A Stochastic Method for Bayesian Estimation of Hidden Markov Random Field Models With Application to a Color Model

François Destrempes, Max Mignotte, and Jean-François Angers

***Abstract*—We propose a new stochastic algorithm for computing useful Bayesian estimators of hidden Markov random field (HMRF) models that we call exploration/selection/estimation (ESE) procedure. The algorithm is based on an optimization algorithm of O. François, called the exploration/selection (E/S) algorithm. The novelty consists of using the *a posteriori* distribution of the HMRF, as exploration distribution in the E/S algorithm. The ESE procedure computes the estimation of the likelihood parameters and the optimal number of region classes, according to global constraints, as well as the segmentation of the image. In our formulation, the total number of region classes is fixed, but classes are allowed or disallowed dynamically. This framework replaces the mechanism of the split-and-merge of regions that can be used in the context of image segmentation. The procedure is applied to the estimation of a HMRF color model for images, whose likelihood is based on multivariate distributions, with each component following a Beta distribution. Meanwhile, a method for computing the maximum likelihood estimators of Beta distributions is presented. Experimental results performed on 100 natural images are reported. We also include a proof of convergence of the E/S algorithm in the case of nonsymmetric exploration graphs.**

***Index Terms*—Bayesian estimation of hidden Markov random field (HMRF) models, color model, exploration/selection (E/S) algorithm, image segmentation, maximum likelihood (ML) estimation of Beta distributions.**

I. Introduction

ESTIMATION of an image model is an important problem in image processing, with applications to higher level tasks (such as object recognition or three-dimensional reconstruction) and is closely related to image segmentation [1]. Since the pioneer work of [2] and [3], hidden Markov random field (HMRF) models have shown to be useful, if not fundamental, in understanding that problem. HMRF models are sufficiently simple to be algorithmically amenable, although that simplicity might be considered as an over-restrictive hypothesis. However, it is known [4] that (first-order) "HMRF models are dense among essentially all finite-state discrete-time stationary processes and finite-state lattice-based stationary random fields" so that they actually offer a nearly universal structure. The Bayesian paradigm has been widely used in the context of estimation of HMRF models and its richness deserves further study.

Various methods have been developed for segmenting an image based on HMRF models. The simulated annealing (SA) algorithm [5] computes asymptotically [2] the optimal segmentation in the sense of the maximum *a posteriori* (MAP) criterion. However, the temperature-cooling schedule depends on the function to minimize (i.e., the image treated). The iterative conditional mode (ICM) algorithm [3], based on a greedy strategy, usually produces a good suboptimal solution. The modes of posterior marginals (MPM) criterion has been proposed as an alternative to the MAP criterion, with the advantage of being easily computed by a Monte Carlo (MC) algorithm [6]. In the context of hierarchical multiscale (HMS) models, the sequential maximum *a posteriori* (SMAP) criterion has been introduced in order to take into account the interscale relations [7]; a recursive algorithm computes an approximate solution [7]. A multitemperature variant of the SA has been extended to the case of HMS models [8]. Other segmentation methods are based on multiresolution (MR) [9] or multigrid (MG) [10] models.

One fundamental aspect of HMRF models is the unsupervised estimation of the model parameters (i.e., without knowing the segmentation) [1]. The adaptive simulated annealing (ASA) algorithm [11] computes a joint estimation of the likelihood parameters of the HMRF model and segmentation of the image, in the sense of the MAP. However, the solution might be suboptimal [11]. A (suboptimal) estimation of the model parameters and segmentation of the image, in the sense of the maximum likelihood (ML), can be computed jointly using a generalization [12] of the expectation maximization (EM) algorithm [13]. Another approach consists of estimating the HMRF model parameters and then performing the segmentation of the image. Under that point of view, the iterated conditional estimation (ICE) procedure has shown to be relevant in estimating a wide variety of HMRF models [14]–[20], although the statistical estimator that it computes is not fully understood as of now.

In all the methods mentioned above, the number of region classes is assumed to be known. More recently, the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm [21] has been used to perform a joint estimation and segmentation of the HMRF model [22]–[24] in the case where the number of region classes is unknown. In [25], a cooling-temperature schedule is imposed on the RJMCMC stochastic process, in order to compute an optimal solution in the sense of the MAP.

Manuscript received January 8, 2004; revised August 18, 2004. This work was supported by the former Fonds formation chercheurs & aide recherche (FCAR), Quebec, 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, Université de Montréal, Montréal, H3C 3J7 QC Canada.

J.-F. Angers is with the Département de mathématiques et de statistique, Université de Montréal, Montréal, H3C 3J7 QC Canada.

Digital Object Identifier 10.1109/TIP.2005.851710

In [26], the split-and-merge strategy that is exploited in the RJMCMC is incorporated into a hybrid genetic algorithm.

In this paper, we consider a useful family of Bayesian estimators for HMRF models that take into account global constraints in the loss function. We propose a method for computing these estimators that we call exploration/selection/estimation (ESE) procedure. This procedure is an instance of the exploration/selection (E/S) algorithm of O. François [27], with the novelty that the *a posteriori* distribution of the HMRF model is used as exploration distribution. The E/S algorithm is an evolutionary optimization algorithm that belongs to the family of the generalized simulated annealing (GSA) algorithm [28]–[30]. Other GSA algorithms include the simulated annealing (SA) itself [5], a parallel version of the SA [31], and the genetic algorithm of R. Cerf [32], [33]. The internal parameters of the E/S algorithm depend (for all practical purposes) on the diameter of an exploration graph, and *not* on the fitness function itself. This appears to be a major advantage over other GSA algorithms.[1] It follows from O. François' theorem [27] that the ESE procedure converges to an optimal solution independently of the initial solution. The ESE procedure computes not only the estimation of the HMRF likelihood parameters and the segmentation of the image, but also the optimal number of region classes, based on global constraints. In our framework, these tasks can be performed jointly, or in two steps (estimation of the likelihood parameters, followed by a segmentation and an estimation of the number of region classes). We view the total number of classes as fixed, but with the possibility of dynamically allowing or disallowing classes; in contrast, one would usually consider a total number of classes that varies [22]–[26]. Our formulation allows the ESE procedure to find the optimal number of (allowed) classes without resorting (explicitly) to the more sophisticated split-and-merge operators.

To keep this paper in its simplest form, we do not consider hierarchical HMRF models. Rather, we apply the ESE procedure to a new statistical HMRF (mono-scale) model for colors, whose likelihood is modeled on multivariate distributions, with each component following a Beta distribution. Incidentally, we note that the log-likelihood function of a Beta distribution is strictly concave, which justifies the use of the Fletcher–Reeves algorithm in the computation of its ML estimators. This observation can be useful in SAR imagery [19], [34], [35], where Beta distributions are commonly used.[2] Other HMRF color models include: a probabilistic model [36] for various color features, which is segmented in the sense of the MAP by Hopfield neural network optimization; a heuristic probabilistic MR model [37] for dissimilarities of color features (based on thresholds), which is segmented in the sense of the MAP by a MR SA; a Gaussian model [38] for spatial interactions of RGB color features, which is estimated in the sense of the ML, and then segmented by a split-and-merge strategy; a Gaussian model [39] for the Luv features, which is estimated in the sense of the ML, and then segmented in the sense of the MAP by the SA; and a Gaussian model [25] for the Luv features, with variable number of classes, which is jointly estimated and segmented in the sense of the MAP by a RJMCMC with temperature-cooling schedule.

The remaining part of this paper is organized as follows. In Section II, we present the HRMF models considered in this paper and the Bayesian estimators that we study. Also, the E/S algorithm is described in detail, as well as its application to Bayesian estimation (i.e., the ESE procedure). Section II ends with a description of the two-step estimation and segmentation variant of the ESE procedure. In Section III, we apply those concepts to the proposed HMRF color model, with a discussion on the computation of the ML estimators. Experimental results are briefly presented in Section IV.

II. Bayesian Estimation of Constrained HMRF Models

A. Constrained HMRF Models Considered in This Paper

Given an image, G will denote the graph whose nodes s are the pixels of the image with neighborhoods given by the usual 8-neighbors. We consider a couple of random fields (X, Y), where $Y = \{Y_s\}$ represents a random field of (continuous) observations located at the sites s of G, and $X = \{X_s\}$ represents the labeling field (i.e., a hidden discrete random field). Typically, in a standard segmentation formulation, we seek an optimal realization (in the sense of some statistical criterion) of X given an observed realization of Y. For the color model presented in Section III, X_s represents a class of regions in the image with "similar color" and takes its values in a finite set of labels $\Lambda = \{e_1, e_2, \ldots, e_K\}$, whereas Y_s is the YIQ color channels based at the pixel s.

In our context, K represents the maximal number of region classes allowed in the image. In our opinion, it is reasonable to set this upper bound according to the image size; indeed, an exceedingly large number of region classes will result in a poor estimation of the model parameters (to be discussed below), due to too few elements in the sample sets. The problem of estimating the exact number of classes will be handled below.

We consider as usual a likelihood $P(y|x)$ defined by a site-wise product

$$\prod_s P(y_s|x_s) \tag{1}$$

i.e., the components of Y are mutually independent given X and, furthermore, $P(y_s|x) = P(y_s|x_s)$. In a typical application, the local distributions $P(y_s|x_s = e_k)$ belong to a specified family of distributions parametrized by a vector $\Phi_{(k)}$ (for instance, a multivariate Gaussian model). The likelihood of the HMRF is then described completely by the parameter vector $\Phi = (\Phi_{(k)})$, $1 \leq k \leq K$. The dependence of the likelihood distribution on the particular values of the parameters is made explicit by using the notations $P(y_s|x_s = e_k, \Phi_{(k)})$ and $P(y|x, \Phi)$. We assume that the distributions $P(y_s|x_s = e_k, \Phi_{(k)})$ are strictly positive and continuous functions of y_s and $\Phi_{(k)}$.

Now, it might be desirable to have actually less classes than the maximal number allowed. We view this option (for reasons

[1]More precisely, the *critical height* H_1 of the E/S algorithm depends on the fitness function, but the exploration diameter D is a good upper bound. For other GSA algorithms, the known upper bounds are impractical.

[2]Gamma distributions are equally used in SAR imagery. However, our color features are *bounded*, and, hence, we chose a family of distributions with bounded support. In particular, the simpler Gaussian distribution hypothesis would not be sound in our context.

that will be clear later) as omitting certain classes, rather than decreasing the actual number of classes. Thus, we introduce a vector v of K bits, that indicates which classes are allowed, with the obvious constraint that at least one of them is allowed (i.e., $\sum_{k=1}^{K} v_k \geq 1$). In particular, the vector of parameters Φ has a fixed size ($\dim(\Phi) = \sum_{k=1}^{K} \dim(\Phi_{(k)}) \geq K$) in our framework.

We model the prior distribution by a two-dimensional isotropic Potts model with a second-order neighborhood in order to favor homogeneous regions with no privileged orientation; more complex models are available in the literature. We also consider a constraint imposed by the vector v of allowed classes; namely, we say that a segmentation x is *allowed* by v, if all labels e_k appearing in x (i.e. $e_k = x_s$ for some pixel s) satisfy $v_k = 1$. Setting $\chi(e_k, v) = v_k$, it follows that $\prod_s \chi(x_s, v) = 1$ if x is allowed by v, and $\prod_s \chi(x_s, v) = 0$, otherwise. Thus, $P(x)$ is modeled by

$$\frac{1}{Z(\beta, v)} \exp\left\{-\beta \sum_{\langle s,t \rangle} (1 - \delta(x_s, x_t))\right\} \prod_s \chi(x_s, v) \quad (2)$$

where summation is taken over all pairs of neighboring sites, $\delta(\cdot)$ is the Kronecker delta function, $\beta > 0$ is a parameter, and $Z(\beta, v)$ is a normalizing constant equal to

$$\sum_x \exp\left\{-\beta \sum_{\langle s,t \rangle} (1 - \delta(x_s, x_t))\right\} \prod_s \chi(x_s, v). \quad (3)$$

So, the prior model depends on the parameter vector $\Psi = (\beta, v)$, and, again, the dependence of the *prior* on Ψ is made explicit by the notation $P(x|\Psi)$.

Altogether, the joint distribution of the couple of random fields (X, Y) is given by $P(x, y|\Phi, \Psi) = P(y|x, \Phi)P(x|\Psi)$. Note that the exact number L of region classes appears implicitly in Ψ; namely, $L = \sum_{k=1}^{K} v_k$. If ever $L < K$, it is understood that the parameter vectors $\Phi_{(k)}$ corresponding to disallowed classes (i.e, $v_k = 0$) become obsolete in subsequent higher level tasks (such as indexing images or localizing objects). However, they turn out to be useful at the intermediate step of estimation (to be discussed below).

B. Bayesian Estimators of HMRF Models

As mentioned in Section II-A, the joint distribution of the HMRF (X, Y) is completely specified by the vector (Φ, Ψ). We want to estimate jointly X, Φ, and Ψ, according to some statistical criterion.

We formulate the estimation of the parameters in a Bayesian framework. We view $\Theta = (X, \Phi, \Psi)$ as the parameters to be estimated. Would it be only for numerical reasons, we find convenient to assume in the sequel that the parameters Φ and Ψ belong to *bounded* domains. The prior distribution on the parameters is defined by

$$P(\Theta) = P(X|\Psi)\mathcal{U}(\Phi)\mathcal{U}(\Psi) \quad (4)$$

where $\mathcal{U}$ denotes the uniform distribution, and $P(X|\Psi)$ is provided by the HMRF model.

Now, image segmentation is not a well-posed problem; it depends on some criteria that favor an over segmentation, or, on the contrary, a region merging. Thus, we consider an energy function $\rho(x)$ that sets a particular global constraint on the segmentation process. In general, that function might depend on meta parameters, based on the particular application one has in mind (for instance, a probabilistic model of the real scene). In this paper, we consider an energy function based on the "cubic law" for region sizes [40]. Namely, assuming a Poisson model for the objects of the real scene that is captured by the image under orthographic projection [40], the area A of disk-like objects has a density proportional to $1/A^2$ (because the radius r has a density proportional to $1/r^3$). We also want to restrict directly the number of regions in the image. So, we consider the energy function

$$\rho(x) = \omega n |G|^{\frac{1}{2}} + \sum_{i=1}^{n} \left(2 \log(|R_i(x)|) + \log\left(1 - \frac{1}{|G|}\right)\right) \quad (5)$$

where $|R_1(x)|, \ldots, |R_n(x)|$ are the sizes of the n connected regions induced by x, $|G|$ is the size of the image, and ω is a meta parameter ($= 0$ or 1 in our tests). More precisely, $R_1(x), \ldots, R_n(x)$ are the connected components of the graph H, whose vertices are the pixels of the image, and whose edges consist of pairs of 8-neighbors with *same* region label. Now, it is crucial to realize that the value of the partition function $Z(\psi)$ increases at an exponential rate with respect to the number of allowed classes (for a fixed value of β). This combinatorial fact makes obsolete the comparison of the prior $P(x|\Psi)$ for different number of classes: allowing just one class would be optimal. So, the constraint function ρ has to counter balance the term $\log Z(\psi)$ appearing in the Gibbs energy of the prior model.[3] We, thus, consider a loss function defined by

$$E(\Theta, \theta) = 1 - e^{-\rho(X,\Psi)} \delta(X, x)\delta(\Phi, \phi)\delta(\Psi, \psi) \quad (6)$$

where $\rho(X, \Psi) = \rho(X) - \log Z(\Psi)$, δ denotes the Dirac distribution for continuous variables, or the Kronecker symbol for discrete variables, with $\Theta = (X, \Phi, \Psi)$ and $\theta = (x, \phi, \psi)$.[4]

Finally, the likelihood $P(y|\Theta) = P(y|X, \Phi)$ is provided by the HRMF model. We are then interested in the generalized Bayesian estimator defined pointwise by

$$\hat{\theta}(y) = \arg\min_{\theta} \int_{\Theta} E(\Theta, \theta) P(\Theta|y) d\Theta \quad (7)$$

$$= \arg\min_{\theta} \int_{\Theta} E(\Theta, \theta) P(\Theta) P(y|\Theta) d\Theta \quad (8)$$

[3]In statistical mechanical physics, the quantity $\lim_{|G|\to\infty}(\log Z(\psi)/|G|)$ corresponds to the *pressure* of the image lattice under the prior distribution $P(x|\psi)$. See [41] and [42], for instance.

[4]One could also include in the constraint ρ a term corresponding to the Bayesian information criterion (BIC) [43] in order to encourage simpler models (i.e., a smaller number of allowed classes).

since $P(y) = \int_\Theta P(\Theta)P(y|\Theta)d\Theta$ does not depend on θ; henceforth

$$\hat{\theta}(y) = (x_*, \Phi_*, \Psi_*) = \arg \max_{(x,\phi,\psi)} e^{-\rho(x,\psi)} P(x|\psi)P(y|x,\phi) \tag{9}$$

as is readily seen. Note that one could include ρ in the *prior* of the HMRF model and obtain the MAP estimator. However, we prefer not to do so, because this would make the Markovian blanket of each pixel extend to the whole image lattice.[5] At any rate, the proposed loss function yields the *weighted mode* of the posterior distribution of θ. The squared error loss function and the absolute error loss function would give respectively the *weighted mean* and the *weighted median* of the posterior distribution of θ (see [44, Ch. 3]).

Now, let $\hat{\Phi}(x,y)$ be the ML estimator for the complete data (x,y). That is, given a realization x and the observed realization y, let $\hat{\Phi}_{(k)}(x,y)$ be the ML estimator of $\Phi_{(k)}$ on the sample set $\{y_s : x_s = e_k\}$, so that $\hat{\Phi}(x,y) = (\hat{\Phi}_{(k)}(x,y))$. Here, it is understood that $\hat{\Phi}_{(k)}(x,y)$ can have *any* value whenever the class e_k is empty in the segmentation x (i.e., $x_s \neq e_k$ for all pixels s). Following [11], we obtain

$$(x_*, \Psi_*) = \arg \max_{x,\psi} e^{-\rho(x,\psi)} P(x|\psi) P\left(y|x, \hat{\Phi}(x,y)\right) \tag{10}$$

$$\Phi_* = \hat{\Phi}(x_*, y) \tag{11}$$

since, for given values of x and ψ, we have $e^{-\rho(x,\psi)}P(x|\psi)P(y|x,\phi) \leq e^{-\rho(x,\psi)}P(x|\psi)P(y|x,\hat{\Phi}(x,y))$ upon using the independence of the variables Y_s conditional to X.

For simplicity, the prior parameter β is fixed to 1 throughout[6] so that Ψ reduces to the vector of allowed classes v. Thus, in that case, the estimation problem is reduced to the minimization of the fitness function

$$f(x,v) = \rho(x) - \log\left(P\left(y|x,\hat{\Phi}(x,y)\right)\right) + \beta \sum_{\langle s,t \rangle} (1 - \delta(x_s, x_t)) \tag{12}$$

on the set A of all realizations (x,v) for which x is allowed by v.[7] In this context, the SA algorithm [2] is intractable. Also, the ASA algorithm [11] might converge to suboptimal solutions. In this paper, we propose a new variant of the E/S algorithm [27] in order to find an *optimal* solution which we now present.

C. Exploration/Selection (E/S) Optimization Algorithm

The aim of the E/S is to minimize a fitness function f on a finite search space A. It relies on a graph structure $\mathcal{G}$ on A, called the *exploration graph*, which is assumed connected and symmetric. For each element $x \in A$, $N(x)$ denotes the neighborhood of x in the graph $\mathcal{G}$. For each $x \in A$, a *positive* distribution $a(x,\cdot)$ is defined on the neighborhood $N(x)$ of x in the graph $\mathcal{G}$. Given $m \geq 2$, an element $\vec{x} = (x_1, \ldots, x_m)$ of the Cartesian product A^m is called a *population* (of solutions). Given a population $\vec{x} = (x_1, \ldots, x_m)$, $\alpha(\vec{x})$ will denote the current best solution with minimal index: $\alpha(\vec{x}) = x_l$ such that $f(x_k) > f(x_l)$, for $1 \leq k < l$, and $f(x_k) \geq f(x_l)$, for $l < k \leq m$. The algorithm can be stated as follows.

1) **Initialization**: Choose randomly the initial population $\vec{x} = (x_1, x_2, \ldots, x_m)$.
2) Repeat until a stopping criterion is met.
   a) **Updating the current best**: Determine $\alpha(\vec{x})$ from the current population $\vec{x}$, according to the fitness function f.
   b) **Exploration/selection**: For each $l = 1, \ldots, m$, replace with probability p, x_l by $x'_l \in N(x_l)$ according to the distribution $a(x_l, x'_l)$; otherwise, replace x_l by $\alpha(\vec{x})$ (with probability $1-p$). Decrease p.

In [27], at the exploration step, the element x'_l is taken in $N(x_l) \backslash \{\alpha(\vec{x})\}$, but this is unnecessary, as is explained in details in Appendix A.[8]

The probability p is called the *probability of exploration* and depends on a parameter $T > 0$, called the *temperature*. Taking $p_T = \exp(-1/T)$, one has to decrease T to 0 sufficiently slowly and assume that the size m of the population is sufficiently large. Let A_* be the set of global minima of the fitness function f. The following result follows directly from Theorem 2 of [27] and will suffice for our purposes.

Corollary 1 (Corollary to O. François' Theorem 2 [27]): Let D be the diameter of the exploration graph. Then, for any $m > D$, and any $\tau \geq D$

$$\lim_{t \to \infty} \max_{\vec{x}} P\left(X(t) \notin A_* | X(0) = \vec{x}\right) = 0$$

whenever $p(t) = (t+2)^{-1/\tau}$ (i.e., $T(t) = (\tau / \log(t+2))$), where $t \geq 0$ is the iteration.

Proof: See [50, p. 43]. ■

Now, we will actually need a slightly modified version of the E/S algorithm. Let B be an auxiliary finite set. We assume that the exploration distribution depends on an element ϕ of B. So, given $x \in A$ and $\phi \in B$, $a_\phi(x,\cdot)$ is a positive distribution on the neighborhood $N(x)$. The modified E/S algorithm can be stated as follows.

1) **Initialization**: Choose randomly the initial population $\vec{x} = (x_1, x_2, \ldots, x_m)$, and choose by some deterministic rule the initial vector of auxiliary elements $\vec{\phi} = (\phi_1, \ldots, \phi_m)$.

[5] The two formulations are perfectly equivalent, and it is just a matter of taste as to which one is preferred.

[6] See, for instance, [3], [12], and [45]–[48] for estimation methods of the *prior* model. Technically, the estimation of the parameters of the *prior* distribution can be included in our framework, but we choose not to discuss that aspect in this paper.

[7] The term $-\log Z(\psi)$ of the constraint function cancels out with the term $\log Z(\psi)$ of the prior so that only $\rho(x)$ appears explicitly. In particular, the fitness function does not depend on v, once restricted to the case where x is allowed by v, but note that $f(x,v) = \infty$, whenever x is not allowed by v.

[8] However, for the variant [49] of the E/S algorithm, one *has* to take $x'_l \neq \alpha(\vec{x})$.

2) Repeat until a stopping criterion is met.
   a) **Updating the current best**: Determine $\alpha(\vec{x})$ from the current population $\vec{x}$, according to the fitness function f.
   b) **Exploration/selection**: For each $l = 1,\ldots,m$, replace with probability p, x_l by $x'_l \in N(x_l)$ according to the distribution $a_{\phi_l}(x_l,\cdot)$; otherwise, replace x_l by $\alpha(\vec{x})$ (with probability $1-p$).
   c) **Updating**: Modify the auxiliary elements $\phi_1,\ldots,\phi_m$ according to some deterministic rule, based on the current values of $x_1,\ldots,x_m$. Decrease the probability of exploration p.

In Appendix A, we show that all the results of [27] also hold for this modified version of the E/S algorithm. An example of "deterministic rule" for modifying the auxiliary elements, is presented in Section II-D.

D. Exploration/Selection/Estimation (ESE) Procedure

We now present a particular instance of the E/S algorithm in the context of Section II-B. We let $\mathcal{G}$ be the complete graph structure on the search space A of all pairs (x,v) for which x is allowed by v. Thus, $D = 1$, and this would yield a very poor algorithm if the exploration distribution were the uniform distribution. So, one has to design carefully an exploration distribution.

Let the auxiliary set B consists of all elements $\phi = (\phi_{(k)})$ of the form $\phi_{(k)} = \hat{\Phi}_{(k)}(x,y)$ for some x (*depending* on k). A simple possibility for the exploration distribution is the *a posteriori* distribution of the HMRF model itself $a_\phi((x,v),(x',v')) = P(x'|y,\phi,v')\mathcal{U}(v')$, which can be simulated (approximatively) using a few sweeps of the Gibbs sampler. Thus, roughly speaking, the new allowed classes are chosen randomly according to a uniform distribution, and the exploration is concentrated around the modes of the posterior distribution $P(x'|y,\phi,v')$. However, for algorithmic reasons, it seems to us more interesting to replace the uniform distribution by a distribution $\lambda(v'|v)$ that modifies only 1/2 bit on average, and to simulate x' according to the classes allowed by v' for only one sweep

$$a_\phi\left((x,v),(x',v')\right) \approx P(x'|y,\phi,v')\lambda(v'|v). \tag{13}$$

Note that we do not mind whether x is allowed by v', as long as x' is. In our implementation, the dependence of $a_\phi((x,v),\cdot)$ on x holds in the fact that x serves as initialization for one sweep of the Gibbs sampler. Also, the deterministic rule for *modifying* ϕ given x, consists of setting $\phi_{(k)} = \hat{\Phi}_{(k)}(x,y)$, whenever a class k appears in x (i.e., $x_s = e_k$ for some s), and keeping the current value of $\phi_{(k)}$, otherwise. Hence, ϕ is not completely determined by x; this prevents us from dropping the dependence of the distribution $a_\phi((x,v),\cdot)$ on ϕ. In other words, writing $\phi = \hat{\Phi}(x,y)$ might leave out some classes, which would be problematic since we want to simulate *any* currently allowed class. This is the whole point in using the auxiliary set B in the E/S algorithm. Note that the exploration distribution is strictly positive because of the assumption made in Section II-A on $P(x|\Psi)$ and $P(y|x,\Phi)$.

In order to speed up convergence, one can use the K-means algorithm described in [51], rather than a random initialization. Altogether, the E/S algorithm can be outlined as follows in our context. Let $m \geq 2$ and $\tau \geq 1$.

1) **Parameter initialization**: Use the K-means algorithm to obtain a raw segmentation $x^{[0]}$ based on the set of color features $\{y_s\}$ into K classes. Set $x_l^{[0]} = x^{[0]}$, for $1 \leq l \leq m$; set $v_l^{[0]} = \vec{1}$, for $1 \leq l \leq m$, where $\vec{1}$ denotes the vector with all bits equal to 1. The estimate $\phi^{[0]}$ is equal to the ML estimator on the complete data $(x^{[0]},y)$. Set $\phi_l^{[0]} = \phi^{[0]}$, for $1 \leq l \leq m$.
2) Then, $(\vec{x}^{[t+1]},\vec{v}^{[t+1]},\vec{\phi}^{[t+1]})$ is computed recursively from $(\vec{x}^{[t]},\vec{v}^{[t]},\vec{\phi}^{[t]})$ until a stopping criterion is met, as follows.
   a) **Updating the current best**: Determine $\alpha(\vec{x}^{[t]},\vec{v}^{[t]})$ from the current population $(\vec{x}^{[t]},\vec{v}^{[t]})$, using the values of the fitness function $f(x_l^{[t]},v_l^{[t]})$, $1 \leq l \leq m$.
   b) For $l = 1,2,\ldots,m$, explore a solution with probability $p = (t+2)^{-1/\tau}$, or else select the current best.
      i) Exploration: Modify each bit of $v_l^{[t]}$ with probability $1/2K$; if all bits become equal to 0, set one of them (randomly) equal to 1. Let $v_l^{[t+1]}$ be the resulting vector of allowed classes. For one sweep, visit the sites of the image lattice G sequentially. At each site s, draw $x_s = e$ according to the weights

$$P\left(y_s|x_s=e,\phi_l^{[t]}\right)\chi\left(e,v_l^{[t+1]}\right) \times \exp\left\{-\beta\sum_{s'\in N(s)}\left(1-\delta(e,x_{s'})\right)\right\} \tag{14}$$

      where $N(s)$ denotes the set of 8-neighbors of s. Let $x_l^{[t+1]}$ be the resulting segmentation.
      ii) Selection: Let $(x_l^{[t+1]},v_l^{[t+1]}) = \alpha(\vec{x}^{[t]},\vec{v}^{[t]})$.
   c) **Estimation**: Set $\phi_l^{[t+1]} = \hat{\Phi}(x_l^{[t+1]},y)$. It is understood that for each class not appearing in $x_l^{[t+1]}$, the former estimation is kept.

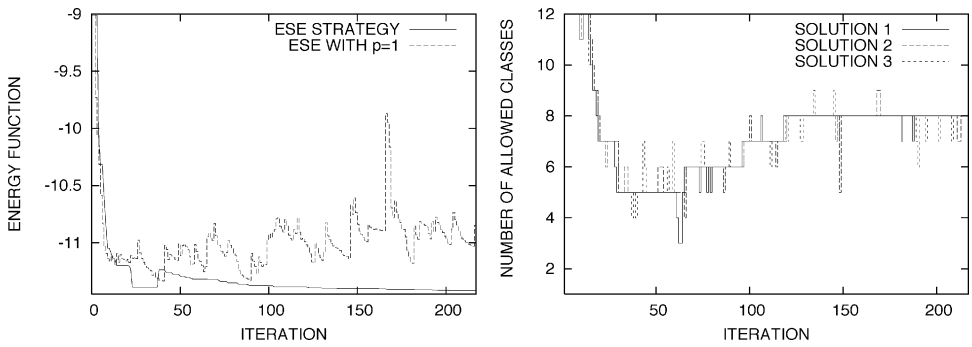

Fig. 1. Left: Example showing the current best value of the fitness function f as a function of the iteration t (the value of the function is normalized by the size of the image); the ESE strategy converges surely to the optimal solution, whereas a simulation-like strategy might take a lot longer before it reaches the optimal solution. Right: Example showing the actual number of allowed classes as a function of the iteration t, for a population of three solutions.

From O. François' theorem, we obtain $\lim_{t\to\infty}(x_l^{[t]}, v_l^{[t]}, \phi_l^{[t]}) = (x_*, v_*, \Phi_*)$, for $1 \leq l \leq m$, with probability 1. The Bayesian estimator sought (x_*, v_*, Φ_*) might not be uniquely defined, but the algorithm will compute one of the optimal solutions. We call this algorithm exploration/selection/estimation (ESE) procedure.

It remains to determine a sensible stopping criterion. The best result known to date in that direction is given by Theorem 3 of [27]. However, the constants R_1 and R_2 appearing there are not known explicitly. Moreover, achieving this task is way beyond the scope of this paper. So, we have decided to fix the final exploration probability empirically. In our tests, we take $m = 3$ and $\tau = 3$, and the final exploration probability is set equal to 1/6. Thus, the procedure is stopped after 217 iterations and an average of 158 explorations are performed. See Fig. 1 for an example showing the current best value of the fitness function f as a function of the iteration t. In that figure, we compare the ESE strategy with a simulation-like strategy, upon setting $p \equiv 1$. Clearly, the ESE procedure seems more promising. Fig. 1 also presents an example of the actual number of allowed classes that were explored, in the case of an upper bound of 12 allowed classes. We note that only 3 to 12 allowed classes were actually explored within the 217 iterations; the minimal Gibbs energy obtained was −9.796 57 and the estimated number of allowed classes was 8. We also performed the estimation procedure with a fixed number of allowed classes varying from 1 to 12. The respective Gibbs energy obtained were: −4.317 26, −5.6363, −9.625 38, −9.800 42, −9.791 46, −9.751 14, −9.742 72, −9.726 39, −9.737 38, −9.741 45, −9.729 46, −9.707. The main point is that an exhaustive search on the number of classes yield a relative improvement of only 0.039% on the Gibbs energy (see Section IV for further discussion).

As seen above, the exploration distribution can be easily simulated using the Gibbs sampler. If ρ were included in the exploration distribution, one would need the MCMC algorithm to simulate the exploration distribution (because, in that case, the Markovian blanket of a pixel would be too large). In that case, the Gibbs sampler could be used to simulate the proposal function, but this is unnecessary in our framework: The acceptance/rejection mechanism of the MCMC is replaced by the exploration/selection mechanism of the E/S.

One could choose the model of variable size for the vector of parameters Φ, corresponding to a variable number of region classes, but, then, one would need the RJMCMC for the simulation of the exploration distribution. In contrast, our framework, based on omission of classes and auxiliary set in the E/S algorithm, allows the use of the Gibbs sampler.

The ESE procedure presents some resemblance with particle filtering (PF) algorithms [52]. One can consider the iteration t as the time, the sequence of estimated parameters (θ_t) as the *signal process*, and the constant sequence $(y_t) = (y)$ as the *observation process*. The exploration distribution $a_\phi((x,v),(x',v'))$ would correspond to the transition kernel of the signal process at consecutive time and the model likelihood to the marginal distribution of the observation conditional to the signal process. The selection step of the ESE would be replaced by the *updating step* (or *resampling*) of the PF. Finally, the estimation step would be replaced by a simulation of the parameters and included in the *prediction step*, together with the exploration step. The main point is that the ESE procedure converges with a fixed number of particles (i.e., solutions) as small as 2, whereas the known convergence results [52] for the PF require that the number of particles tend to infinity.

E. Variants of the ESE Procedure

In the case where the model is very complex, it might be preferable to perform the estimation and the segmentation of the model in two steps. In a first step, the estimation is performed without omitting any class, nor considering any global constraint. In a second step, the segmentation is performed according to the full model, but using the parameters of the likelihood previously estimated. We now give the details.

1) Estimation With No Class Omitted: In order to omit no class, it suffices to consider for the prior model the distribution $P(v) = \delta_{\vec{1}}(v)$, where δ is the Kronecker delta symbol. Moreover, the global constraint is not considered, upon setting $\rho(x) = 0$. This approach amounts to computing the Bayesian estimator

$$(x_*, \Phi_*) = \arg\max_{(x,\phi)} P(x, y|\phi, v = \vec{1}). \tag{15}$$

The ESE procedure is modified accordingly upon letting the search space A consists of all realizations x of the hidden random field X.

2) Segmentation Based on the Likelihood Parameters: Once the vector of parameters Φ_* of the model is estimated, one can estimate once again x_* itself, but using this time the global constraint $\rho(x)$ and permitting the omission of classes. This amounts to computing the Bayesian estimator

$$(x_*, v_*) = \arg\max_{(x,v)} e^{-\rho(x,(\beta,v))} P(x, y|\Phi_*, v). \tag{16}$$

Thus, $P(\Phi) = \delta_{\Phi_*}(\Phi)$. Accordingly, one can modify the ESE procedure upon letting the auxiliary set B consists of the only element Φ_*.

Note that the resulting estimated parameters x_*, Φ_*, v_* are *not* equal to the ones computed in Section II-D. Nevertheless, they also constitute reasonable and (hopefully) useful estimators of the model.

III. Statistical Model for Colors

We now apply the general concepts presented in Section II to an original statistical model for colors. We adopt the same formalism as in Section II-A. Namely, G denotes the image lattice, Y is the observable random field of YIQ color channels on G, and X is the hidden random field of color labels that belong to a finite set Λ of K region classes.

A. Description of the Color Features

The raw data I_s represents the RGB channels at the pixel located at the site s. We compute the YIQ coordinates $y'_s = MI_s$ using the transition matrix [53]

$$M = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{pmatrix}. \tag{17}$$

With the convention that each component of I_s takes its values in $[0, 255] \subset (-1, 256)$, we deduce that $-1 < y'_{s,1} < 256$, $-0.596 \times 257 < y'_{s,2} < 0.596 \times 257$, and $-0.523 \times 257 < y'_{s,3} < 0.523 \times 257$. Based on these bounds, each component of y'_s is normalized between 0 and 1. This yields the transformed data y_s.[9]

B. Statistical Model for the Color Features

For each site s of G, and each color class $e_k \in \Lambda$, we model the distribution $P(y_s|x_s = e_k)$ by a multivariate Beta model, that we now describe. First, we consider the diffeomorphism $\xi : (0,1)^d \to \mathbb{R}^d$ defined by $\tanh^{-1}(2x-1)$ on each component $x \in (0,1)$, where $d = 3$. A few examples convinced us that the variable $\xi(y_s)$ does not quite follow a Gaussian distribution. We chose to model y_s by considering the random vector of dimension d equal to

$$u_s = \pi(y_s) = \xi^{-1}\left(W^t_{(k)}\left(\xi(y_s) - \nu_{(k)}\right) + \nu_{(k)}\right) \tag{18}$$

where $\nu_{(k)}$ is the average d-dimensional vector of the transformed features $\xi(y_s)$, and $W_{(k)}$ is a $d \times d$ orthogonal (decorrelation) matrix for $\xi(y_s)$. Thus, after a suitable rotation, the components of the variable $W^t_{(k)}(\xi(y_s) - \nu_{(k)}) + \nu_{(k)}$ are assumed independent, and the same holds true for the components of u_s.

We model independently each variable $u_{s,r}$ by a Beta distribution $\mathcal{B}(u; a_r, b_r)$, where

$$\mathcal{B}(u; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}, \quad u \in (0,1) \tag{19}$$

with $a, b > 0$. Here, $\Gamma(x) = \int_0^1 (-\log u)^{x-1} du = \int_0^\infty t^{x-1} e^{-t} dt$ is the Euler (1729) gamma function.[10] Now, it does not seem suitable to allow an arbitrarily small value for the standard deviation of the Beta distribution, since one might end up with arbitrarily large values for the shape parameters a, b. Indeed, we have $\sigma^2(\mathcal{B}(a,b)) = (ab)/((a+b)^2(1+a+b))$ and, hence, $\lim_{a\to\infty} \sigma^2(\mathcal{B}(a,a)) = 0$. So, we impose the condition that σ be no less than a fixed value σ_-. This condition implies that a, b are bounded. Thus, our requirement that the likelihood vector of parameters Φ be defined over a bounded domain is fulfilled (see Section II-B).

The values of $\sigma_{-,r}$, for $r = 1, \ldots, d$, are established as follows. We compute $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$, where $\hat{\mu}_r$ is the estimated mean of $u_{s,r}$ over the sample set. We consider the derivative D of the map π evaluated at the point $y_s = \pi^{-1}(\hat{\mu})$, and we set $\sigma_{-,r} = \varepsilon \sum_{r'=1}^d |D_{r,r'}|$, for some fixed-value $\varepsilon > 0$. With that choice, the image of the box $\prod_{r=1}^d [\hat{\mu}_r - \sigma_{-,r}, \hat{\mu}_r + \sigma_{-,r}]$ under the map π^{-1}, *covers* the box centered at $\pi^{-1}(\hat{\mu})$ of radius ε (with respect to the norm $\|\cdot\|_\infty$). Thus, roughly speaking, at least 99% of the distribution of y_s covers for each r an interval of length no less than 6ε. In our tests, we chose $\varepsilon = (1/6 \times 257)$ in order to cover one unit of the RGB channels (on a scale of 0 to 255). Since, the RGB channels actually vary between 0 and 255, rather than -1 and 256 (see Section III-A), the variances obtained are indeed bounded.

Altogether, $P(y_s|x_s = e)$ is modeled by $\prod_{r=1}^d \mathcal{P}_r(u_{s,r})$, where $u_s = \pi(y_s)$ and $\mathcal{P}_r(u_{s,r})$ stands for $\mathcal{B}(u_{s,r}; a_r, b_r)$. See Fig. 2 for an example of empirical distributions for the decorrelated color features.

C. ML Estimators

Let $y_1, y_2, \ldots, y_n$ be a sample of i.i.d. observations drawn according to the multivariate Beta model $(\nu, W, \mathcal{P}_1, \ldots, \mathcal{P}_d)$. The first step in computing the ML estimators of the model is the estimation of the decorrelation operator (ν, W). Here, we use the principal component analysis (PCA) estimators

$$\hat{\nu} = \frac{1}{n}\sum_{l=1}^n \xi(y_l), \quad \hat{W} = U_d \tag{20}$$

where the columns of U_d span the principal subspace of the sample covariance matrix of the sample $\xi(y_l)$ (with corresponding eigenvalues in decreasing order).

Next, the pseudo-decorrelated features $u_l = \pi(y_l)$ are computed. For each fixed index r, we estimate the corresponding Beta distribution, using the method explained in Appendix B.

[9]One could also consider nonlinear transformations of the RGB channels [54], such as the Luv coordinates.

[10]By theorem, the Euler Beta function $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1} du$ is equal to $\Gamma(a)\Gamma(b)/\Gamma(a+b)$.

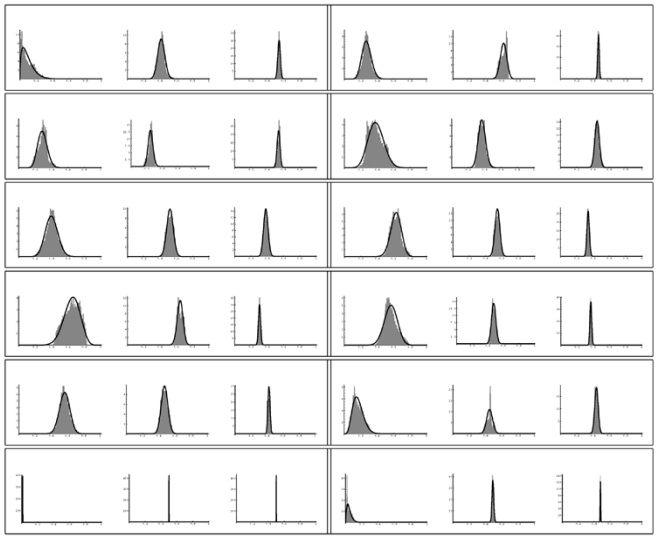

Fig. 2. Example of empirical distributions for the decorrelated color features, based on the segmentation and the parameters estimated by the ESE procedure on the first image of Fig. 4, with a maximum of 12 allowed color classes. Here, we show two classes per line. The histogram of each normalized decorrelated color feature is compared with the corresponding estimated Beta distribution.

D. Estimation and Segmentation Based on the Color Model

Given an image, the statistical model for colors is described completely by the parameter vectors

$$\Phi = \left(\Phi_{(k)}\right) = \left(\nu_{(k)}, W_{(k)}, \mathcal{P}_{(k),1}, \ldots, \mathcal{P}_{(k),d}\right), \quad \Psi = (\beta, v) \tag{21}$$

where $1 \leq k \leq K$. As in Section II-B, we fix $\beta = 1$ throughout, so that Ψ reduces to v. The ESE procedure described in Section II-D is used in order to perform a joint estimation and segmentation. Alternatively, one can use the two-step variant of Section II-E.

E. Simulation of the Color Model

Given a color class e_k, and a statistical parameter vector $\Phi_{(k)} = (\nu_{(k)}, W_{(k)}, \mathcal{P}_{(k),1}, \ldots, \mathcal{P}_{(k),d})$, we proceed as follows to simulate a color region of that class. For each pixel s with label k, simulate each component $u_{s,r}$ according to the given distribution $\mathcal{P}_{(k),r}$, and set $y_s = \pi^{-1}(u_s)$. Then, compute the vector y'_s corresponding to y_s before normalization and set $I_s = M^{-1} y'_s$. This process is repeated until $0 \leq I_{s,r} \leq 255$, for $r = 1, 2, 3$.

IV. Experimental Results

We have tested the proposed method of estimation and segmentation on 100 natural images taken from the database The Big Box of Art. 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 511×768. We have performed two series of tests, with the cubic law of sizes as global constraint.

In the first series of tests, we performed for each natural image I, a joint estimation and segmentation (x_*, Φ_*, v_*), with a maximal number of $K = 12$ allowed classes, and $\omega = 0$ or 1. We then simulated a synthetic image I' based on that estimation. Thus, I' and (x_*, Φ_*, v_*) were considered as ground truth. The RGB channels of that image were saved in floats, rather than in the format ppm, in order to preserve the distributions. Next, we performed a joint estimation and segmentation (x'_*, Φ'_*, v'_*) for the synthetic image, with a maximal number of $K' = 12$ allowed classes. We evaluated the estimation error with the measure

$$\Delta_1 = \frac{|g(x_*, \Phi_*) - g(x'_*, \Phi'_*)|}{|g(x_*, \Phi_*)|} \times 100\% \tag{22}$$

where $g(x, \phi) = \rho(x) - \log(P(y'|x, \phi)) + \beta \sum_{\langle s,t \rangle} \delta(x_s \neq x_t)$, and y' is the observed random field for the synthetic image. See Fig. 3 for a histogram of Δ_1 over the dataset and Fig. 4 for examples of simulated images.

The average number of allowed classes was 11.91 with $\omega = 0$, and 7.18 with $\omega = 1$. This does not necessarily mean that the algorithm failed in finding an optimal reduced number of classes. It could just mean that the optimal number of classes, according to the color model and the global constraint, is not so low. In order to clarify that important point, we performed a second series of tests, with $K = 4$, $K' = 12$, and $\omega = 0$. We compared the two segmentations with the following measure:

$$\Delta_2 = \arg\min_h \frac{1}{|G|} \sum_s \delta\left(x'_*(s) \neq h\left(x_*(s)\right)\right) \times 100\% \tag{23}$$

where h ranges over all one-to-one maps from Λ into Λ'. Thus, that measure represents the classification error, after an optimal match of classes. Δ_2 indicates whether the ESE procedure is capable of estimating the right number of classes, in the difficult situation where the algorithm has to reach four classes, starting with 12 of them. The average number of classes was 5.57, but note that Δ_2 takes into account the proportion occupied by extra classes in the image and had an average value of 0.5%. See Fig. 3 for a histogram of Δ_2 over the dataset.

In the case of synthetic images produced with $K = 4$, we estimated each image with a *fixed* number of four classes. We then compared the optimal Gibbs energy with the one obtained when $K' = 12$. The relative error was only 0.20% on average. Thus, one would not gain much by performing an exhaustive search on the number of classes. The point is that, as in [55], all that matters for higher level tasks, is the Gibbs energy of the model.

Note that specifying the value of ω (i.e., the global constraint) does not amount to fixing the number of allowed classes. Indeed, once the synthetic images are obtained upon setting $K = 12$ or $K = 3$, one obtains an average of 11.91 classes, and 4.89 classes, respectively, with a fixed value of $\omega = 0$. The point is that once the global constraint is fixed, the number of classes found by the proposed model depends on the constraint *and* the observed data. That being said, modifying the global constraint (e.g., taking $\omega = 1$ instead of $\omega = 0$) does affect the number of allowed classes. As in [24], the choice of a global constraint could be guided by a generic model of the image acquisition (e.g., [40]), a statistical criterion (e.g., [43]), or a learning phase performed on a database of images. It would remain to test the

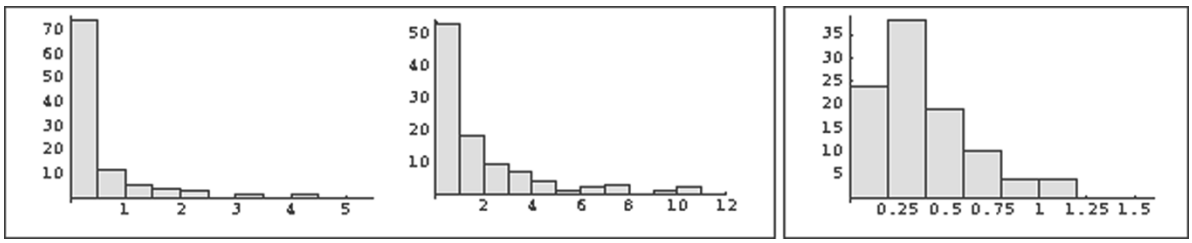

Fig. 3. Histograms of evaluation measures over the dataset. In the usual order. Δ_1 for $K = 12$ and $\omega = 0$; mean: 0.47%. Δ_1 for $K = 12$ and $\omega = 1$; mean: 1.73%. Δ_2 for $K = 4$ and $\omega = 0$; mean: 0.50%.

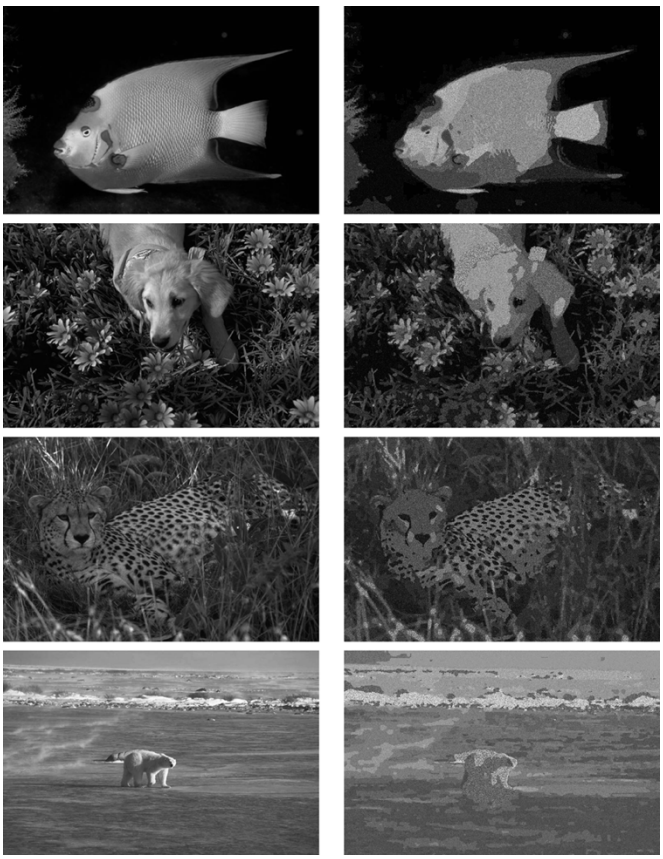

Fig. 4. Unsupervised estimation and segmentation of images using the ESE procedure, according to the multivariate Beta color model. Left: Image. Right: Simulation based on the resulting Bayesian estimators, using the cubic law of region sizes, with $\omega = 0$ and $K = 12$.

robustness of the proposed method with respect to a calibration or estimation of the global constraint parameters.

The ESE procedure is stopped when the exploration probability reaches 1/6 (i.e., after 217 iterations) and takes about 48 min. on a Workstation 1.2 GHz when $K = 12$. This represents an average of no more than 18.5 s/exploration, for a total of 158 explorations. In [24], it is reported that an image of size 350×250 takes from 10 to 30 min *after* some preprocessing step, on a Pentium-III PC. In our case, an image of size 366×250 takes about 10 min. The CPU time is not available for [25], but we know that 200 explorations were performed. Thus, we are inclined to think that the computational complexity of the ESE procedure is equivalent to that of the RJMCMC [24], [25]. Also, the ASA [11] would take the same time per iteration (and might yield a suboptimal solution). The main point is that the known internal parameters of the ESE procedure that ensure convergence are *practical*, whereas for other state-of-the-art algorithms (ASA, RJMCMC with stochastic relaxation), the known bounds are *impractical* (e.g., what should be the initial temperature that would *ensure* convergence?). Furthermore, as mentioned in [11], a joint estimation and segmentation with a plain SA is out of the question, since this would require at each iteration one estimation of the model parameters per pixel for each color label. In our case, this represents about 44 days and 6 h of CPU time per exploration step (i.e., an increment by a factor of about 2.06×10^5). Thus, it seems to us that the computational load of the ESE procedure compares favorably to state-of-the-art algorithms for joint segmentation and estimation, with a clear advantage of having practical optimal internal parameters.

V. Conclusion

The ESE procedure is a general method for estimating the weighted modes of HMRF models, with global constraints taken into account. The optimal internal parameters of the algorithm (i.e., that insure asymptotic convergence to an optimal solution) are known explicitly and are practical. The split-and-merge mechanism is handled implicitly by the procedure, thus yielding a relatively easy implementation. The tests reported in this paper indicate that the ESE procedure succeeds in finding the optimal solution of the proposed color model, within a relative error bound of less than 1.73% on average.

As for the color model itself, it remains to be tested in various higher level tasks, such as indexing or localization of shapes, in combination with models for additional aspects of image analysis. For instance, it is agreed that image segmentation should also include texture analysis and edge detection. See [24] and [56], for instance, but, in this paper, we wanted to test the estimation method on a simple model. Future work will include an extension of the ESE procedure to a hierarchical HMRF model [57], in view of texture segmentation.

Appendix I

The E/S algorithm simulates a in-homogeneous Markov chain on the set A^m, since the temperature depends on the iteration $t \geq 0$. X_t^T will denote the state of the vector $\vec{x}$ at iteration t, where $T = T(t)$. We let q_T be the Markov

transition matrix associated with the chain (X_t^T); i.e., $q_T(\vec{x},\vec{x}') = P(X_{t+1}^T = \vec{x}' | X_t^T = \vec{x})$. We then have

$$q_T(\vec{x},\vec{x}') = \prod_{l=1}^{m} \left(p_T a_{\phi_l(T)}(x_l, x_l') + (1-p_T)\delta(\alpha(\vec{x}), x_l') \right)$$

where δ denotes the Kronecker symbol. Let $I(\vec{x},\vec{x}') = \{l : 1 \le l \le m, x_l' \ne \alpha(\vec{x})\}$. We obtain

$$\prod_{l \in I(\vec{x},\vec{x}')} a_{\phi_l(T)}(x_l, x_l') \, p_T^{|I(\vec{x},\vec{x}')|} (1-p_T)^m \le q_T(\vec{x},\vec{x}') \le \prod_{l \in I(\vec{x},\vec{x}')} a_{\phi_l(T)}(x_l, x_l') \, p_T^{|I(\vec{x},\vec{x}')|}.$$

Let $C > 1$ be a constant such that $(1/C)\mathcal{U}_{N(x)}(x') \le a_\phi(x,x') \le C\mathcal{U}_{N(x)}(x')$, for any $x \in A$, $\phi \in B$, and $x' \in N(x)$, where $\mathcal{U}_{N(x)}$ denotes the uniform distribution on $N(x)$. Such a constant exists because all sets involved are *finite*, and the distributions are *positive* on $N(x)$. Then, we have

$$\frac{1}{\kappa}\pi(\vec{x},\vec{x}') \exp\left(-\frac{V_1(\vec{x},\vec{x}')}{T}\right) \le q_T(\vec{x},\vec{x}') \le \kappa \pi(\vec{x},\vec{x}') \exp\left(-\frac{V_1(\vec{x},\vec{x}')}{T}\right)$$

where

$$\pi(\vec{x},\vec{x}') = \prod_{l \in I(\vec{x},\vec{x}')} \mathcal{U}_{N(x_l)}(x_l')$$

$$V_1(\vec{x},\vec{x}') = \begin{cases} |I(\vec{x},\vec{x}'))|, & \text{if } \pi(\vec{x},\vec{x}') > 0, \\ \infty & \text{otherwise} \end{cases}$$

$$\kappa = \left(1 - p_{T(0)}\right)^{-m} C^m.$$

In particular, π is irreducible (since $\mathcal{G}$ is connected), and the function V_1 is exactly as in [27]. Hence, we are exactly in the same relevant setting as [27], and all the results there apply directly.

We now turn to the case where the exploration graph $\mathcal{G}$ is not necessarily symmetric. We recall from [58] that a $\vec{x}$-graph on A^m consists of a set of arrows $\vec{y} \to \vec{z}$ $(\vec{y},\vec{z} \in A^m, \vec{y} \ne \vec{x},\vec{z})$ such that every point of $A^m \setminus \{\vec{x}\}$ is the initial point of exactly one arrow, and leads to $\vec{x}$ through a sequence of arrows. If $\vec{x} \in A^m$, the set of all $\vec{x}$-graphs is denoted by $G(\vec{x})$. Also, the communication cost from $\vec{x}$ to $\vec{y}$ is defined by

$$V(\vec{x},\vec{y}) = \min\left\{ \sum_{k=0}^{r-1} V_1(\vec{x}_k, \vec{x}_{k+1}) : \vec{x}_0 = \vec{x}, \vec{x}_r = \vec{y}, r \ge 1 \right\}.$$

The virtual energy of $\vec{x}$ is then defined by

$$W(\vec{x}) = \min_{g \in G(\vec{x})} V(g)$$

where

$$V(g) = \sum_{\vec{y} \to \vec{z} \in g} V(\vec{y},\vec{z}).$$

The set of minima of W on A^m is denoted by $\mathcal{W}_*$ and the minimal value by W_*. Let U denote the set $\{\vec{x} \in A^m : x_1 = x_2 = \ldots = x_m\}$. We identify A_* with its natural embedding into U.

The asymptotic behavior of the algorithm is determined by the critical height H_1. We refer the reader to [30] for a detailed definition of this concept, as well as the notion of cycles and exit height of a cycle. If π is a cycle, $H_e(\pi)$ denotes its exit height. H_1 is then defined as $\max_{\pi \cap \mathcal{W}_* = \emptyset} H_e(\pi)$. The importance of the critical height is expressed by the following theorem valid for any GSA.

Theorem 1 (Trouvé [30]): (a) For all decreasing cooling schedules $(T(t))_{t \ge 0}$ converging to 0, we have

$$\lim_{t \to \infty} \sup_x \text{Prob}\left(X(t) \notin \mathcal{W}_* | X(0) = \vec{x}\right) = 0$$

if and only if $\sum_{t=0}^{\infty} \exp(-H_1/T(t)) = \infty$.

The corollary of Section II-C in the case of not necessarily symmetric graphs follows from Trouvé's theorem, upon proving the following two propositions. See [27] for similar results and proofs in the symmetric case.

Proposition 1: If $m > D$, then $\mathcal{W}_* \subset A_*$.

Proposition 2: If $m > D$, then $H_1 \le D$.

Lemma 1: Let $\vec{x} \in A^m$. Then, there exists $g \in G(\vec{x})$ such that $V(g) = W(\vec{x})$ and for all $\vec{y} \to \vec{z} \in g$ either 1) $\vec{y} \in A^m \setminus U$, $\vec{z} \in U$, $V(\vec{y},\vec{z}) = 0$ or 2) $\vec{y} \in U$, $\vec{z} \in U \cup \{\vec{x}\}$.

Proof: This is a special case of [32, Lemma 5.9] with $H = U$, since $V(\vec{x},\alpha(\vec{x})) = 0$ for any $\vec{x} \in A^m$. ∎

Lemma 2: Let $a_* \in A_*$. and $\vec{x} \in A^m$. Then, $V(\vec{x},a_*) \le D$.

Proof: This follows from [49, Lemma 6.1] (with identical proof in the nonsymmetric case), since $V(\vec{x},\alpha(\vec{x})) = 0$. ∎

Lemma 3: Let $\vec{x} \in A^m \setminus U$. Then, $W(\vec{x}) > W(\alpha(\vec{x}))$.

Proof: Let g be an $\vec{x}$ graph as in Lemma 1. There exists $\vec{y} \ne \alpha(\vec{x}) \in U \cup \{\vec{x}\}$ such that $\alpha(\vec{x}) \to \vec{y} \in g$. Remove that edge and introduce the edge $\vec{x} \to \alpha(\vec{x})$. This gives a $\alpha(\vec{x})$ graph so that

$$W(\alpha(\vec{x})) \le V(g) - V(\alpha(\vec{x}),\vec{y}) + V(\vec{x},\alpha(\vec{x})) = W(\vec{x}) - V(\alpha(\vec{x}),\vec{y})$$

but $V(\alpha(\vec{x}),\vec{y}) > 0$, since $\vec{y} \ne \alpha(\vec{x})$. ∎

Proof: (proposition 1): First, consider any $\vec{x}$ with $\alpha(\vec{x}) \in A_*$. Let g be an $\vec{x}$ graph as in Lemma 1. There exists $a_* \in A_*$ and $\vec{y} \in (U \setminus A_*) \cup \{\vec{x}\}$ such that $a_* \to \vec{y} \in g$. Remove that edge and introduce the edge $\vec{x} \to a_*$. This gives a a_* graph.

Now, $V(a_*,\vec{y}) \ge m$, since $\alpha(\vec{y}) \notin A_*$. Moreover, from Lemma 2, $V(\vec{x},a_*) < m$. Hence

$$W(a_*) \le V(g) - V(a_*,\vec{y}) + V(\vec{x},a_*) < W(\vec{x}).$$

Now use Lemma 3. ∎

Lemma 4: For any cycle, $\pi \ne A^m$, $\vec{x} \in \pi$, $\vec{y} \notin \pi$, we have

$$H_e(\pi) + W(\pi) \le W(\vec{x}) + V(\vec{x},\vec{y})$$

where $W(\pi) = \min_{\vec{z} \in \pi} W(\vec{z})$.

Proof: This is established within the proof of [30, Prop. 2.16]. ∎

Proof: (proposition 2): Let $\pi \cap \mathcal{W}_* = \emptyset$ and pick $a_* \in \mathcal{W}_* \subset A_*$, as well as $\vec{x} \in \pi$, such that $W(\pi) = W(\vec{x})$. By Lemma 4, $H_e(\pi) \leq V(\vec{x}, a_*)$. Now use Lemma 2.

APPENDIX II

Lemma 5: Consider the Euler Beta function

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}du.$$

Then, $\log B$ is strictly convex on its domain $D = (0,\infty) \times (0,\infty)$; i.e.,

$$\log B\left(t(a_1,b_1)+(1-t)(a_2,b_2)\right) < t\log B(a_1,b_1) + (1-t)\log B(a_2,b_2)$$

for any $(a_1,b_1) \neq (a_2,b_2)$ in D, and $t \in (0,1)$.

Proof: Fixing $(a_1,b_1) \neq (a_2,b_2)$, we have to show that

$$\log \int_0^1 u^{ta_1+(1-t)a_2-1}(1-u)^{tb_1+(1-t)b_2-1}du < t\log \int_0^1 u^{a_1-1}(1-u)^{b_1-1}du + (1-t)\log \times \int_0^1 u^{a_2-1}(1-u)^{b_2-1}du, \quad t \in (0,1).$$

This inequality amounts to

$$\int_0^1 \left(u^{a_1-1}(1-u)^{b_1-1}\right)^t \left(u^{a_2-1}(1-u)^{b_2-1}\right)^{1-t} du < \left(\int_0^1 u^{a_u-1}(1-u)^{b_1-1}du\right)^t \left(\int_0^1 u^{a_2-1}(1-u)^{b_2-1}du\right)^{1-t}$$

$t \in (0,1)$. Now, write $g(u) = \sqrt{u^{a_1-1}(1-u)^{b_1-1}}$ and $h(u) = \sqrt{u^{a_2-1}(1-u)^{b^2-1}}$. Then, in the case where $t = 1/2$, the inequality reads as

$$\int_0^1 g(u)h(u)du < \left(\int_0^1 g(u)^2du\right)^{\frac{1}{2}} \left(\int_0^1 h(u)^2du\right)^{\frac{1}{2}}$$

but this is just Cauchy–Schwartz inequality, since $h(u)$ is of the form $cg(u)$ only if $(a_1,b_1) = (a_2,b_2)$. Thus, $\log B$ is strictly convex in the Jensen sense (J-convex) on D

$$\log B\left(\frac{a_1+a_2}{2}, \frac{b_1+b_2}{2}\right) \frac{1}{2}\left(\log B(a_1,b_1) + \log B(a_2,b_2)\right)$$

for $(a_1,b_1) \neq (a_2,b_2)$. Since $\log B$ is continuous, we conclude that it is strictly convex. ∎

Let $u_1, u_2, \ldots u_n$ be a sample of i.i.d. observations drawn according to a Beta distribution

$$\mathcal{B}(u;a,b) = \frac{1}{B(a,b)}u^{a-1}(1-u)^{b-1}, \quad u \in (0,1).$$

From [59], the log-likelihood function of the distribution $\mathcal{B}(a,b)$ is given by

$$\mathcal{L}(a,b) = \log\Gamma(a+b) - \log\Gamma(a) - \log\Gamma(b) + (a-1)\lambda_1 + (b-1)\lambda_2$$

where $\lambda_1 = (1/n)\sum_{l=1}^n \log(u_l)$, $\lambda_2 = (1/n)\sum_{l-1}^n \log(1-u_l)$.

Corollary 2: The log-likelihood function of the Beta distribution is strictly concave and has a unique global maximum, which is its unique critical point.

Proof: Since the function $(a-1)\lambda_1 + (b-1)\lambda_2$ is affine, we conclude from the lemma, that $\mathcal{L}(a,b)$ is strictly concave. Furthermore, setting $\mathcal{L}_l(a,b) = -\log(B(a,b)) + (a-1)\log(u_l) + (b-1)\log(1-u_l)$, we have that $\lim_{a\to\infty}\lim_{b\to\infty}\mathcal{L}_l(a,b) = \lim_{a\to 0}\mathcal{L}_l(a,b) = \lim_{b\to 0}\mathcal{L}_l(a,b) = -\infty$. Thus, $\mathcal{L}$ has a global maximum on its domain. Furthermore, using strict concavity, this is the unique critical point of $\mathcal{L}$ on its domain. ∎

Following [59], we obtain

$$\frac{\partial\mathcal{L}}{\partial a} = \psi(1+b) - \psi(a) + \lambda_1, \quad \frac{\partial\mathcal{L}}{\partial b} = \psi(a+b) - \psi(b) + \lambda_2$$

where $\psi(x)$ is the digamma function $\Gamma'(x)/\Gamma(x)$. For an initial approximation of the ML estimators, let

$$\hat{\mu} = \frac{1}{n}\sum_{l=1}^n u_l, \quad \hat{\eta} = \frac{1}{n}\sum_{l=1}^n u_l(1-u_l), \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{l=1}^n (u_l-\hat{\mu})^2$$

and set $a_0 = \hat{\mu}(\hat{\eta}/\hat{\sigma}^2)$, $b_0 = (1-\hat{\mu})(\hat{\eta}/\hat{\sigma}^2)$. If ever $\hat{\sigma} < \sigma_-$, replace the former by the latter. In [59], it is recommended to use Newton–Raphson's method in order to refine the solution, but, by the corollary, it is more appropriate to use a method such as Fletcher–Reeves algorithm for the optimization of the log-likelihood function $\mathcal{L}$. Using *strict concavity*, this algorithm will converge to the optimal solution, even if the initial solution is somewhat far from the optimal one. This gives us the estimated Beta distribution $\mathcal{B}(\hat{a},\hat{b})$. In our implementation in C++, we use the GNU scientific library of functions for the log-gamma and digamma functions, as well as for the Fletcher–Reeves method (with a tolerance of 10^{-4} as stopping criterion). If ever $ab/((a+b)^2(1+a+b)) < \sigma_-^2$, the procedure is stopped. We admit that this is rather ad hoc, but, in this manner, we avoid working directly with the constraint.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and all of the anonymous reviewers for their comments and questions that helped improve both the technical content and the presentation quality of this paper. In particular, they would like to thank the reviewer who pointed out the omission of the partition function in the loss function in the first draft of this paper, and who mentioned the relevance of other loss functions. They would also like to thank the reviewer who made various suggestions to extend the experimental results section, and who asked to present the link of the proposed method with particle filtering algorithms. Finally, they would like to thank the Associate Editor for mentioning [4].

REFERENCES

[1] P. C. Chen and T. Pavlidis, "Image segmentation as an estimation problem," *Compu. Graph. Image Understand.*, vol. 12, pp. 153–172, 1980.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Jun. 1984.

[3] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. B*, vol. 48, pp. 259–302, 1986.

[4] H. Künsch, S. Geman, and A. Kehagias, "Hidden Markov random fields," *Ann. Appl. Prob.*, vol. 5, pp. 577–602, 1995.

[5] B. Hajek, "Cooling schedule for optimal annealing," *Math. Open Res.*, vol. 13, pp. 311–329, 1988.

[6] J. Maroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computation vision," *J. Amer. Stat. Assoc.*, vol. 82, no. 397, pp. 76–89, 1987.

[7] C. Bouman and M. Shapiro, "A multiscale image model for Bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 3, no. 2, pp. 162–177, Feb. 1994.

[8] Z. Kato, M. Berthod, and J. Zerubia, "A hierarchical Markov random field model and multi-temperature annealing for parallel image classification," *Graph. Mod. Image Process.*, vol. 58, no. 1, pp. 18–37, 1996.

[9] M. G. Bello, "A combined Markov random field and wave-packet transform-based approach for image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 3, pp. 618–633, May 1993.

[10] P. Pérez, "Champs markoviens et anaylse multirésolution de l'image: application à l'analyse du mouvement," Ph.D. dissertation, Inst. Recherche en Informatique et Systèmes Aléatoires (IRISA), Univ. Rennes 1, Rennes, France, 1993.

[11] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 8, pp. 799–813, Aug. 1989.

[12] B. Chalmond, "An iterative Gibbsian technique for reconstruction of m-ary images," *Pattern Recognit.*, vol. 22, no. 6, pp. 747–761, 1989.

[13] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Roy. Stat. Soc.*, pp. 1–38, 1976.

[14] B. Braathen, P. Masson, and W. Pieczynski, "Global and local methods of unsupervised Bayesian segmentation of images," *Mach. Graph. Vis.*, vol. 2, no. 1, pp. 39–52, 1993.

[15] A. Peng and W. Pieczynski, "Adaptive mixture estimation and unsupervised local Bayesian image segmentation," *CVGIP: Graph. Models Image Process.*, vol. 57, no. 5, pp. 389–399, 1995.

[16] H. Caillol, W. Pieczynski, and A. Hilton, "Estimation of fuzzy gaussian mixture and unsupervised statistical image segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 425–440, Mar. 1997.

[17] N. Giordana and W. Pieczynski, "Estimation of generalized multisensor hidden Markov chains and unsupervised image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 465–475, May 1997.

[18] F. Salzenstein and W. Pieczynski, "Parameter estimation in hidden fuzzy Markov random fields and image segmentation," *CVGIP: Graph. Models Image Process.*, vol. 59, no. 4, pp. 205–220, 1997.

[19] Y. Delignon, A. Marzouki, and W. Pieczynski, "Estimation of generalized mixture and its application in image segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 10, pp. 1364–1375, Oct. 1997.

[20] 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.

[21] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–731, 1995.

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

[23] S. A. Barker and P. J. W. Rayner, "Unsupervised image segmentation using Markov random field models," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. New York: Springer-Verlag, 1997, pp. 165–178.

[24] 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.

[25] Z. Kato, "Bayesian Color Image segmentation using reversible jump Markov chain Monte Carlo," ERCIM, Res. Rep. 01/99-R055, 1999.

[26] C.-T. Li and R. Chiao, "Unsupervised texture segmentation using multiresolution hybrid genetic algorithm," presented at the 10th IEEE Int. Conf. Image Processing, Barcelona, Spain, Sep. 2003.

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

[28] C.-R. Hwang and S.-J. Sheu, "Singular perturbed Markov chains and exact behaviors of simulated annealing process," *J. Theoret. Prob.*, vol. 5, pp. 223–249, 1992.

[29] A. Trouvé, "Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithm," *Ann. Inst. Henri Poincare Prob. Stat.*, vol. 32, pp. 299–348, 1996.

[30] ——, "Cycle decomposition and simulated annealing," *SIAM J. Control Optim.*, vol. 34, no. 3, pp. 966–986, 1996.

[31] ——, "Partially parallel simulated annealing: low and high temperature approach to the invariant measure," presented at the US-French Workshop on Applied Stochastic Analysis, Apr./May 1991.

[32] R. Cerf, "The dynamics of mutation-selection algorithms with large population sizes," *Ann. Inst. Henri Poincare Prob. Stat.*, vol. 32, no. 4, pp. 455–508, 1996.

[33] ——, "A new genetic algorithm," *Ann. Appl. Prob.*, vol. 6, no. 3, pp. 778–817, 1996.

[34] A. L. Maffett and C. C. Wackerman, "The modified beta density function as a model for synthetic aperture radar clutter statistics," *IEEE Trans. Geosci. Remote Sens.*, vol. 29, no. 2, pp. 277–283, Mar. 1991.

[35] Y. Delignon and W. Pieczynski, "Modeling non-Rayleigh speckle distribution in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 6, pp. 1430–1435, Jun. 2002.

[36] M. J. Daily, "Color image segmentation using Markov random fields," presented at the DARPA Image Understanding, 1989.

[37] J. Liu and Y. H. Yang, "Multiresolution color image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 7, pp. 689–700, Jul. 1994.

[38] D. K. Panjwani and G. Healey, "Markov random field models for unsupervised segmentation of textured color images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 939–954, Oct. 1995.

[39] Z. Kato, T. C. Pong, and J. C. M. Lee, "Motion compensated color video classification using Markov random fields," in *Proc. ACCV*, vol. I, R. Chin and T. C. Pong, Eds., Hong Kong, China, Jan. 1998, pp. 738–745.

[40] 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.

[41] R. Griffiths and D. Ruelle, "Strict convexity (continuity) of the pressure in lattice systems," *Commun. Math. Phys.*, vol. 23, pp. 169–175, 1971.

[42] F. Comets, "On consistency of a class of estimators for exponential families of Markov random fields on the lattice," *Ann. Stat.*, vol. 20, no. 1, pp. 455–468, 1992.

[43] G. Schwartz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[44] T. N. Herzog, *Introduction to Credibility Theory*. Winsted, CT: ACTEX, 1994.

[45] J. Besag, "Statistical analysis of nonlattice data," *The Statistician*, vol. 24, pp. 179–195, 1977.

[46] L. Younes, *Parametric Inference for Imperfectly Observed Gibbsian Fields*. New York: Springer-Verlag, 1989, vol. 82, pp. 625–645.

[47] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 1, pp. 39–55, Jan. 1987.

[48] X. Descombes, R. D. Morris, J. Zerubia, and M. Berthod, "Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood," *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 954–962, Jul. 1999.

[49] O. François, "An evolutionary strategy for global minimization and its Markov chain analysis," *IEEE Trans. Evol. Comput.*, vol. 2, no. 3, pp. 77–90, Mar. 1998.

[50] F. Destrempes, "Détection nonsupervisée de contours et localisation de formes à l'aide de modèles statistiques," M.S. thesis, Dept. d'informatique et de recherche opérationnelle (DIRO), Univ. Montréal, Montréal, QC, Canada, Apr. 2002.

[51] S. Banks, *Signal Processing, Image Processing and Pattern Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1990.

[52] D. Crisan and A. Doucet, "A survey of convergence results on particle filtering methods for practitioners," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 736–746, Mar. 2002.

[53] J. D. Foley, A. van Darn, S. K. Feiner, and J. F. Hughes, *Computer Graphics, Principles and Practice*, 2 ed. New York: Addison-Wesley, 1996.

[54] G. Healy and T. O. Binford, "A color metric for computer vision," presented at the DARPA Image Understanding, 1988.
[55] D. Leman and B. Jedynak, "An active testing model for tracking roads in satellite images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 1, pp. 1–14, Jan. 1996.
[56] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[57] F. Destrempes and M. Mignotte, "Unsupervised texture segmentation using a statistical wavelet-based hierarchical multi data model," in *Proc. 10th IEEE Int. Conf. Image Processing*, vol. II, Barcelona, Spain, Sep. 2003, pp. 1053–1056.
[58] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*. New York: Springer-Verlag, 1984.
[59] R. Gnanadesikan, R. S. Pinkhan, and L. P. Hughes, "Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics," *Technometr.*, vol. 9, no. 4, pp. 607–620, 1967.

François Destrempes received the B.Sc. degree in mathematics from 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 and the M.Sc. degree in computer science from Université de Montréal, in 2000 and 2002, respectively, where he is currently pursuing the Ph.D. degree in computer science.

He was a Postdoctoral fellow at Centre de Recherche Mathématiques (CRM), Université de Montréal, from 1990 to 1992. He has taught mathematics at Concordia University, Montréal; the University of Ottawa, Ottawa, ON, Canada; the 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, and the localization of shapes and applications of stochastic optimization to computer vision.

Max Mignotte received the DEA (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), France, and the Digital Signal Laboratory (GTS) of the French Naval Academy in 1998.

He was an INRIA Postdoctoral Fellow at the Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, Montréal, QC, Canada, from 1998 to 1999. He is currently with DIRO at the Computer Vision and Geometric Modeling Laboratory 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, France, 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, parameter estimation, tracking, classification, shape recognition, deconvolution, three-dimensional reconstruction, and restoration problems.

Jean-François Angers received the B.Sc. and M.Sc. degrees in applied mathematics from the Université de Sherbrooke, Sherbrooke, QC, Canada, in 1981 and 1984, respectively, 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 Full Professor at the Université de Montréal, Montréal, QC. He is also member of the Centre de recherche mathématique and the Centre de recherche sur les transports, Université de Montréal. His current research interests include nonparametric Bayesian functional estimation, hierarchical Bayesian models, and Bayesian robust estimation.