid \beta, v)$

$$P(\beta \mid v) \propto 1. \tag{24}$$

5) Prior on the Number of Allocated Regions: We consider the number of allocated region labels $|v|$ as a random variable of the form $1 + p$, where p is a Poisson variable with mean λ. So, $\mathcal{P}(p \mid \lambda) = (\lambda^p e^{-\lambda}/p!)$. We set Jeffrey's prior on the mean: $P(\lambda) = \left(1/\sqrt{\lambda}\right)$. With that model, we obtain $P(p) = \int P(p \mid \lambda)P(\lambda)d\lambda = (1/p!)\int \lambda^{p-1/2}e^{-\lambda}d\lambda = (\Gamma(p + 1/2)/p!) = (\Gamma(|v| - 1/2)/\Gamma(|v|))$.

Taking also the number of permutations of the region labels into account, we obtain the following prior distribution on the vector of allowed region classes v defined by

$$P(v) \propto \frac{1}{|v|!}\frac{\Gamma\left(|v| - \frac{1}{2}\right)}{\Gamma(|v|)} = \frac{\Gamma\left(|v| - \frac{1}{2}\right)}{\Gamma(|v| + 1)\Gamma(|v|)}. \tag{25}$$

See Fig. 4 for an illustration of the shape of the proposed distribution.

6) Prior Distribution on the Parameters: Altogether, we obtain the following prior on the parameters θ:

$$P(\theta) \propto P(\pi \mid v)P(\mu, \Sigma)P(x \mid \beta, v)P(\beta \mid v)P(v) \tag{26}$$

as defined in (19), (23), (12), (24), and (25).

7) Likelihood: We find convenient to consider the augmented data $(\hat{y}, \hat{c})$. The joint distribution of the augmented model is described by (8), whereas the marginal distribution of $\hat{y}$ given x is described by (9). Thus, the corresponding likelihoods can be expressed as

$$P(\hat{y}, \hat{c} \mid \theta) = P(\hat{y} \mid \hat{c}, \mu, \Sigma)P(\hat{c} \mid x, \pi) \tag{27}$$

$$P(\hat{y} \mid \theta) = P(\hat{y} \mid x, \mu, \Sigma, \pi). \tag{28}$$

8) Posterior Distribution on the Parameters: Finally, the corresponding posterior distributions on the parameters θ are expressed as

$$P(\theta, \hat{c} \mid \hat{y}) \propto P(\hat{y}, \hat{c} \mid \theta)P(\theta) \tag{29}$$

$$P(\theta \mid \hat{y}) \propto P(\hat{y} \mid \theta)P(\theta). \tag{30}$$

The directed acyclic graph (DAG) presented in Fig. 5 summarizes the proposed Bayesian model.

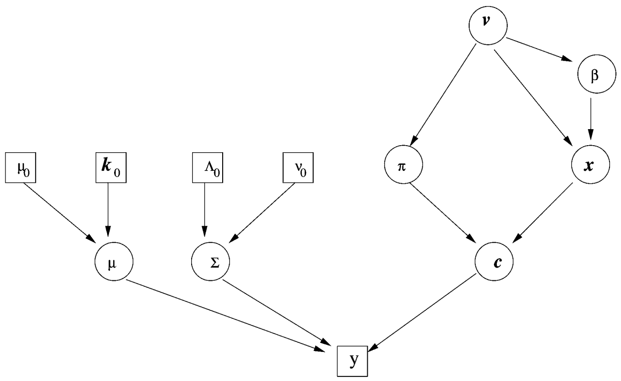

Fig. 5. DAG of the proposed fusion model.

9) Global Constraint: Given a segmentation x, let $R_1(x),\ldots,R_\ell(x)$ be the ℓ connected regions induced by x. That is, let G' be the graph with nodes the pixels of the image, in which two pixels s and t are connected if they are 8-neighbors and if they have the same label ($x_s = x_t$). Then, $R_1(x),\ldots,R_\ell(x)$ are the connected components of the graph G'. Under the "cubic law" of 3-D-object sizes [37], the probability of observing a connected component of size $|R|$ is proportional to $(1/|R|^2)$. If we let the size $|R|$ vary from 1 to the size $|G|$ of the image, the constant of proportionality is $(1/(1-|G|^{-1}))$.

Also, we consider the number of connected regions ℓ to be of the form $1+p$ where p is a Poisson variable of unknown mean λ. As in Section III-A5, the marginal distribution of ℓ is equal to $((\Gamma(\ell-(1/2)))/(\Gamma(\ell)))$.

Taking $-\log$ of the combined probabilities, we obtain a global constraint on the region process x of the form

$$\rho(x) = \sum_{i=1}^{\ell}\left\{2\log|R_i(x)| + \log\left(\frac{1-1}{|G|}\right)\right\} + \log\Gamma(\ell) - \log\Gamma\left(\ell-\frac{1}{2}\right). \quad (31)$$

Fig. 4 illustrates the shape of the proposed global constraint, in the special case of equal size components (i.e., $|R_1(x)| = \cdots = |R_\ell(x)| = |G|/\ell$). In this case, we took $|G| = 38400$ pixels. In practice, only the increasing part of the curve is relevant, since the function starts to decrease after as much as 12000 connected regions. The likelihood of a natural image prevents this case to occur. In our tests, the average number of connected regions was 521.89.

10) Weighted Map Estimator: Due to the intractable computation of the partition function $Z(\beta, v)$, the ML estimator of β cannot be computed. So, we replace the likelihood (as is often done) by the pseudolikelihood [38]. Since the pseudolikelihood estimator of an MRF is *consistent* [39], nothing is lost in the estimation of the hyper-parameter β, at least for sufficiently large images. The pseudolikelihood estimator $\hat{\beta}(x)$ is the maximum of the function

$$-\frac{1}{2}\sum_s \sum_{t\in N(s)} \beta U_{\langle s,t\rangle}(x) - \frac{1}{2}\sum_s \log\left(\sum_{k:v_k=1} e^{-\sum_{t\in N(s)}\beta U_{\langle s,t\rangle}(x|x_s=e_k)}\right) \quad (32)$$

where $N(s)$ is the set of neighbors of the pixel s. The factor 1/2 takes into account the fact that each binary clique is counted twice in the pseudolikelihood term.

Therefore, we propose the following *weighted* maximum *a posteriori* (MAP) estimator $\theta_* = (x_*, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*)$ of the fusion model of (30): the values of the parameters that maximize the function

$$P(x,\mu,\Sigma,\pi,\beta,v \mid \hat{y})\frac{Z(\beta,v)}{Z_p(x,\beta,v)}e^{-\rho(x)} \quad (33)$$

where the pseudopartition function $Z_p(x,\beta,v)$ is equal to

$$\prod_s \left(\sum_{k:v_k=1} e^{-\sum_{t\in N(s)}\beta U_{\langle s,t\rangle}(x|x_s=e_k)}\right)^{1/2}. \quad (34)$$

Equivalently, $(x_*, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*)$ is a global minimum of the energy function $\bar{f}(x,\mu,\Sigma,\pi,\beta,v \mid \hat{y})$ defined by

$$-\sum_{n=1}^{M}\log P(y_n \mid x, \mu_n, \Sigma_n, \pi_n) + \sum_{n=1}^{M}\log P(\pi_n \mid v) + \log P(\mu_n, \Sigma_n) + \beta U(x) + \log Z_p(x,\beta,v) - \log P(v) + \rho(x). \quad (35)$$

Note that β_* must be equal to the pseudolikelihood estimator $\hat{\beta}(x_*)$ on the optimal segmentation x_*. Also, the dependence of $\bar{f}$ on v is only implicit. Namely, v is the set of labels appearing in x (otherwise, $\bar{f}(x,\mu,\Sigma,\pi,\beta,v \mid \hat{y}) = \infty$). See Table VII for an algorithm that computes the pseudolikelihood estimator. This algorithm works because the function defined by (32) is a concave function.

11) Identifiability: An important property in statistics is the *identifiability* of the model (cf. [40]). In the context of the proposed model, this property can be stated as follows.

Theorem 1: Let $\theta_i = (x_i, \mu_i, \Sigma_i, \pi_i, \beta_i, v_i)$ with $\beta_i = \hat{\beta}(x_i)$, for $i = 1, 2$, be two vectors of parameters that induce the same values of the energy function $\bar{f}(\theta \mid \hat{y})$ *for any* observable data $\hat{y}$. Then, $\theta_1 = \theta_2$ (up to permutation of the indices).

The proof of the Theorem is postponed until Appendix I. A practical consequence is that, for large images, the parameters are uniquely determined by the observed data. In particular, the weighted MAP is practically unique.

B. Stochastic Algorithm

We find convenient to use the augmented data $\hat{c}$ in the calculation of $(x_*, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*)$. Also, we consider an auxiliary integer $1 \le L \le K$ that represents the number of allocated region labels so far. The role of L in the algorithm will appear clearer later. Thus, we consider the augmented vector $\psi = (x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L)$ and we define a function $f(x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L) = \bar{f}(x, \mu, \Sigma, \pi, \beta, v)$. Clearly, for any $\hat{c}$ and L, the vector $(x_*, \hat{c}, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*, L_*)$ is optimal for f if and only if $(x_*, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*)$ is optimal for $\bar{f}$. Thus, we want to optimize the augmented function f on the (augmented) search space A consisting of all admissible 8-tuples

TABLE VII
COMPUTATION OF THE MAXIMUM PSEUDOLIKELIHOOD PARAMETER

```
Let $x$ be a realization of the region process. Let $|v|$ be
the number of region classes appearing in $x$. Let $p'(\beta)$
be the derivative of equation (32) with respect to $\beta$.
  if $|v| = 1$ then
    Return $\beta = 0$.
  else
    Set $\beta_1 = 0$, and $\beta_2 = 1$.
    while $p'(\beta_2) > 0$ do
      Multiply $\beta_2$ by 2.
    end while
    while $\beta_2 - \beta_1 > 0.01$ do
      Set $\beta = (\beta_1 + \beta_2)/2$.
      if $p'(\beta) > 0$ then
        Set $\beta_1 = \beta$.
      else
        Set $\beta_2 = \beta$.
      end if
    end while
    Return $\beta$.
  end if
```

$(x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L)$; i.e., $v_k = 1$ if and only if $x_s = e_k$ for some pixel s, $v_k = 0$ for $k > L$, and $\beta = \hat{\beta}(x)$. Note that we can view the space A as finite, upon using the ε-machine of the computer. That is, in reality, we are working with a finite set!

In order to solve this optimization problem, we resort to O. François' Exploration/Selection () algorithm [16]. The goal of the algorithm is to minimize a function f on a finite set A. The algorithm relies on an exploration kernel $a(\psi, \psi')$, with $\psi, \psi' \in A$, which gives the probability of reaching ψ' from ψ under an exploration operator.

Given $m > 1$, a vector of m solutions in the Cartesian product A^m is denoted $\vec{\psi}$. Such a vector is called a population of solutions. Given a population of solutions $\vec{\psi} = (\psi_1, \dots, \psi_m)$, $\alpha(\vec{\psi})$ denotes the best solution (with minimal index in case of a tie) among $\psi_1, \dots, \psi_m$; i.e., $\alpha(\vec{\psi}) = \psi_l$, where $f(\psi_k) > f(\psi_l)$ for $k < l$, and $f(\psi_k) \geq f(\psi_l)$ for $k \geq l$.

At each step, two operators are available. An exploration operator that draws a new solution ψ'_l according the kernel $a(\psi_l, \cdot)$; a selection operator that replaces ψ_l by the best current solution $\alpha(\vec{\psi})$. The exploration operator is performed with probability p_t at iteration t. The probability of exploration p_t is set equal to $t^{-1/\tau}$, where $\tau \geq 1$ is a parameter of the algorithm. The random state of the vector $\vec{\psi}$ at iteration t is denoted $\psi^{[t]}$. The algorithm is summarized in Table VIII.

In the original version of the algorithm [16], the exploration kernel $a(\psi, \psi')$ is a uniform distribution on a neighborhood $N(\psi)$ of ψ with $\alpha(\vec{\psi})$ deleted. In this paper, the exploration kernel can be any distribution that satisfies the following hypothesis:

$$\text{For any } \psi, \psi' \in A, \quad a(\psi, \psi') > 0. \tag{36}$$

TABLE VIII
ES ALGORITHM IN ITS GENERAL FORM

```
Let $m > 1$ and $\tau \geq 1$. Goal: to minimize a function $f$
on a finite set $A$.
  Parameter initialization: Initialize randomly $\psi_l^{[0]}$, for
  $l = 1, ..., m$. Set $t = 1$.
  while a stopping criterion is not met do
    Update $t \leftarrow t + 1$.
    Determine the best current solution $\alpha(\vec{\psi}^{[t]}) = \psi_l^{[t]}$,
    where $f(\psi_k^{[t]}) > f(\psi_l^{[t]})$ for $k < l$, and $f(\psi_k^{[t]}) \geq$
    $f(\psi_l^{[t]})$ for $k > l$.
    for $l = 1, 2, ..., m$ do
      Let $u$ be a random number between 0 and 1.
      if $u \leq p_t = t^{-\frac{1}{\tau}}$ then
        Exploration: Replace $\psi_l^{[t]}$ by $\psi_l^{[t+1]}$ drawn
        according to an exploration kernel $a(\psi_l^{[t]}, \cdot)$.
      else
        Selection: Replace $\psi_l^{[t]}$ by $\alpha(\vec{\psi}^{[t]})$.
      end if
    end for
  end while
```

Theorem 2: Let hypothesis (36) hold. For any $t \geq 1$ and any $\varepsilon > 0$

$$\begin{aligned} P&\left\{f(\alpha(\vec{\psi}^{[t+1]})) \geq f_* + \varepsilon\right\} \\ &\leq t^{-(m/\tau)}(1 - P_\varepsilon)^m \left(1 - P\left\{f(\alpha(\vec{\psi}^{[t]})) \geq f_* + \varepsilon\right\}\right) \\ &\quad + (1 - P_\varepsilon t^{-(1/\tau)})^m P\left\{f(\alpha(\vec{\psi}^{[t]})) \geq f_* + \varepsilon\right\} \end{aligned}$$

where f_* is the global minimum of f on A, and $P_\varepsilon = \min_\psi P_{a(\psi,\cdot)}(f < f_* + \varepsilon)$.

Theorem 2 is useful because we obtain a sequence converging to 0, as shown in the following Lemma.

Lemma 1: Let b_t be the sequence defined recursively by

$$\begin{aligned} b_1 &= 1; \\ b_{t+1} &= t^{-(m/\tau)}(1 - P)^m(1 - b_t) + (1 - Pt^{-(1/\tau)})^m b_t \end{aligned}$$

where $0 < P \leq 1$, $m \geq 2$, and $\tau \geq 1$. Then, $\lim_{t \to \infty} b_t = 0$.

Corollary 1: For any $\varepsilon > 0$, $\lim_{t \to \infty} P\{f(\alpha(\vec{\psi}^{[t]})) \geq f_* + \varepsilon\} = 0$.

Proof: This follows directly from Theorem 2 and Lemma 1, upon observing that hypothesis (36) implies that $P_\varepsilon > 0$. ■

Thus, we recover a new proof of the asymptotic convergence of the ES algorithm under hypothesis (36). But Theorem 2 actually gives unilateral confidence intervals for $f(\alpha(\vec{\psi}^{[t]}))$. For instance, with $m = 6$, $\tau = 3$, and $P_\varepsilon = 0.01$, we obtain $P\left\{f(\alpha(\vec{\psi}^{[t]})) \geq f_* + \varepsilon\right\} \leq 1.225 \times 10^{-4}$ if $t \geq 1465$. If $t \geq 8352$, we obtain $P\left\{f(\alpha(\vec{\psi}^{[t]})) \geq f_* + \varepsilon\right\} \leq 4.912 \times 10^{-6}$.

In practice, condition (36) can be met as follows. Let $a(\psi, \psi')$ be any exploration kernel that satisfies the hypothesis

$$\begin{aligned} &\text{For any } \psi, \psi' \in A, \text{there exists } \psi_0 = \psi, \psi_1, \dots, \psi_D = \psi' \\ &\text{such that } a(\psi_i, \psi_{i+1}) > 0 \end{aligned} \tag{37}$$

TABLE IX
MODIFICATION OF AN ES EXPLORATION KERNEL a SATIFYING HYPOTHESIS (37) INTO A KERNEL $\tilde{a}$ THAT SATISFIES HYPOTHESIS (36)

```
Modified exploration kernel $\tilde{a}(\psi, \psi')$ from a kernel $a(\psi, \psi')$ satisfying hypothesis (37).
  Draw $N$ according to a Binomial distribution $\mathcal{B}(p, n)$ with $p = 0.1/(D-1)$ and $n = D-1$. Set $\psi_0 = \psi$.
  for $i = 0, ..., N$ do
    Draw $\psi_{i+1}$ according to the kernel $a(\psi_i, \cdot)$.
  end for
  Set $\psi' = \psi_{N+1}$.
```

where $D \geq 1$ is called the diameter. If $D = 1$, the kernel a satisfies itself condition (36). If $D > 1$, consider the modified kernel $\tilde{a}(\psi, \psi')$ defined by the algorithm of Table IX. The idea is to simply repeat the kernel a a random number of times between 1 and D. Then, clearly, the modified kernel satisfies condition (36), because there is a positive probability of performing consecutively D times the exploration according to the kernel a.

There is a closed connection between the ES algorithm and the simulated annealing. To see this, let $p_t = \exp(-(1/T))$, where $T > 0$ is called the temperature. Then, if $T = T(t)$ is given by the usual simulated annealing temperature schedule $T = (\tau/\log t)$, we recover $p_t = t^{-1/\tau}$ as in Table VIII. In fact, it is shown in [16] that the algorithm converges to an optimal solution (under hypothesis (37)) if and only if T is of the form $(\tau/\log t)$, with an appropriate value of τ (in particular, it is sufficient that $\tau \geq D$).

The ESE procedure [15] is a variant of the algorithm designed in the case where A is a space of parameters and f is $-\log$ the posterior distribution of the parameters conditional to the observed data. Again, after digitization of the space, A can be viewed as finite. The main idea is to use an MCMC kernel of the posterior distribution as exploration kernel. In practice, this crucial idea helps the algorithm perform efficiently; in particular, using a uniform distribution would yield a very poor algorithm in our case.

We can build systematically the MCMC kernel upon using the Gibbs sampler. Namely, one transition consists in performing the following sampling steps.

1) $v^{[t+1]} \sim v \mid x^{[t]}, c^{[t]}, \mu^{[t]}, \Sigma^{[t]}, \pi^{[t]}, \beta^{[t]}$.
2) $x^{[t+1]} \sim x \mid \hat{c}^{[t]}, \mu^{[t]}, \Sigma^{[t]}, \pi^{[t]}, \beta^{[t]}, v^{[t+1]}$.
3) $\hat{c}^{[t+1]} \sim \hat{c} \mid x^{[t+1]}, \mu^{[t]}, \Sigma^{[t]}, \pi^{[t]}, \beta^{[t]}, v^{[t+1]}$.
4) $\pi^{[t+1]} \sim \pi \mid x^{[t+1]}, \hat{c}^{[t+1]}, \mu^{[t]}, \Sigma^{[t]}, \beta^{[t]}, v^{[t+1]}$.
5) $\mu^{[t+1]}, \Sigma^{[t+1]} \sim \mu, \Sigma \mid x^{[t+1]}, \hat{c}^{[t+1]}, \pi^{[t+1]}, \beta^{[t]}, v^{[t+1]}$.
6) $\beta^{[t+1]} \sim \beta \mid x^{[t+1]}, \hat{c}^{[t+1]}, \mu^{[t+1]}, \Sigma^{[t+1]}, \pi^{[t+1]}, v^{[t+1]}$.

However, we need to bring some modifications to this general scheme.

- In step 2), one needs to perform many sweeps of the image, and moreover take into account the global constraint $\rho(x)$ in a Metropolis-Hastings (M-H) strategy. But this is unnecessary in our case, because the proposal/disposal mechanism of the M-H algorithm is replaced by the exploration/selection mechanism of the ES algorithm. The point is that our goal in this paper is not to simulate the posterior distribution of ψ, but rather to compute the MAP estimator.
- Step 1) should be combined with step 2) in a RJMCMC strategy in order to simulate the posterior distribution of (x, v) with jumps in dimension. Note that engineering a RJMCMC kernel that offers sufficiently high rates of acceptance is a hard task in practice. But again, all we need is an exploration kernel that satisfies hypothesis (37) (for some value of D). So, we replace step 1) by an *ad hoc* exploration $v^{[t+1]} \sim v \mid v^{[t]}$, as described in Table X. The point is that in the sampling of the region labels $x^{[t+1]}$ in step 2), only those labels allowed by $v^{[t+1]}$ are used. In this manner, an artificial jump in dimension is performed.
- In step 4), whenever a region class e_k is empty, we do not simulate the mixture proportions according to the prior distribution. Rather, we keep the former values of the mixture proportions for this region class intact for a subsequent iteration. Similarly, for step 5).
- As explained in Section III-A10, in step 6), we are actually interested in taking $\beta^{[t+1]} = \hat{\beta}(x^{[t+1]})$ instead of simulating β.

The resulting modified Gibbs sampler is presented in Table X.

Now, we want to start with one region (i.e., $L = 1$) and let the number of allowed regions grow gradually until it reaches the maximal value K (i.e., $L = K$). In order to do so, we consider an operator of birth of a region explained in Table XI. The idea of using such an operator can be found in [22]. Altogether, the exploration kernel $a(\cdot, \cdot)$ used in this paper consists of Table XI followed by Table X. The resulting kernel satisfies hypothesis (37) with $D = K$. Then, use Table IX with $D = K$ to obtain a kernel $\tilde{a}$ that satisfies hypothesis (36). Note that at the intermediate steps, the vector ψ might not be admissible, but that at the output ψ *is* admissible.

Finally, we present the initialization steps in Table XII. The results above imply that the whole procedure converges asymptotically to the weighted MAP $(x_*, \mu_*, \Sigma_*, \pi_*, \beta_*, v_*)$, with probability 1. In our tests, we took $m = 6$ and $\tau = 3$ in Table VIII. Furthermore, we waited for the first 10 iterations before increasing the number of allowed regions (cf. Table XI).

IV. EXPERIMENTAL RESULTS

We have tested the proposed method of estimation and segmentation on the University of California at Berkeley (UCB) test dataset [41] of 100 natural images in ".jpg" format. We think that all of them are optical images obtained by electronic acquisition, though we do not have that information at hand. The typical size of an image was 481 × 321. Each image was reduced by a factor of 50%. In our implementation in C++, we use the GNU scientific library of functions.

We performed for each natural image I, a joint estimation and segmentation $(x_*, \phi_*, \pi_*, \beta_*, v_*)$ based on the observed channels data $\hat{y}(I)$, with a maximal number of $K = 40$ allowed classes, and a fixed number of $K_1 = 16$ color classes and $K_2 = 16$ texton classes. This represents a task of estimating 38 400 color labels, 38 400 texton labels, 38 400 region labels, 144 Gaussian color parameters, 864 Gaussian texton parameters, 30 mixture parameters per region class, and one Markovian hyper-parameter. We then simulated the image channels $\hat{y}'$ based on that estimation. Thus, $\hat{y}'$ and $(x_*, \phi_*, \pi_*, \beta_*, v_*)$ were considered as ground-truth. Note that the image I' itself

TABLE X
MODIFIED GIBBS SAMPLER FOR THE ESE PROCEDURE

Let $\psi = (x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L)$. Modified Gibbs sampler: $\psi^{[t]} \rightarrow \psi^{[t+1]}$.

1) ($v^{[t+1]} \sim v \mid v^{[t]}$) Let $L^{[t]}$ be the number of labels allocated so far. Modify each bit of $v^{[t]}$, with index $k \leq L^{[t]}$, with probability $\frac{1}{2L^{[t]}}$. If they are all equal to 0, set one of them equal to 1 randomly. Let $v^{[t+1]}$ be the resulting vector of allowed classes, and set $L^{[t+1]} = L^{[t]}$.

2) ($x^{[t+1]} \sim x \mid \hat{c}^{[t]}, \mu^{[t]}, \Sigma^{[t]}, \pi^{[t]}, \beta^{[t]}, v^{[t+1]}$) For two sweeps, set $x_s = e_k$ at each site s of the image graph V according to the weights

$$\exp\left\{-\beta^{[t]} \sum_{s' \in N(s)} U_{\langle s,s' \rangle}(x \mid x_s = e_k)\right\} \prod_{n=1}^{M} \pi^{[t]}_{(i_n,k),n},$$

where $N(s)$ denotes the set of 8-neighbors of s in the image graph G, $c_{s,n} = g_{i_n,n}$, and e_k ranges over the set of allocated classes by $v^{[t+1]}$. Let $x^{[t+1]}$ be the updated segmentation. Readjust $v^{[t+1]}$ so that $v_k^{[t+1]} = 1$ if and only if $x_s^{[t+1]} = e_k$ for some pixel s.

3) ($\hat{c}^{[t+1]} \sim \hat{c} | x^{[t+1]}, \mu^{[t]}, \Sigma^{[t]}, \pi^{[t]}, \beta^{[t]}, v^{[t+1]}$) At each level n, for each pixel s, draw $c_{s,n}$ according to the weights

$$P(y_{s,n} \mid c_{s,n} = g_{i,n}, \mu^{[t]}_{i,n}, \Sigma^{[t]}_{i,n})\pi^{[t]}_{(i,k),n},$$

where $x_s = e_k$. Let $c^{[t+1]}$ be the updated segmentations.

4) ($\pi^{[t+1]} \sim \pi \mid x^{[t+1]}, \hat{c}^{[t+1]}, \mu^{[t]}, \Sigma^{[t]}, \beta^{[t]}, v^{[t+1]} = \pi \mid x^{[t+1]}, \hat{c}^{[t+1]}, v^{[t+1]}$) For each level n and for each allowed region class e_k, simulate $\pi_{(i,k),n}$ according to the posterior distribution of Table II. But, if ever the class is empty, the former value of $\pi_{(i,k),n}$ is kept. Let $\pi^{[t+1]}$ be the resulting mixture proportions.

5) ($\mu^{[t+1]}, \Sigma^{[t+1]} \sim \mu, \Sigma \mid x^{[t+1]}, \hat{c}^{[t+1]}, \pi^{[t+1]}, \beta^{[t]}, v^{[t+1]} = \mu, \Sigma \mid \hat{c}^{[t+1]}$) For each level n and each non-empty class $g_{i,n}$, simulate $\mu_{i,n}, \Sigma_{i,n}$ according to the posterior distribution of Table VI. But, if ever $c_{s,n} \neq g_{i,n}$ for all $s \in V$, the former values of $\mu_{i,n}, \Sigma_{i,n}$ are kept. Let $\mu^{[t+1]}, \Sigma^{[t+1]}$ be the resulting likelihood parameters.

6) Compute the pseudo-likelihood estimator $\beta^{[t+1]} = \hat{\beta}(x^{[t+1]})$ as in Table VII.

TABLE XI
OPERATOR OF BIRTH OF A REGION FOR THE ESE PROCEDURE

```
Let $\psi = (x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L)$. Birth of a region: $\psi^{[t]} \rightarrow \psi^{[t+1]}$.
  Let $L^{[t]}$ be the number of classes allocated in $v^{[t]}$ so far. Set $L^{[t+1]} = L^{[t]} + 1$, provided $L^{[t]} < K$.
  Otherwise, set $L^{[t+1]} = L^{[t]}$ and return.
  Set $v_k^{[t+1]} = v_k^{[t]}$, for $k \leq L^{[t]}$, and $v_{L^{[t+1]}}^{[t+1]} = 1$.
  for each level of analysis $n$ do
    Sample $(\pi_{(1,L^{[t+1]}),n}, ..., \pi_{(K_n,L^{[t+1]}),n})$ according to the Dirichlet prior $\mathcal{D}(A_0, \alpha_1, ..., \alpha_{K_n})$.
  end for
  Let $\pi_{L^{[t+1]}}^{[t+1]}$ be the resulting proportions and set $\pi_k^{[t+1]} = \pi_k^{[t]}$ for $k \leq L^{[t]}$.
```

TABLE XII
INITIALIZATION OF THE ESE PROCEDURE

Let $\psi = (x, \hat{c}, \mu, \Sigma, \pi, \beta, v, L)$. Initialization of $\psi^{[0]}$.

1) Set $v_1^{[0]} = 1$, and $v_k^{[0]} = 0$ for $k \geq 2$. Set $L^{[0]} = 1$.
2) Set $x_s^{[0]} = e_1$, for all pixels s.
3) Random initialization of $\hat{c}^{[0]}$.
4) Perform steps 4 to 6 of the exploration kernel of Table X.

was not simulated. Next, we performed a joint estimation and segmentation $(x'_*, \phi'_*, \pi'_*, \beta'_*, v'_*)$ for the synthetic image, with again $K = 40$ and $K_1 = K_2 = 16$. We took $m = 6$ and $\tau = 3$ as internal parameters of the ESE procedure. The procedure was stopped after $t = 1465$ iterations (see Section III-B).

We evaluated the estimation error with the measure

$$\Delta_0 = \left|\bar{f}\left(x_*, \phi_*, \pi_*, \beta_*, v_* \mid \hat{y}'\right) - \bar{f}\left(x'_*, \phi'_*, \pi'_*, \beta'_*, v'_* \mid \hat{y}'\right)\right|$$

where $\bar{f}$ is defined by (35), as well as the relative measure proposed in [15]

$$\Delta_1 = \frac{\left|\bar{f}\left(x_*, \phi_*, \pi_*, \beta_*, v_* \mid \hat{y}'\right) - \bar{f}\left(x'_*, \phi'_*, \pi'_*, \beta'_*, v'_* \mid \hat{y}'\right)\right|}{\left|\bar{f}\left(x_*, \phi_*, \pi_*, \beta_*, v_* \mid \hat{y}'\right)\right|} \times 100\%.$$

For the first estimation, the average number of region classes was $31.08 \pm 4.208SD$ (the maximum 40 was reached for only one image out of 100), while the average number of connected regions was $489.79 \pm 418.18SD$ (or $349.12 \pm 234.34SD$ if singletons are omitted). For the second estimation, the average number of region classes was $32.03 \pm 6.509SD$ (the maximum 40 was never reached), while the average number of connected regions was $467.83 \pm 408.16SD$ (or $236.12 \pm 159.14SD$ if singletons are omitted). The ESE procedure took on average 3 h and 26 min. on a Workstation 2.4 GHz for an average of 1046.44 explorations. This represents roughly 11.81 s/exploration. The complexity of each exploration is actually linear in the size of the image times the number of region classes.

See Fig. 7 for a histogram of Δ_0 and Δ_1 over the dataset, and Figs. 1, 6 and 8 for examples of segmentations. The three images 175043.jpg, 38082.jpg, and 69040.jpg were totally missed ($\Delta_0 = 2.145700, 2.371300, 1.792500$, and $\Delta_1 = 6.954818\%, , 7.005151\%, 5.494504\%$, respectively). In fact, the current number of allowed region classes was only 1 at iteration 1465. We increased the number of iterations to 8352 and obtained successfully $\Delta_0 = 0.499901, 0.152901, 0.212399$ and $\Delta_1 = 1.538261\%, 0.423909\%, 0.620983\%$, respectively.

V. CONCLUSION

We have presented an HMRF data-fusion model based on Julesz ensembles and applied it to the segmentation of natural images. The ESE procedure [15] is a general method for estimating the weighted modes of HMRF models, with global

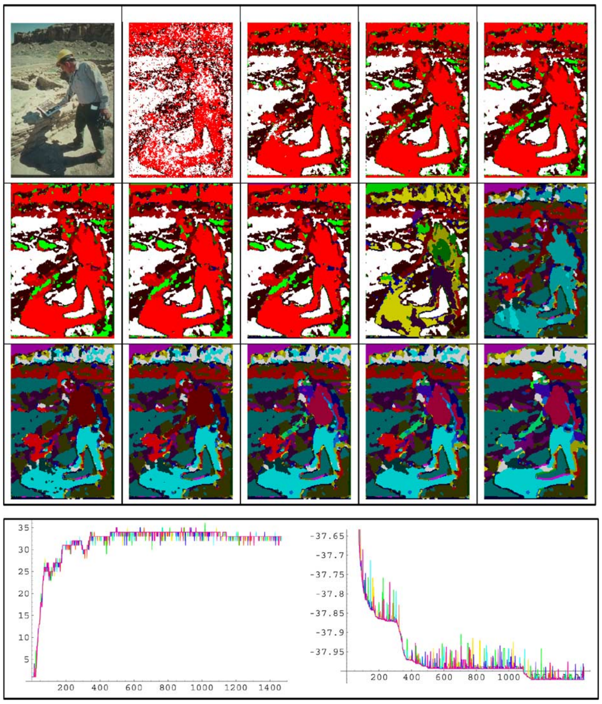

Fig. 6. Top: A natural image and the formation of its region process at iterations $25, 30, 35, \ldots, 85$, and 1465. Bottom: evolution of the number of region classes and the Gibbs energy of six solutions as a function of the iteration.

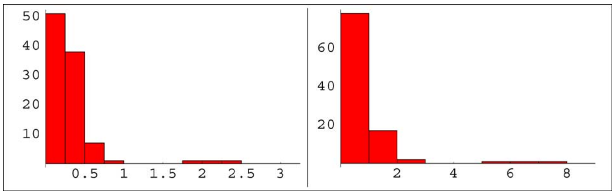

Fig. 7. Histograms of the evaluation measures Δ_0 and Δ_1 over the dataset. Mean of Δ_0: 0.308919. Mean of Δ_1: 0.84%.

constraints taken into account. We have shown how to adapt it to the proposed data fusion model. Not only the parameters of the Gaussian kernels and the region labels were estimated, but also the mixture proportions, the number of regions, and the Markov hyper-parameter. The internal parameters of the algorithm that insure asymptotic convergence to an optimal solution are known explicitly and are practical [15]. Furthermore, we have presented new finite time bounds for the rate of convergence. The tests reported in this paper indicate that the ESE procedure succeeds in finding the optimal solution of the proposed fusion model, within a relative error bound of less than 0.87% on average.

It remains to test the fusion model in various higher-level tasks, such as image indexing, 3-D-reconstruction, motion detection, or localization of shapes, in combination with prior knowledge on the particular problem.

Appendix I

We present in this Appendix a proof of Theorem 1 of Section III-A11.

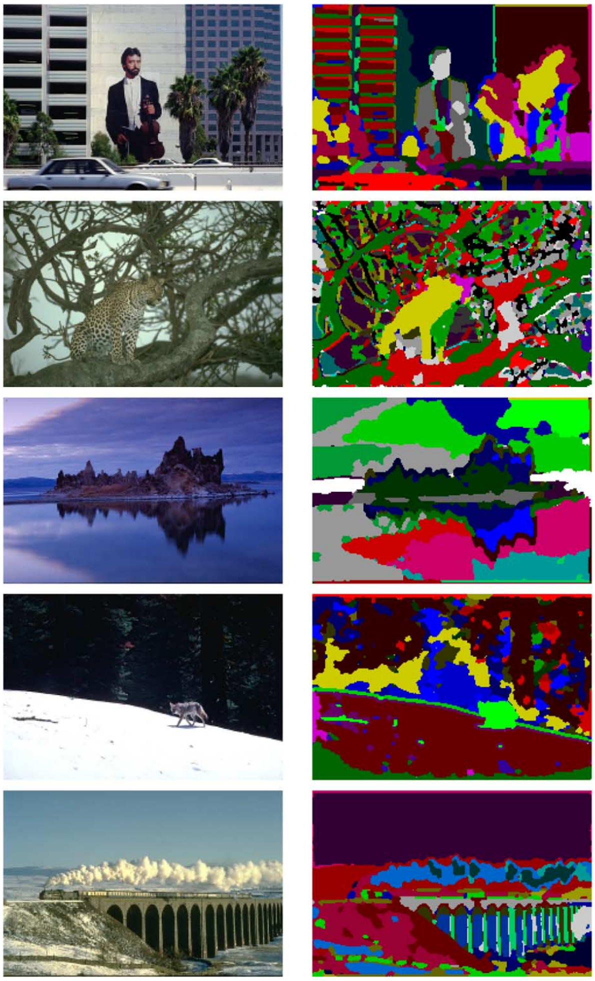

Fig. 8. Examples of segmentations based on a fusion of colors and textons.

Let $\theta_i = (x_i, \mu_i, \Sigma_i, \pi_i, \beta_i, v_i)$, $i = 1, 2$ be two vectors of parameters, with $\beta_i = \hat{\beta}(x_i)$. Assume that $\bar{f}(\theta_1 \mid \hat{y}) = \bar{f}(\theta_2 \mid \hat{y})$ for all $\hat{y}$. This means that

$$\begin{aligned}\prod_{n=1}^{M} \prod_{s} & P(y_{s,n} \mid x_1, \mu_{1,n}, \Sigma_{1,n}, \pi_{1,n}) \\ &= \prod_{n=1}^{M} \prod_{s} P(y_{s,n} \mid x_2, \mu_{2,n}, \Sigma_{2,n}, \pi_{2,n}) \\ &\quad \times \frac{P(\pi_{2,n} \mid v_2) P(\mu_{2,n}, \Sigma_{2,n})}{P(\pi_{1,n} \mid v_1) P(\mu_{1,n}, \Sigma_{1,n})} \\ &\quad \times \frac{P(\beta_2 \mid v_2) e^{-\beta_2 U(x_2)} Z_p(x_2, \beta_2, v_2)^{-1} e^{-\rho(x_2)} P(v_2)}{P(\beta_1 \mid v_1) e^{-\beta_1 U(x_1)} Z_p(x_1, \beta_1, v_1)^{-1} e^{-\rho(x_1)} P(v_1)}.\end{aligned}$$

In particular, we obtain an equality of distributions

$$\prod_{n=1}^{M} \prod_{s} P(y_{s,n} \mid x_1, \mu_{1,n}, \Sigma_{1,n}, \pi_{1,n}) = \prod_{n=1}^{M} \prod_{s} P(y_{s,n} \mid x_2, \mu_{2,n}, \Sigma_{2,n}, \pi_{2,n})$$

as well as the equality

$$1 = \frac{P(\pi_{2,n} \mid v_2)P(\mu_{2,n}, \Sigma_{2,n})}{P(\pi_{1,n} \mid v_1)P(\mu_{1,n}, \Sigma_{1,n})} \times \frac{P(\beta_2 \mid v_2)e^{-\beta_2 U(x_2)} Z_p(x_2, \beta_2, v_2)^{-1} e^{-\rho(x_2)} P(v_2)}{P(\beta_1 \mid v_1)e^{-\beta_1 U(x_1)} Z_p(x_1, \beta_1, v_1)^{-1} e^{-\rho(x_1)} P(v_1)}.$$

Considering the marginals, we deduce

$$P(y_{s,n} \mid x_1, \mu_{1,n}, \Sigma_{1,n}, \pi_{1,n}) = P(y_{s,n} \mid x_2, \mu_{2,n}, \Sigma_{2,n}, \pi_{2,n})$$

for each pixel s and each model n. Indeed, the variables $y_{s,n}$ are mutually independent. It follows at once from the identifiability property [40] of mixtures of Gaussian distributions, that after relabeling of indices, $\pi_1 = \pi_2$ on allowed region classes, and $\mu_1 = \mu_2$, $\Sigma_1 = \Sigma_2$ on cue classes. We deduce immediately that $x_1 = x_2$ with probability (w. p.) 1, since distinct regions have distinct mixture proportions w. p. 1. Furthermore, $v_1 = v_2$, since v indicates which region labels are present in the segmentation x.

Finally, we conclude that $\beta_1 = \beta_2$ because β_i is the pseudo-likelihood estimator of x_i.

Appendix II

The purpose of this appendix is to give an upper bound on the rate of convergence of the ES algorithm of Table VIII. The hypothesis is given in (36) and the notation is as in Section III-B. Our approach is inspired by previous work on genetic algorithms [42] and [43].

We now present the proof of the main result. After normalization, we may assume without loss of generality that the global minimum of the function f is equal to $f_* = 0$.

*Proof: (of Theorem 2)*Let q_t be the Markov transition matrix associated with the chain $(\vec{\psi}^{[t]})$; i.e., $q_t(\vec{\psi}, \vec{\psi}') = P(\vec{\psi}^{[t+1]} = \vec{\psi}' \mid \vec{\psi}^{[t]} = \vec{\psi})$. From Table VIII, we have for any $\vec{\psi}, \vec{\psi}' \in A^m$

$$q_t(\vec{\psi}, \vec{\psi}') = \prod_{l=1}^{m} \left(p_t a\left(\psi_{[l]}, \psi'_{[l]}\right) + (1 - p_t)\delta\left(\alpha(\vec{\psi}), \psi'_{[l]}\right)\right) \tag{38}$$

where δ denotes the Kronecker symbol.

Given $\varepsilon > 0$, let χ_ε be the characteristic function of the event $\{f(\psi) \geq \varepsilon\}$; i.e., $\chi(\psi) = 1$ if $f(\psi) \geq \varepsilon$, and $\chi(\psi) = 0$ otherwise. Then, $P\{f(\alpha(\vec{\psi}^{[t+1]})) \geq \varepsilon\} = E[\chi_\varepsilon(\alpha(\vec{\psi}^{[t+1]}))]$. Let $\pi^{[t]}$ denote the distribution of $\vec{\psi}^{[t]}$. Since $\pi^{[t+1]}(\vec{\psi}') = \int_{\vec{\psi}} q_t(\vec{\psi}, \vec{\psi}')\pi^{[t]}(\vec{\psi})d\vec{\psi}$, we compute

$$\begin{aligned} E&\left[\chi_\varepsilon(\alpha(\vec{\psi}^{[t+1]}))\right] \\ &= \int_{\vec{\psi}'} \chi_\varepsilon(\alpha(\vec{\psi}'))\pi^{[t+1]}(\vec{\psi}')d\vec{\psi}' \\ &= \int_{\vec{\psi}'} \chi_\varepsilon(\alpha(\vec{\psi}')) \int_{\vec{\psi}} q_t(\vec{\psi}, \vec{\psi}')\pi^{[t]}(\vec{\psi})d\vec{\psi}d\vec{\psi}' \\ &= \int_{\vec{\psi}} \left\{ \int_{\vec{\psi}'} \chi_\varepsilon(\alpha(\vec{\psi}'))q_t(\vec{\psi}, \vec{\psi}')d\vec{\psi}' \right\} \pi^{[t]}(\vec{\psi})d\vec{\psi} \\ &= \int_{\vec{\psi}\in E_0} \left\{ \int_{\vec{\psi}'\in E_1} \chi_\varepsilon(\alpha(\vec{\psi}'))q_t(\vec{\psi}, \vec{\psi}')d\vec{\psi}' \right\} \\ &\quad \times \pi^{[t]}(\vec{\psi})d\vec{\psi} && (39) \\ &\quad + \int_{\vec{\psi}\in E_1} \left\{ \int_{\vec{\psi}'\in E_1} \chi_\varepsilon(\alpha(\vec{\psi}'))q_t(\vec{\psi}, \vec{\psi}')d\vec{\psi}' \right\} \\ &\quad \times \pi^{[t]}(\vec{\psi})d\vec{\psi} && (40) \end{aligned}$$

where $E_0 = \{\vec{\psi} \mid \chi(\alpha(\vec{\psi})) = 0\}$ and $E_1 = \{\vec{\psi} \mid \chi(\alpha(\vec{\psi})) = 1\}$. In the first term (39), we have $f(\alpha(\vec{\psi})) < \varepsilon$, whereas $f(\psi'_l) \geq \varepsilon$, for any l. Thus, the equality $\psi'_l = \alpha(\vec{\psi})$ never occurs. It follows from (38) that $q_t(\vec{\psi}, \vec{\psi}') = \prod_{l=1}^{m} p_t a(\psi_{[l]}, \psi'_{[l]})$. Thus, we obtain

$$\begin{aligned} &\int_{\vec{\psi}\in E_0} \left\{ \int_{\vec{\psi}'\in E_1} \chi_\varepsilon(\alpha(\vec{\psi}'))q_t(\vec{\psi}, \vec{\psi}')d\vec{\psi}' \right\} \pi^{[t]}(\vec{\psi})d\vec{\psi} \\ &\quad \leq \int_{\vec{\psi}\in E_0} p_t^m (1 - P_\varepsilon)^m \pi^{[t]}(\vec{\psi})d\vec{\psi} \\ &\quad \leq p_t^m (1 - P_\varepsilon)^m \left(1 - P\left\{f\left(\alpha(\vec{\psi}^{[t]})\right) \geq \varepsilon\right\}\right). \end{aligned} \tag{41}$$

In the second term (40), we have $f(\psi'_l) \geq \varepsilon$, for any l. Thus, we obtain

$$\begin{aligned} &\int_{\vec{\psi}\in E_1} \left\{ \int_{\vec{\psi}'\in E_1} \chi_\varepsilon(\alpha(\vec{\psi}'))q_t(\vec{\psi}, \vec{\psi}')d\vec{\psi}' \right\} \pi^{[t]}(\vec{\psi})d\vec{\psi} \\ &\quad \leq \int_{\vec{\psi}\in E_1} \{p_t(1 - P_\varepsilon) + (1 - p_t)\}^m \pi^{[t]}(\vec{\psi})d\vec{\psi} \\ &\quad = (1 - P_\varepsilon p_t)^m P\left\{f\left(\alpha(\vec{\psi}^{[t]})\right) \geq \varepsilon\right\}. \end{aligned} \tag{42}$$

This completes the proof of the Theorem, upon setting $p_t = t^{-1/\tau}$.

Finally, we prove the lemma.

*Proof: (of Lemma 1)*We rewrite the recursion for b_t as follows:

$$\begin{aligned} b_t &= 1, \text{ for } t = 1 \\ b_{t+1} &= \alpha(t)(1 - b_t) + \beta(t)b_t, \text{ for } t \geq 1 \end{aligned}$$

where $\alpha(t) = t^{-(m/\tau)}(1 - P)^m$ and $\beta(t) = (1 - Pt^{-(1/\tau)})^m$.

First of all, we claim that $0 \leq b_t \leq 1$, for all $t \geq 1$. For $t = 1$, the property holds by definition. Assume that the property holds for t; i.e., $b_t \in [0, 1]$. Then, b_{t+1} is located *between* the numbers $\alpha(t)$ and $\beta(t)$. Since both of them are in the interval $[0, 1]$, the same holds true for b_{t+1}.

Next, we claim that it is sufficient that $\lim_{t\to\infty} b_{2t} = 0$. Indeed, we have $b_{2t+1} \leq (2t)^{-(m/\tau)} + b_{2t}$. Thus, it follows that $\lim_{t\to\infty} b_{2t+1} = 0$. We now show that $\lim_{t\to\infty} b_{2t} = 0$.

Case 1: $\tau > 1$.

Since $\lim_{t\to\infty} \alpha(t) = 0$, and $\lim_{t\to\infty} \beta(2t) = 1$, we can take t sufficiently large so that $\alpha(t) < \beta(2t)$. Note also that the sequence $\alpha(t)$ is decreasing, whereas the sequence $\beta(t)$ is increasing.

Fix t, and consider the sequence

$$\begin{aligned} c_k &= b_t, \text{ for } k = t \\ c_{k+1} &= \alpha(t)(1 - c_k) + \beta(2t)c_k, \text{ for } k \geq t. \end{aligned}$$

We claim that $b_k \leq c_k$ for $k \in [t, 2t]$. For $k = t$, the property is immediate. Assume the property true for some $k \in [t, 2t)$. We compute

$$\begin{aligned} b_{k+1} &= \alpha(k)(1-b_k) + \beta(k)b_k \leq \alpha(t)(1-b_k) + \beta(2t)b_k \\ &= \alpha(t) + (\beta(2t)-\alpha(t))b_k \leq \alpha(t) + (\beta(2t)-\alpha(t))c_k \\ &= c_{k+1} \end{aligned}$$

which proves the claim. In particular, $b_{2t} \leq c_{2t}$.

Now, the sequence c_k can be solved explicitly. Namely, we have

$$\begin{aligned} c_k = &\left(b_t - \frac{\alpha(t)}{1-\beta(2t)+\alpha(t)}\right)(\beta(2t)-\alpha(t))^{(k-t)} \\ &+ \frac{\alpha(t)}{1-\beta(2t)+\alpha(t)}, \text{ for } k \geq t. \end{aligned}$$

Indeed, this sequence satisfies the recursive definition of c_k. In particular, we obtain

$$b_{2t} \leq \beta(2t)^t + \frac{\alpha(t)}{1-\beta(2t)+\alpha(t)}$$

upon making use of the facts that $b_t \leq 1$ and $\alpha(t) \leq \beta(2t)$.

One can check that $\tau > 1$ implies that $\lim_{t\to\infty} t \log \beta(2t) = -\infty$; also, $m > 1$ implies that $\lim_{t\to\infty}(\alpha(t)/(1-\beta(2t)+\alpha(t))) = 0$. Therefore, the right-hand side converges to 0, and we are done. One can actually show more: the right-hand side is of the same order as $t^{-(m-1/\tau)}$, where the constant of proportionality depends on P_ε. We skip the details here.

a) Case 2: $\tau = 1$.

We then have $b_{k+1} \leq k^{-m} + (1-Pk^{-1})^m b_k$. Fixing t, we thus obtain $b_{t+N} \leq \sum_{k=t}^{N-1} k^{-m} + \prod_{k=t}^{N-1}(1-Pk^{-1})^m b_t$. But now, $\prod_{k=t}^{\infty}(1-Pk^{-1}) = 0$ since $\sum_{k=t}^{\infty}\log(1-Pk^{-1}) = -\infty$. Thus, $\limsup b_t \leq \sum_{k=t}^{\infty} k^{-m}$. Since $m > 1$, the series $\sum_{k=1}^{\infty} k^{-m}$ converges. Thus, $\lim_{t\to\infty}\sum_{k=t}^{\infty} k^{-m} = 0$, and we are done.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and questions that helped them improve both the technical content and the presentation quality of this paper.

References

[1] A. Mohammad-Djafari, "Probabilistic methods for data fusion," in *Proc. Maximum Entropy Bayesian Methods*, Boise, ID, 1998, pp. 57–69.

[2] S. C. Zhu, X. W. Liu, and Y. N. Wu, "Exploring texture ensembles by efficient Markov chain Monte Carlo—Toward a trichromacy theory of texture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 554–569, Jun. 2000.

[3] O. Viveros-Cancino, X. Descombes, J. Zerubia, and N. Baghdadi, "Fusion of radiometry and textural information for SIR-C image classification," presented at the 9th IEEE Int. Conf. Image Processing Rochester, NY, Sep. 2002.

[4] O. Féron and A. Mohammad-Djafari, "A hidden Markov model for Bayesian fusion of multivariate signals," presented at the 5th Int. Triennial Calcutta Symp. Probability and Statistics Kolkata, India, Dec. 2003.

[5] Z. Kato and T. C. Pong, W. Skarbek, Ed., "A Markov random field image segmentation model using combined color and texture features," in *Proc. Int. Conf. Computer Analysis of Images and Patterns*, Warsaw, Poland, Sep. 2001, pp. 547–554.

[6] W. Pieczynski, J. Bouvrais, and C. Michel, "Estimation of generalized mixture in the case of correlated sensors," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 308–311, Feb. 2000.

[7] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[8] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Comput.*, vol. 9, no. 8, 1997.

[9] ——, "Filters, random fields and maximum entropy (FRAME)," *Int. J. Comput. Vis.*, vol. 27, no. 2, 1998.

[10] ——, "Equivalence of Julesz ensembles and Frame models," *Int. J. Comput. Vis.*, vol. 38, no. 3, pp. 247–265, 2000.

[11] Z. W. Tu and S. C. Zhu, "Image segmentation by data-driven Markov Chain Monte Carlo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 657–673, May 2002.

[12] Z. Kato, T.-C. Pong, and S. G. Qiang, "Unsupervised segmentation of color textures images using a multi-layer MRF model," presented at the 10th IEEE Int. Conf. Image Processing Barcelona, Spain, Sep. 2003.

[13] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, "Multisensor image segmentation using Dempster-Shafer fusion in Markov fields context," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 8, pp. 25–32, Aug. 2001.

[14] A. Sarkar, A. Banerjee, N. Banerjee, S. Brahma, B. Kartkeyan, M. Chakraborty, and K. L. Majumder, "Landcover classification in MRF context using Dempster-Shafer fusion for multisensor imagery," *IEEE Trans. Image Process.*, vol. 14, no. 5, pp. 634–645, May 2005.

[15] F. Destrempes, M. Mignotte, and J.-F. Angers, "A stochastic method for Bayesian estimation of hidden Markov random field models with application to a color model," *IEEE Trans. Image Process.*, vol. 14, no. 8, pp. 1096–1124, Aug. 2005.

[16] O. François, "Global optimization with exploration/selection algorithms and simulated annealing," *Ann. Appl. Probab.*, vol. 12, no. 1, pp. 248–271, 2002.

[17] P. D. Moral, *Genealogical and Interacting Particle Approximations*, ser. Probability and Applications. New York: Springer, 2004.

[18] P. D. Moral and L. Miclo, "On the convergence and the applications of the generalized simulated annealing," *SIAM J. Control Optim.*, vol. 37, no. 4, pp. 1222–1250, 1999.

[19] P. J. Green, "Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[20] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unkown number of components," *J. Roy. Statist. Soc.*, vol. 59, no. 4, pp. 731–792, 1997.

[21] Z. Kato, Bayesian Color Image Segmentation Using Reversible Jump Markov Chain Monte Carlo, ERCIM, 1999, Res. Rep.01/99-R055.

[22] M. Stephens, "Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods," *Ann. Statist.*, vol. 28, no. 1, pp. 40–74, 2000.

[23] P. J. Green and A. Mira, "Delayed rejection in reversible jump metropolis-hastings," *J. Roy. Statist. Soc. B*, vol. 88, no. 4, pp. 1035–1053, 2001.

[24] O. Cappé, C. P. Robert, and T. Rydén, "Reversible jump, birth-and-death and more general continuous time Markov Chain Monte Carlo samplers," *J. Roy. Statist. Soc. B*, vol. 65, pp. 679–700, 2003.

[25] A. Barbu and S.-C. Zhu, "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1239–1253, Aug. 2005.

[26] R. H. Swendsen and J. S. Wang, "Nonuniversal critical dynamics in Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 58, no. 2, pp. 86–88, 1987.

[27] S. C. Zhu, C. Guo, Y. Wang, and Z. Xu, "What are textons," *Int. J. Comput. Vis.*, vol. 62, no. 1, pp. 121–143, 2005.

[28] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 7–27, 2001.

[29] C. A. Bouman and K. Sauer, "A unified approach to statistical tomography using coordinate descent optimization," *IEEE Trans. Image Process.*, vol. 5, no. 3, pp. 480–492, Mar. 1996.

[30] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics, Principles and Practice, Second Edition in C*. New York: Addison-Wesley, 1996.

[31] Coloremetry Publication CIE 15.2-1986, 2nd.

[32] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, 1992.

[33] J. G. Daughman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer.*, vol. 2, no. 7, pp. 1160–1169, 1985.

[34] M. Escobar, "Estimating normal means with a Dirichlet process prior," *J. Amer. Statist. Assoc.*, vol. 89, no. 425, pp. 268–277, 1993.

[35] T. W. Anderson*, An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1971.

[36] M. C. Jones, "Generating inverse Wishart matrices," *Commun. Statist. Simul. Comput.*, vol. 14, no. 2, pp. 511–514, 1985.

[37] A. B. Lee, J. G. Huang, and D. B. Mumford, "Occlusion models for natural images," *Int. J. Comput. Vis.*, vol. 41, pp. 33–59, 2001.

[38] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, vol. 24, pp. 179–195, 1977.

[39] F. Comets, "On consistency of a class of estimators for exponential fanilies of Markov random fields on the lattice," *Ann. Statist.*, vol. 20, pp. 455–468, 1992.

[40] D. M. Titterington, A. F. Smith, and U. E. Makov*, Statistical Analysis of Finite Mixture Distributions*, ser. Probability and Mathematical Statistics. New York: Wiley, 1992.

[41] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Computer Vision*, Jul. 2001, vol. 2, pp. 416–423.

[42] Y. Rabinovich and A. Wigderson, "Techniques for bounding the convergence rate of genetic algorithms," *Random Struct. Alg.*, vol. 14, no. 2, pp. 111–138, 1999.

[43] S. Y. Yuen and B. K. S. Cheung, "Bounds for probability of success of classical genetic algorithm based on Vose-Liepens model," *Les Cahiers du GERAD*, vol. 0711-2440, 2003.

François Destrempes received the B.Sc. degree in mathematics from the Université de Montréal, Montréal, QC, Canada, in 1985, and the M.Sc. and Ph.D. degrees in mathematics from Cornell University, Ithaca, NY, in 1987 and 1990, respectively. He also received a Postgraduate degree in applied computer science in 2000 and the M.Sc. degree in computer science in 2002 from the Université de Montréal, where he is currenlty pursuing the Ph.D. degree in computer science.

He was a Postdoctoral Fellow at the Centre de Recherche Mathématiques (CRM), Université de Montréal, from 1990 to 1992. He has taught mathematics at Concordia University, Montréal; University of Ottawa, Ottawa, ON, Canada; University of Toronto, Toronto, ON; and the University of Alberta, Edmonton, AB, Canada. His current research interests include statistical methods for image segmentation, parameters estimation, detection of contours, localization of shapes, and applications of stochastic optimization to computer vision.

Jean-François Angers received the B.Sc. degree in 1981 and the M.Sc. degree in 1984, both in applied mathematics, from the Université de Sherbrooke, Sherbrooke, QC, Canada, and the Ph.D. degree in statistics from Purdue University, West Lafayette, IN, in 1987.

He was an Assistant Professor of Statistics at the Université de Sherbrooke from 1987 to 1990. He is currently a Full Professor at the Université de Montréal and a member of the Centre de Recherche Mathématique and the Centre de Recherche sur les Transports. His research interests include nonparametric Bayesian functional estimation, hierarchical Bayesian models, and Bayesian robust estimation.

Max Mignotte received the D.E.A. (Postgraduate degree) in digital signal, image, and speech processing from the INPG University, Grenoble, France, in 1993 and the Ph.D. degree in electronics and computer engineering from the University of Bretagne Occidentale (UBO) and the Digital Signal Laboratory (GTS) of the French Naval Academy, in 1998.

He was an INRIA Postdoctoral Fellow at the University of Montréal (DIRO), QC, Canada, from 1998 to 1999. He is currently with DIRO at the Computer Vision and Geometric Modeling Lab as an Assistant Professor (Professeur Adjoint). He is also a member of the Laboratoire de recherche en Imagerie et Orthopédie (LIO), Centre de Recherche du CHUM, Hôpital Notre-Dame) and a researcher at CHUM. His current research interests include statistical methods and Bayesian inference for image segmentation (with hierarchical Markovian, statistical templates, or active contour models), hierarchical models for high-dimensional inverse problems from early vision, parameters estimation, tracking, classification, shape recognition, deconvolution, 3-D reconstruction and restoration problems.