PERGAMON

Pattern Recognition 33 (2000) 573–586

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Noniterative manipulation of discrete energy-based models for image analysis

Patrick Pérez*, Annabelle Chardin, Jean-Marc Laferté

IRISA/INRIA, Campus de Beaulieu, F-35042 Rennes Cedex, France

Received 15 March 1999

Abstract

With emphasis on the graph structure of energy-based models devoted to image analysis, we investigate efficient procedures for sampling and inferring. We show that triangulated graphs, whom trees are simple instances of, always support causal models for which noniterative procedures can be devised to minimize the energy, to extract probabilistic descriptions, to sample from corresponding prior and posterior distributions, or to infer from local marginals. The relevance and efficiency of these procedures are illustrated for classification problems. © 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Energy-based models; Independence graph; Causality; Triangulated graphs; Trees; Noniterative procedures

1. Introduction and background

Many issues of image analysis can be modeled and coped with by designing an *energy* function $U(x,y)$ which captures the interaction between the unknown variables $x=(x_i)_i$ to be estimated, and the observed variables – the measurements or data – , $y=(y_j)_j$. A standard, although complex in general, problem is then the minimization of this function with respect to x, y being known. Other intricate issues, such as estimation of internal parameters or validation of selected models, also arise within this framework.

Depending on the number of variables, the nature of the single variable state space (discrete or not), and the properties of the function (convex or not, local or not), various situations with specific types of difficulty occur. The number of variables, which may be extremely large in usual image analysis problems, remains however a generic source of concern.

Such an energetic modeling is encountered in various fields (e.g., statistical physics, multivariate statistics, combinatorial optimization, artificial intelligence). We are here interested in its use in different approaches to image processing problems, such as in Markov random fields (MRFs)-based approaches [1,2] and partial differential equations (PDEs)-based approaches [3].[1] In the former class of approaches x and y are random vectors and the energy function is naturally related to the *joint distribution*[2] through $\mathsf{P}(x,y) \propto \exp\{-U(x,y)\}$. In the latter one, U is a functional of continuous functions x and y which are discretized afterward (e.g., within the function minimization process). In the following we shall only refer to energy functions of a finite number of variables,

* Corresponding author. Tel: + 33-299-847273; fax: + 33-299-847171

E-mail addresses: perez@irisa.fr (P. Pérez), achardin@irisa.fr (A. Chardin)

[1] Many aspects of energetic modeling do not need to be related to any probabilistic framework, as in PDFs-based approaches, although we think it is a good thing to do. Thus, in order to remain quite general, we tried in this paper to emphasis, when possible, the non-probabilistic aspect of issues of interest. However, certain issues are intrinsically probabilistic, such as drawing samples for instance.

[2] We adopt the convention that all the probability masses will be denoted by P(.). We shall refer to them simply as "distributions".

PII: S0031-3203(99)00073-4

while keeping in mind the strong (and sometimes straight-forward) connection to continuous energetic models. We shall also assume that x_i's take values in a finite set Λ.

The first critical step toward energetic modeling obviously relies on the choice of the energy form. A tailor-made parameterized class of functions is generally chosen. An important ingredient of functions usually used, is their decomposition as a sum of simple *interaction potentials* depending just on a few variables. Thus, by specifying very simple local interactions (possibly non-linear and involving variables of different natures) which sum up as an energy function of all variables and parameters, one can define a global model. This local/global duality is behind the flexibility and power of these modeling approaches.

With such a setting, each variable only directly interacts with a few other "neighboring" variables. From a more global point of view, all variables are mutually dependent, but only through the combination of successive local interactions. This key notion of local functional dependencies is naturally captured by defining an *independence graph* associated to U. It is an undirected graph for which i and j are neighbors if x_i and x_j appear within a same local component of the chosen energy decomposition. This graph structure turns out to be a powerful tool to account for important local and global structural properties of the model. As we shall see, in some specific cases it suffices to deduce *causality* properties, thus allowing the design of efficient estimation algorithms. This paper is particularly dedicated to the exploration of this type of situations in case of discrete models (i.e., x_i's take values in a finite set Λ).

After the specification of an energetic model, one deals with the actual use of it for modeling a class of problems and for solving them. At that point, three main general issues may be of interest:

1. *Sampling*: in order to evaluate the statistical properties of the specified energy-based model, one might want to draw samples from the prior and posterior distributions ($\mathsf{P}(x)$ and $\mathsf{P}(x|y)$ respectively) associated to the energy function. It is then a purely probabilistic issue;
2. *Inferring*: one of the primary goals in early vision problems is to infer the "best" estimate of x given y, with respect to a criterion to be devised.
3. *Learning*: also, one has to tune the parameters involved in the definition of U. The estimation of the optimal parameter vector is tricky since the whole energy landscape depends on it. Apart from manual tuning, consistent estimation procedures (e.g., EM-type algorithms [4,5]) exist, but they remain extremely heavy, if practicable.

In general, there is no way to directly draw samples from the prior and posterior distributions. Among other problems, these distributions are known up to proportionality constants (the *partition functions*) whose computation (by summing up exponentials over all possible values of x) is not tractable in general. One has to use iterative Monte Carlo Markov chain (MCMC) methods (where these constant do not appear) to get samples from distributions converging toward the target distribution. This Monte Carlo framework then allows in addition to compute approximations of partition functions or any other expression involving sums over very large set of configurations (like posterior marginals or posterior expectations). However, the overall procedure is computation demanding due to slow convergence.

As for estimation of x in case of discrete model, there exist two standard estimators stemming from Bayesian estimation theory. The *maximum a posteriori* (MAP) estimator which is the most widely used, makes the best estimate of the most probable x given y: $\hat{x} = \arg\max_x \mathsf{P}(x|y) = \arg\min_x U(x, y)$. It corresponds to the global minimizer of the energy function, and its estimation can therefore be seen as a non probabilistic problem, the energy simply being a "cost" function to be minimized. The second estimator is as for itself intrinsically probabilistic. It is the so-called MPM (for *marginal posterior modes*) which defines site-wise estimate as the most probable given y: $\forall i$, $\hat{x}_i = \arg\max_{x_i} \mathsf{P}(x_i|y)$.

The global minimization necessary to get MAP estimate is not possible in general. Various iterative algorithms can be devised to cope with the problem, but they only provide approximate estimates in general (i.e., "local" energy minima). As for MPM estimates, they rely on the computation of the posterior marginals which is not tractable in general. Aforementioned MCMC iterative techniques can provide us with approximations of these marginals.

Problem of parameter estimation is even more complicated. Standard *Expectation–Maximization* (EM)-type iterative methods require the knowledge of prior partition function, as well as local posterior expectations relative to the current parameter fit [4,5]. Both ingredients are out of reach in general, and (Monte Carlo) approximations are once again necessary [6].

It turns out that for most energy-based models suitable for image analysis problems, one has to devise deterministic or stochastic iterative algorithms exploiting the locality of the model. While permitting tractable single-step computations, the locality results in a very slow propagation of information. As a consequence, these iterative procedures may converge very slowly. It is particularly unbearable for stochastic (sampling or minimization) algorithms. This motivates the search for specific models allowing noniterative or efficient handling of the different listed issues.

In this spirit, probabilistic *causal* models have been already thoroughly studied [7–15]. The class of *causal autoregressive fields*, *unilateral* MRFs, *mesh* MRFs, and *mutually compatible* MRFs on bidimensional lattices has thus been introduced. As we shall recall later, these models rely on a probabilistic causality concept captured by the factorization of $\mathsf{P}(x)$ in terms of *causal transition kernels*.

We examine here the causality from a more graphical point of view, in order to identify causal models *at first sight*, based on simple characteristics of the independence graph of the model. We then explain how noniterative two-sweep algorithms can be devised on these nice graph structures whose simplest instances are *trees*. In particular, we present two algorithms for the *exact* computation of MAP and MPM estimates. They are respectively related to Viterbi algorithm [16] and Baum algorithm [17] which both stem from Hidden Markov (chain) Models (HMMs) [18].

On trees, the general setting we intend to develop here is very much related to discrete classification models by Bouman et al. [19] and by Laferté et al. [20,21]. It is also formally similar to Gaussian models on trees designed by Chou et al. in seminal papers [22,23], and which have been applied to various image processing problems (optical flow estimation [24], texture analysis [25], remote sensing [26,27]).

The paper is organized as follows. In Section 2, we define the independence graph structure which can be associated to any energy-based interacting model, and show how the graph is transformed through three basic mechanisms: freezing of variables (i.e., partial conditioning), energy minimizing with respect to some variables (i.e., conditional MAP estimation) and summing over all possible values of some variables (i.e., marginal computation). In Section 3, standard causality is first defined as a probabilistic concept which can be functionally characterized, and then examined from an alternate graph-theoretic point of view. It is shown that certain constraints on the independence graph ensure at once that noniterative computations are reachable. In Section 4, we detail such efficient computations on trees. The relevance and efficiency of the different procedures are then illustrated in Section 5 for classification tasks.

2. Independence graph and Markovian properties

In the coming section and the following one, we choose a general notational setting in which some variables of interest (observed or not) are gathered into a vector $z=(z_i)_{i=1}^n$ associated to an energy function $U(z)$.

2.1. Definition and properties

As we said in the introduction, an important characteristic of the energy function is its usual decomposition as a sum of local terms:

$$U(z)=\sum_{c\in\mathscr{C}} v_c(z_c), \tag{1}$$

where elements of $\mathscr{C}$ are "small" subsets of indices (usually one or two), and the *interaction potential* v_c only depends on $z_c \triangleq (z_i)_{i\in c}$. Equivalently, the joint distribution of z factorizes into a product of positive *factor potentials*:

$$\mathsf{P}(z)\propto\prod_c g_c(z_c), \tag{2}$$

where $g_c \triangleq \exp\{-v_c\}$. The interaction structure such introduced is conveniently captured by a graph [2,28]:

Definition. The *independence graph* associated to energy decomposition $U(z)=\sum_c v_c(z_c)$ is the simple undirected graph $\mathbb{G}=[S,E]$ with vertex set $S=\{1,\dots,n\}$, and edge set E defined as $\{i,j\}\in E \Leftrightarrow \exists c\in\mathscr{C}\colon \{i,j\}\subset c$.

As a consequence of this definition, elements of $\mathscr{C}$ are *cliques* of $\mathbb{G}$ (i.e., subsets on which $\mathbb{G}$ generates *complete subgraphs*). In the following, we will always assume that the energy function is such that its independence graph is *connected*. This graph structure is equivalently characterized by its *neighborhood system* $\mathscr{N}\triangleq\{\mathsf{n}(i)\}_i$ defined as: $i\in\mathsf{n}(j)\Leftrightarrow j\in\mathsf{n}(i)\Leftrightarrow\{i,j\}\in E$. The independence graph will be equivalently denoted as $\mathbb{G}=[S,\mathscr{N}]$. The vertex set $\mathsf{n}(i)$ contains the neighbors of i in $\mathbb{G}$. For practical convenience, the neighborhoods must be small, i.e., $\mathbb{G}$ should be of reduced degree.

Since a same joint distribution can obviously be defined by different energy functions [1,28], and a same energy function can be decomposed in a number of different ways, different independence graphs can be assigned to $\mathsf{P}(z)$. However, a unique *minimal* independence graph can be defined for this distribution [29]: the neighborhood of i in this graph is the intersection of neighborhoods of i in all possible independence graphs of joint distribution $\mathsf{P}(z)$. As a consequence, the key probabilistic information conveyed by an independence graph $\mathbb{G}$ about two vertices, relies on the *absence* of edge between them: this absence will remain in the minimal graph. One can easily show that in this case, random variables z_i and z_j are *independent given all the remaining variables*:

$$\{i,j\}\notin E \Rightarrow \mathsf{P}(z_i,z_j|z_{S-\{i,j\}}) = \mathsf{P}(z_i|z_{S-\{i,j\}})\times\mathsf{P}(z_j|z_{S-\{i,j\}}). \tag{3}$$

For the minimal independence graph, the implication is replaced by an equivalence. This probabilistic statement constitutes the pairwise Markov property. To prove it, it suffices to note that the distribution of z factorizes into a product of two functions, one of which not depending on z_i, and the other one not depending on z_j. More generally, one can prove the following global Markov property [30–32]: If a vertex subset *a separates* two other disjoint subsets b and d in $\mathbb{G}$ (i.e., all chains from $i \in b$ to $j \in d$ intersect a), then random vectors z_b and z_d are independent given z_a: $\mathsf{P}(z_b, z_d|z_a) = \mathsf{P}(z_b|z_a) \times \mathsf{P}(z_d|z_a)$ and $\mathsf{P}(z_b|z_a, z_d) = \mathsf{P}(z_b|z_a)$. The particular case where $b = \{i\}$ and $a = \mathsf{n}(i)$ constitutes the local Markov property according to which:

$$\mathsf{P}(z_i|z_j, j \neq i) = \mathsf{P}(z_i|z_{\mathsf{n}(i)}) \propto \prod_{c: i \in c} g_c(z_c).$$

2.2. Graphical mechanisms

When handling an energy-based model, three mechanisms are extensively used: (i) *freezing variables*: a set of variables is fixed in a given state, either definitively (e.g., the observations), or momentarily for practical convenience (e.g., in case of alternate sampling). The frozen variables then become like parameters of the energy function; (ii) *summing out*: to compute probabilistic quantities or distributions, one has to sum $\exp\{-U(z)\}$ over all possible values of one or several variables which then "disappear" from the model; (iii) *maximizing out*: when dealing with MAP estimation, the global maximization of $\exp\{-U\}$ is often performed through coordinate-wise maximizations (i.e., w.r.t. a few variables at a time).

We now examine the structural transformations (if any) generated on independence graphs by these basic mechanisms.

Freezing – From independence graph definition, it is straightforward to see that the subset of variables z_a (with $a \subset S$) with energy function deduced from U by freezing other variables $z_{\bar a}$ in a given state (with $\bar a \triangleq S - a$) exhibits the subgraph generated by $\mathbb{G}$ on a as an independence graph. From the probabilistic point of view, it is a matter of *conditioning*: this means that an independence graph of z_a given $z_{\bar a}$ is simply obtained from $\mathbb{G}$ by deleting edges with at least one endpoint in $\bar a$ (see Fig. 1a).

Summing and maximizing – It has been shown in [29] that the marginal $\mathsf{P}(z_a)$ for some subset $a \subset S$ has an independence graph in which two sites are neighbors if they are neighbors in $\mathbb{G}$, or if they belong to the neighborhood of a same connected component of $\bar a = S - a$. This results from the summation of $\mathsf{P}(z) = \mathsf{P}(z_a, z_{\bar a})$ with respect to $z_{\bar a}$ which provides the marginal distribution. It turns out that the same graphical property holds in case of maximization of $\mathsf{P}(z_a, z_{\bar a})$ with respect to $z_{\bar a}$. Let us briefly sketch the similar proofs of these two properties.

Let $\{\bar a_k\}_k$ be the connected components of $\bar a$ in $\mathbb{G}$. The neighborhood $\mathsf{n}(\bar a_k) \triangleq \{i \in S - \bar a_k\colon \mathsf{n}(i) \cap \bar a_k \neq \emptyset\}$ of $\bar a_k$ belongs to a and separates $\bar a_k$ from the rest. Consequently, $\mathsf{P}(z)$ factorizes into:

$$\mathsf{P}(z) \propto g_a(z_a) \prod_k g_k(z_{\bar a_k}, z_{\mathsf{n}(\bar a_k)}). \tag{4}$$

It follows that:

$$\sum_{z_{\bar a}} \mathsf{P}(z) \propto g_a(z_a) \times \prod_k \underbrace{\sum_{z_{\bar a_k}} g_k(z_{\bar a_k}, z_{\mathsf{n}(\bar a_k)})}_{\triangleq G_k(z_{\mathsf{n}(\bar a_k)})}, \tag{5}$$

$$\max_{z_{\bar a}} \mathsf{P}(z) \propto g_a(z_a) \times \prod_k \underbrace{\max_{z_{\bar a_k}} g_k(z_{\bar a_k}, z_{\mathsf{n}(\bar a_k)})}_{\triangleq \mathscr{G}_k(z_{\mathsf{n}(\bar a_k)})}. \tag{6}$$

This means that, in both cases, the components of each $z_{\mathsf{n}(\bar a_k)}$ become *in general* mutually dependent through function G_k or $\mathscr{G}_k$. In case $\bar a_k$ reduces to a single site ($\bar a_k = \{i\}$), the neighbors of i become mutually neighboring through summation or maximization of the joint distribution w.r.t. z_i (see Fig. 1b).

This of course remains a *graphical* viewpoint. Depending on the analytical form of the original distribution (even for a fixed graph structure), simplifications may occur either in G_k's or in $\mathscr{G}_k$'s (factorization, or actual dependence of these functions on *less variables*), thus reducing the actual number of appearing edges (if any). Such simplifications occur with causal models, as we shall see. However, it is very unlikely that simplifications simultaneously occur in both G_k's and $\mathscr{G}_k$'s.

Within estimation issues concerned by this paper, the energy function is generally of the following form (or can be rewritten that way): $U(x, y) = \sum_c v_c(x_c) + \sum_i l_i(y_i, x_i)$. This corresponds to *pointwise* measurements, i.e., components of x and y are in one-to-one correspondence within independence graph of (x, y), and with a mild abuse of notation, they are indexed identically, even though associated to *different* vertices of the joint graph (Fig. 2a). From the above description of graphical mechanisms, one concludes that *a priori* distribution $\mathsf{P}(x)$ and *a posteriori* distribution $\mathsf{P}(x|y)$ have a common independence graph, namely the one deduced from energy term $\sum_c v_c(x_c)$. As for the exact form of prior distribution, it is associated to this energy term (i.e., $\mathsf{P}(x) \propto \exp\{-\sum_c v_c(x_c)\}$) only if for all i, $\sum_{y_i} \exp\{-l_i(y_i, x_i)\}$ is a

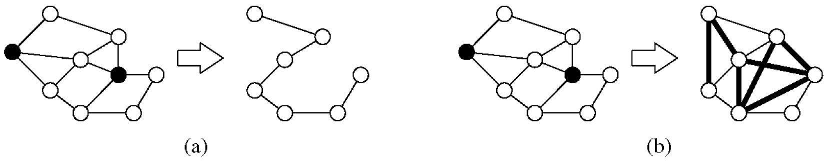


Fig. 1. Graphical consequence of (a) freezing the variables on ● sites and (b) summing or maximizing out the variables on ● sites.

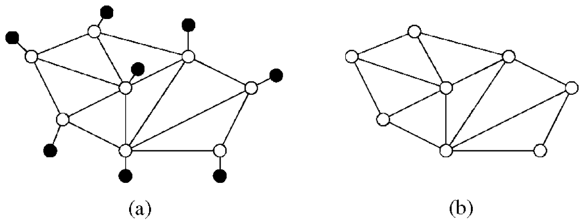


Fig. 2. (a) Example of (x, y) independence graph for pointwise measurements, with x_i's on ○ sites and y_i's on ● sites; (b) the independence graph of x and $x|y$ is obtained by removing ● sites and edges connected to them.

constant. It is always possible to make this assumption hold. In this case $\sum_c v_c$ constitutes the so-called *prior energy*.

3. Causality and graphs

The concept of causality relies on an ordering of sites, and expresses that the conditional distribution of a component z_i given its "past" reduces to the conditional distribution given a "small" neighborhood in the past. To make it precise, one has to introduce a *total ordering* according to which the variables z_i's are now (re)indexed from 1 to n, and seeks the following property [30,31,33]:

$$\forall i > 1, \quad \mathsf{P}(z_i | z_{i-1}, \dots, z_1) = \mathsf{P}(z_i | z_{\tilde{\mathsf{n}}(i)}), \tag{7}$$

where $\tilde{\mathsf{n}}(i)$ is a small subset of i's past $\mathsf{pa}(i) \triangleq \{1, \dots, i-1\}$.[3] If (7) holds, the distribution of $(z_1, \dots, z_k)$, for any k takes the factorized form:

$$\mathsf{P}(z_1, \dots, z_k) = \prod_{i=1}^{k} \mathsf{P}(z_i | z_{\tilde{\mathsf{n}}(i)}), \tag{8}$$

where, for notational convenience, we let $\tilde{\mathsf{n}}(1) = \emptyset$, which means that $z_{\tilde{\mathsf{n}}(1)}$ has to be ignored. There are no unknown normalizing constants within the joint distribution (8), and a noniterative forward recursive sampling of this Markov chain-type distribution can be easily performed.

Given the nice properties offered by causality, it is worth addressing the following issue: the set of sites S being ordered and $U(z) = \sum_c v_c(z_c)$ being an energy function with independence graph $\mathbb{G} = [S, \mathscr{N}]$, could random vector z with distribution $\mathsf{P}(z) \propto \exp\{-U(z)\} \propto \prod_c g_c(z)$ be causal with *small* past neighborhoods? It is not the case in general. As explained in Section 2, for any node i, an independence graph $\mathbb{G}_i$ for marginal distribution $\mathsf{P}(z_1, \dots, z_i)$ can be easily derived from $\mathbb{G}$. In this graph, the neighborhood of i is composed of $\mathsf{n}(i) \cap \mathsf{pa}(i)$ *and* of all sites of $\mathsf{pa}(i)$ that are connected to i in $\mathbb{G}$ through $\{i+1, \dots, n\}$. This neighborhood can be far larger than $\mathsf{n}(i)$ (e.g., if $\mathbb{G}$ is a $M \times N$ grid lexicographically ordered and equipped with the first-order neighborhood system, the previous graphical technique predicts a neighborhood of $M-1$ sites for i in marginal $\mathsf{P}(z_1, \dots, z_i)$, in case site i is away from the border). In this case independence relation (7) *a priori* holds only for a large set $\tilde{\mathsf{n}}(i)$ of predecessors, which makes the causal representation hardly useful.

By successively considering marginals of vectors $z_{\mathsf{pa}(n)}, z_{\mathsf{pa}(n-1)}, \dots, z_1$, one can however establish two cases where causal representation turns out to be at least as local as the original non-causal one. Before we detail them, note that $\prod_c g_c$ can be rearranged as $\prod_i g_i$, where g_i is the product of g_c's for all c containing i and no further site: $g_i \triangleq \prod_{c : \max c = i} g_c$. Then g_i depends only on z_i and on variables attached to the subset $\bigcup_{c : \max c = i} c - \{i\}$ of $\mathsf{n}(i) \cap \mathsf{pa}(i)$. Also, by convention $g_i \equiv 1$ if no clique c of $\mathscr{C}$ verifies $\max c = i$.

3.1. Functional characterization

The marginal independence graph $\mathbb{G}_i$ previously considered is a sort of "upper bound". However, in certain cases depending on the expression of the factors under concern, simplifications might occur in marginal $\mathsf{P}(z_1, \dots, z_i)$, leading to a simpler independence graph. If

$$\forall i, \quad \sum_{z_i} \prod_{c : \max c = i} g_c(z_c) \equiv k_i \text{ (constant)}, \tag{9}$$

it is easy to show, by successive marginalizations, that $\mathsf{P}(z_1, \dots, z_i) \propto \prod_{k=1}^{i} g_k$. The local Markov property indicates that the corresponding independence graph associates $\tilde{\mathsf{n}}(i) \triangleq \bigcup_{c : \max c = i} c - \{i\} \subset \mathsf{n}(i) \cap \mathsf{pa}(i)$ as a neighborhood of i, and:

$$\mathsf{P}(z_i | z_{\mathsf{pa}(i)}) = \mathsf{P}(z_i | z_{\tilde{\mathsf{n}}(i)}) = \frac{\prod_{c : \max c = i} g_c(z_c)}{k_i}.$$

The model then verifies Eq. (7) with past neighborhoods $\tilde{\mathsf{n}}(i) \subset \mathsf{n}(i) \cap \mathsf{pa}(i)$ which are small for $\mathsf{n}(i)$'s are.

This way to introducing causality is at the heart of the various bidimensional causal representations [7,8,9–15,34]. As we already said, this causal probabilistic decomposition allows to recursively draw samples from $\mathsf{P}(z)$, starting from node 1, and all marginals can be exactly computed. However, when Eq. (9) holds for the prior model ($z \equiv x$) (and therefore for the joint model $z \equiv (x, y)$ in case of pointwise measurements), it does not hold in general for the posterior model ($z \equiv x|y$), although prior and posterior independence graphs are the same! This is particularly harmful since posterior model is at the heart of inference and sampling procedures (at least in inverse problems).

[3] "Past" neighborhood system $\tilde{\mathscr{N}} \triangleq \{\tilde{\mathsf{n}}(i)\}_i$ defines an *oriented* independence graph $\tilde{\mathbb{G}} = [S, \tilde{E}]$ on S according to: $(i, j) \in \tilde{E} \Leftrightarrow i \in \tilde{\mathsf{n}}(j)$. See [30,31,33] for more material about directed independence graphs and their semantics.

3.2. Graphical characterization

Graphical considerations will allow to point out an important class of interaction models for which the same conclusion (i.e., causality relative to some in-past parts of the original non-causal neighborhoods) systematically holds, whatever the actual factors are.

To make notations general and simpler to handle in the following, we now note $g_i(z_i, z_{\mathsf{n}(i)\cap\mathsf{pa}(i)})$ even though some of the components of $z_{\mathsf{n}(i)\cap\mathsf{pa}(i)}$ might be absent from the arguments of the function (i.e., if $\bigcup_{c:\max c=i} c - \{i\} \subsetneq \mathsf{n}(i)\cap\mathsf{pa}(i)$).

Return to successive marginalizations from z_n to z_1. As explained in Section 2, the summation of $\prod_i g_i$ w.r.t. z_n makes all sites of $\mathsf{n}(n)\cap\mathsf{pa}(n) = \mathsf{n}(n)$ mutually neighbors through function $G_n(z_{\mathsf{n}(n)}) \triangleq \sum_{z_n} g_n(z_n, z_{\mathsf{n}(n)})$. A particular situation for which this structural change has no incidence is when $\mathsf{n}(n)$ is *already a clique*.

As a consequence random vector $z_{\mathsf{pa}(n)}$ exhibits the subgraph generated by $\mathbb{G}$ on $\mathsf{pa}(n)$ as an independence graph. Its joint distribution is proportional to $\prod_{i=1}^{n-1} g_i \times G_n$. Let $\bar{n} \triangleq \max \mathsf{n}(n)$ be the "greater" vertex of $\mathsf{n}(n)$. Function G_n depends on $z_{\mathsf{n}(n)}$ where $\mathsf{n}(n) \subset (\mathsf{n}(\bar{n})\cap\mathsf{pa}(\bar{n}))\cup\{\bar{n}\}$, since $\mathsf{n}(n)$ is a clique. Therefore G_n and $g_{\bar{n}}$ both depend on $z_{\bar{n}}$ and on some subset of the variables attached to the neighboring predecessors of $\bar{n}$. These two functions can then be "aggregated" in the joint distribution of $z_{\mathsf{pa}(n)}$, to form a single factor function of $(z_{\bar{n}}, z_{\mathsf{n}(\bar{n})\cap\mathsf{pa}(\bar{n})})$. We can start again with reduced vector $z_{\mathsf{pa}(n)}$. By backward induction,[4] one shows that, if

$$\forall i, \quad \mathsf{n}(i)\cap\mathsf{pa}(i) \text{ is a clique of } \mathbb{G}, \tag{10}$$

then any random vector admitting $\mathbb{G}$ as an independence graph is causal relative to oriented neighborhoods defined by $\tilde{\mathsf{n}}(i) \triangleq \mathsf{n}(i)\cap\mathsf{pa}(i)$. Fig. 3 shows a graph that satisfies this condition w.r.t. the indicated site ordering (other orderings were possible). *Any* energy function of six variables admitting this graph as its independence graph defines a causal model relative to $\{\tilde{\mathsf{n}}(i) = \mathsf{n}(i)\cap\mathsf{pa}(i)\}_i$. A nice result shows that the graphs for which an ordering of sites verifying (10) exists are *triangulated* (or *chordal*), i.e., they contain no cycles of length $\geqslant 4$ without a chord. This is obviously the case of graph in Fig. 3. A complete proof can be found in [31].

Let us make more precise the recursion behind the above induction. Functions G_i's are recursively defined for all i as:

$$G_i(z_{\tilde{\mathsf{n}}(i)}) \triangleq \begin{cases} \sum_{z_i} g_i(z_i, z_{\tilde{\mathsf{n}}(i)}) & \text{if } \underline{i} = \emptyset, \\ \sum_{z_i} [g_i(z_i, z_{\tilde{\mathsf{n}}(i)}) \prod_{j\in\underline{i}} G_j(z_{\tilde{\mathsf{n}}(j)})], & \text{otherwise} \end{cases} \tag{11}$$

where $\underline{i} \triangleq \{j \in S: \max \tilde{\mathsf{n}}(j) = i\}$. One can show that Eq. (10) along with the connectedness of $\mathbb{G}$ ensures that $\tilde{\mathsf{n}}(i) \neq \emptyset$, $\forall i > 1$ (therefore $\bar{\imath} \triangleq \max \tilde{\mathsf{n}}(i)$ exists). In this case, it clearly turns out that one deals with a "leaves-to-root" recursion on a tree structure $T_{\mathbb{G}}$ defined as follows: $\forall i > 1$, its parent is $\bar{\imath}$ and its child set is $\underline{i}$. The root is vertex 1. The nodes for which $\underline{i} = \emptyset$ (the first to be considered) are the "leaves" (Fig. 3e). The relevant recursive structure for defining algorithms is not anymore the ordering of sites, but the underlying tree structure which can be defined if (10) holds.

The root prior distribution and transition kernels are then obtained:

$$\mathsf{P}(z_1) = \frac{g_1(z_1)}{G_1} \prod_{j\in\underline{1}} G_j(z_1),$$

$$\mathsf{P}(z_i|z_{\tilde{\mathsf{n}}(i)}) = \frac{g_i(z_i, z_{\tilde{\mathsf{n}}(i)})}{G_i(z_{\tilde{\mathsf{n}}(i)})} \prod_{j\in\underline{i}} G_j(z_{\tilde{\mathsf{n}}(j)}),$$

from which the exact joint distribution is deduced. From Section 2, we know that the completeness of $\mathsf{n}(i)\cap\mathsf{pa}(i)$'s ensures that the maximization counterpart of these derivations exists as well. It yields a noniterative way to find energy minimizers which generalizes chain-based Viterbi minimization algorithm [16]. Also, the upward recursive definition (11) holds both for prior model ($z \equiv x$) and for posterior model ($z \equiv x|y$).[5] In the latter case, it allows either to compute (and sample from) posterior distribution, or to compute (and maximize) local posterior marginals, within a single downward sweep.

The functional characterization of causality is the most general, but necessitates the prior definition of a site ordering and of all transition probabilities (up to multiplicative constants). It therefore relies more on the form of potential than on structural information. In particular, causality of this type for the prior model is not inherited by the posterior model, due to energy modification by data-based terms. Besides, it is strongly related to the ordering originally defined. A new causal representation w.r.t. another ordering of sites is not possible in general.

By contrast, graphical viewpoint allows in some cases to identify at first glance (without need of any computational or probabilistic argument) interaction structures that always support causal models. Hence, noniterative sampling, energy minimization, marginal computation and normalization constant calculation are possible either for prior model or for posterior one if they are respectively defined on triangulated independence graphs. In particular, for any compatible ordering (there are several of them in general), one can get back to the classical causal representation (8) based on transition

[4] In the course of the recursion, other G_i's (if any) such that $\max \mathsf{n}(i)\cap\mathsf{pa}(i) = \bar{n}$ will similarly "aggregate" to $g_{\bar{n}}$.

[5] In these cases, we shall change notation G_i into F_i and $\mathbb{F}_i$, respectively.

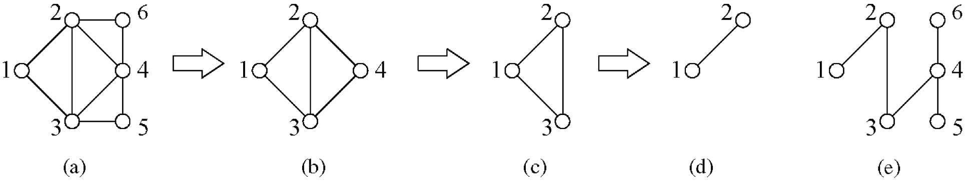


Fig. 3. (a) Example of graph $\mathbb{G} = [S, \mathcal{N}]$ supporting causal energy-based models since in-past neighborhoods $\mathsf{n}(2) \cap \mathsf{pa}(2) = \{1\}$, $\mathsf{n}(3) \cap \mathsf{pa}(3) = \{1,2\}$, $\mathsf{n}(4) \cap \mathsf{pa}(4) = \{2,3\}$, $\mathsf{n}(5) \cap \mathsf{pa}(5) = \{3,4\}$, and $\mathsf{n}(6) \cap \mathsf{pa}(6) = \mathsf{n}(6) = \{2,4\}$ are cliques of $\mathbb{G}$; (b-c-d) successive subgraphs obtained by summing out (or maximizing out) (z_5, z_6), z_4, and z_3 successively; (e) the associated algorithmic tree $T_{\mathbb{G}}$.

kernels, by means of simple noniterative computations. This probabilistic representation which can be derived afterward is used for noniterative sampling and marginal computation/maximization.

In the following, we will focus on particular case of trees. They are triangulated for they do not have cycles by definition. For them, each $\tilde{\mathsf{n}}(i)$ reduces to a singleton with the parent of i, and $\mathbb{G} \equiv T_{\mathbb{G}}$ obviously.

4. Models on trees

Consider an energy-base joint model defined on a tree as:

$$\exp\{-U(x,y)\} = \prod_i f_i(x_i, x_{\bar{\imath}}) h_i(y_i, x_i),$$

with $\sum_{y_i} h_i(y_i, x_i) \equiv m_i$. This means that $\mathsf{P}(y|x) = \prod_i \mathsf{P}(y_i \,|\, x_i) = \prod_i \frac{h_i(y_i,\, x_i)}{m_i}$ and $\mathsf{P}(x) \propto \prod_i f_i$. Recall $\bar{\imath}$ denotes the unique parent of i (with convention $\bar{1} = \emptyset$, and $x_{\bar{1}}$ having to be ignored) and $\underline{i}$ is the set of i's children. Also introduce *ancestor* site set $\bar{\bar{\imath}}$ composed of the sites of the chain between i and 1 (except i itself) and *descendant* site set $\underline{\underline{i}} \triangleq \{j : i \in \bar{\bar{\jmath}}\}$ (see Fig. 4).

As in Eq. (11) (with $z \equiv x$, $\tilde{\mathsf{n}}(i) = \{\bar{\imath}\}$, and appropriate changes of notations) we can recursively define functions F_i's. The causal probabilistic specification of the prior model is then obtained:

$$\mathsf{P}(x_i | x_{\bar{\imath}}) = \frac{f_i(x_i, x_{\bar{1}})}{F_i(x_{\bar{\imath}})} \prod_{j \in \underline{i}} F_j(x_i),$$

and

$$\mathsf{P}(x) = \frac{1}{F_1} \prod_i f_i(x_i, x_{\bar{\imath}}).$$

This allows to draw easily samples from the prior distribution according to a root-to-leaves recursive procedure.

If in addition, $\sum_{x_i} f_i(x_i, x_{\bar{\imath}}) \equiv k_i$, which is usually the case for labeling priors used in detection, segmentation, and classification problems, we turn back to the setting of Section 3.1. This yields the simple causal description $\mathsf{P}(x_i \,|\, x_{\bar{\imath}}) = f_i(x_i, x_{\bar{\imath}})/k_i$. Moreover, if f_i's also verify $f_1 \equiv k'_1$ and $\forall i > 1$, $\sum_{x_{\bar{\imath}}} f_i(x_i, x_{\bar{\imath}}) \equiv k'_i$ (which is also often the case, with $k_i = k'_i$) then it comes that all prior marginals are uniform. In this case, coming derivations are greatly simplified.

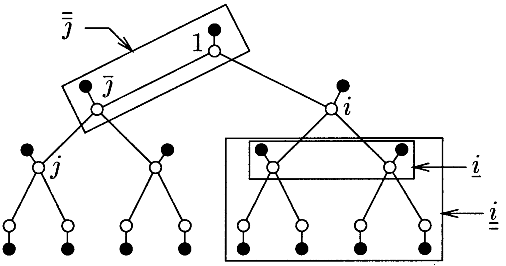


Fig. 4. Independence graph whose prior component is a (dyadic) tree: ○ vertices are for x_i's while ● vertices are for pointwise measurements y_i's.

4.1. *Leaves-to-root maximizations and global energy minimizer*

Using maximization instead of summation in the upward scheme provides a two-sweep Viterbi-like method that minimizes energy $U(x,y)$ w.r.t. x [35]. The maximization counterpart of (11) applied to posterior model provides functions $\mathcal{F}_i$'s which, in this case, "collect" dependencies to more and more data as the recursion proceeds: $\mathcal{F}_i$ not only depends on $x_{\bar{\imath}}$, but also on $(y_i, y_{\underline{\underline{i}}}) \triangleq y_{\underline{\underline{i}}}^{+}$.

The MAP estimate has then to be recovered component by component according to a downward recursion where one has simply to read look-up tables built during the previous sweep (Fig. 5a):

Two-sweep MAP computation on a tree

▲ upward sweep

Leaves

$$\begin{cases} \mathcal{F}_i(x_{\bar{\imath}}, y_i) = \max_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{\imath}}) \\ x_i^*(x_{\bar{\imath}}, y_i) = \arg\max_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{\imath}}) \end{cases}$$

Recursion

$$\begin{cases} \mathscr{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+}) = \max_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{i}}) \prod_{j \in \underline{i}} \mathscr{F}_j(x_i, y_{\underline{\underline{j}}}^{+}) \\ x_i^*(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+}) = \arg\max_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{i}}) \prod_{j \in \underline{i}} \mathscr{F}_j(x_i, y_{\underline{\underline{j}}}^{+}) \end{cases}$$

Root

$$\begin{cases} \mathscr{F}_1(y) = \max_{x_1} h_1(y_1, x_1) f_1(x_1) \prod_{j \in \underline{1}} \mathscr{F}_j(x_1, y_{\underline{\underline{j}}}^{+}) \\ x_1^*(y) = \arg\max_{x_1} h_1(y_1, x_1) f_1(x_1) \prod_{j \in \underline{1}} \mathscr{F}_j(x_1, y_{\underline{\underline{j}}}^{+}) \end{cases}$$

▼ downward sweep

$$\hat{x}_1 = x_1^*(y) \text{ and } \forall i > 1,\ \hat{x}_i = x_i^*(\hat{x}_{\bar{i}}, y_{\underline{\underline{i}}}^{+}).$$

The procedure can equivalently be expressed in terms of interaction potentials $v_i = -\log f_i$ and $l_i = -\log h_i$, products being replaced by sums. Function $-\ln \mathscr{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+})$ is the minimum value for the piece of energy $\sum_{k \in \{i\} \cup \underline{\underline{i}}} l_k + v_k$ for a fixed value of $x_{\bar{i}}$, and the resulting Viterbi-like algorithm provides a global minimizer of U since

$$\begin{aligned} \hat{x}_i &= \arg\min_{x_i} \left[\min_{x_k, k \neq i} U(x, y) \right] \\ &= \arg\min_{x_i} \left[l_i(y_i, x_i) + v_i(x_i, \hat{x}_{\bar{i}}) + \min_{x_{\underline{\underline{i}}}} \sum_{k \in \underline{\underline{i}}} l_k + v_k \right] \\ &= \arg\min_{x_i} [l_i + v_i(x_i, \hat{x}_{\bar{i}}) \\ &\quad + \sum_{j \in \underline{i}} \min_{x_j} \left[v_j + l_j + \min_{x_{\underline{\underline{j}}}} \sum_{k \in \underline{\underline{j}}} l_k + v_k \right] \\ &= \arg\min_{x_i} \left[l_i + v_i(x_i, \hat{x}_{\bar{i}}) + \sum_{j \in \underline{i}} -\log \mathscr{F}_j \right]. \end{aligned}$$

4.2. Leaves-to-root summations given data

In the same spirit as "forward–backward" Baum algorithm on chains [17], a two-sweep procedure can be devised to compute exactly the sitewise posterior marginals (from which the MPM estimate can be deduced). Let us start by introducing $x_{\bar{i}}$ within local posterior marginal $\mathsf{P}(x_i|y)$:

$$\forall i > 1, \quad \mathsf{P}(x_i|y) = \sum_{x_{\bar{i}}} \mathsf{P}(x_i|x_{\bar{i}}, y)\mathsf{P}(x_{\bar{i}}|y), \tag{12}$$

where $\mathsf{P}(x_i|x_{\bar{i}}, y) = \mathsf{P}(x_i|x_{\bar{i}}, y_{\underline{\underline{i}}}^{+})$ due to separation property. This makes appear a downward recursion on site-wise posterior marginal, provided that the posterior marginal at root, $\mathsf{P}(x_1|y)$, and the posterior transition probabilities $\mathsf{P}(x_i|x_{\bar{i}}, y_{\underline{\underline{i}}}^{+})$ are available. These quantities are provided by a previous upward sweep corresponding to successively summing out x_i's from leaves to vertex 1, as in (11) for $z \equiv x|y$ (i.e., $\forall i$, $g_i \equiv h_i f_i$) and $\tilde{\mathsf{n}}(i) = \{\bar{i}\}$, yielding functions $\mathbb{F}_i$'s. The Markov chain-type representation is then obtained as with prior model:

$$\begin{aligned} \mathsf{P}(x_i|x_{\bar{i}}, y) &= \mathsf{P}(x_i|x_{\bar{i}}, y_{\underline{\underline{i}}}^{+}) \\ &= \frac{h_i(y_i, x_i) f_i(x_i, x_{\bar{i}})}{\mathbb{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}^{+}})} \prod_{j \in \underline{\underline{i}}} \mathbb{F}_j(x_i, y_{\underline{\underline{j}}}^{+}), \end{aligned}$$

and

$$\mathsf{P}(x|y) = \frac{1}{\mathbb{F}_1(y)} \prod_i h_i(y_i, x_i) f_i(x_i, x_{\bar{i}}). \tag{13}$$

A noniterative sampling from the posterior distribution can be performed thanks to this probabilistic representation. Also the joint likelihood of data $\mathsf{P}(y)$ is accessible: from $\mathsf{P}(y|x) = \prod_i \frac{h_i(y_i,\, x_i)}{m_i}$, $\mathsf{P}(x) = \frac{1}{F_1} \prod_i f(x_i, x_{\bar{i}})$ and $\mathsf{P}(x|y)$ given above, it comes $\mathsf{P}(y) = \frac{\mathbb{F}_1(y)}{F_1 \times \prod_i m_i}$.

The upward sweep computing $\mathbb{F}_i$'s provides the necessary ingredients for the downward recursion (12). We end up with the following two-sweep procedure (Fig. 5b):

Two-sweep computation of local posterior marginals on a tree

▲ downward sweep

Leaves

$$\mathbb{F}_i(x_{\bar{i}}, y_i) = \sum_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{i}})$$

Recursion

$$\mathbb{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+}) = \sum_{x_i} h_i(y_i, x_i) f_i(x_i, x_{\bar{i}}) \prod_{j \in \underline{i}} \mathbb{F}_j(x_i, y_{\underline{\underline{j}}}^{+})$$

Root

$$\mathbb{F}_1(y) = \sum_{x_1} h_1(y_1, x_1) f_1(x_1) \prod_{j \in \underline{1}} \mathbb{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$$

▼ downward sweep

Initialization

$$\mathsf{P}(x_1|y) = \frac{h_1(y_1, x_1) f_1(x_1)}{\mathbb{F}_1(y)} \prod_{j \in \underline{1}} \mathbb{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$$

Recursion

$$\mathsf{P}(x_i|y) = \sum_{x_{\bar{i}}} \mathsf{P}(x_{\bar{i}}|y) \frac{h_i(y_i, x_i) f_i(x_i, x_{\bar{i}})}{\mathbb{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+})} \prod_{j \in \underline{i}} \mathbb{F}_j(x_i, y_{\underline{\underline{j}}}^{+})$$

Leaves

$$\mathsf{P}(x_i|y) = \sum_{x_{\bar{i}}} \mathsf{P}(x_{\bar{i}}|y) \frac{h_i(y_i, x_i) f_i(x_i, x_{\bar{i}})}{\mathbb{F}_i(x_{\bar{i}}, y_{\underline{\underline{i}}}^{+})}$$

where the MPM estimates are obtained within the top-down part by maximizing the sitewise posterior marginals.[6]

[6] Note that a slightly different, but more complex, procedure can be devised based on the upward propagation of partial posterior marginals $\mathsf{P}(x_i|y_{\underline{\underline{i}}}^{+})$ [21]. It is the exact counterpart of Gaussian inference on a tree based on upward Kalman filtering [22]. Contrary to what we propose here, this method requires however an explicit knowledge of the prior marginals $\mathsf{P}(x_i)$ and of the *child-to-parent* transition probabilities $\mathsf{P}(x_{\bar{i}}|x_i)$.

It is interesting noting that Bouman et al. [19] define a very similar noniterative estimator on tree structure. Starting from an original Bayesian estimator they call "sequential MAP", devised to improve MAP estimates, they obtain a downward recursive *approximation* of it which goes as follows:

$$\hat{x}_1 \approx \arg\max_{x_1} \mathsf{P}(x_1|y),$$

$$\hat{x}_i \approx \arg\max_{x_i} \mathsf{P}(x_i|\hat{x}_{\bar{\imath}}, y_{\underline{\underline{i}}}^{+}) \quad \forall i > 1. \tag{14}$$

At root, their estimator provides the MPM estimate. As for the estimates at other sites, the influence of observations which are not on descendants is simply replaced by the dependency with respect to the parent variable, set at its optimal value already computed. This inference scheme can be plugged into our two-sweep summation procedure to produce an alternate estimator close to the MPM, that we could refer to as "*semi*-MPM". Note that the corresponding (exact) top-down recursive estimation is formally very similar to the one of the MAP estimation (see Fig. 5a): in both cases, the estimate at a site i is obtained by maximizing a function of the estimated value $\hat{x}_{\bar{\imath}}$ on the parent vertex (contrary to MPM estimation), and of the data $y_{\underline{\underline{i}}}^{+}$.

Table 1 gathers in a structured and synthetic way the different two-sweep procedures presented so far, but within the general setting of triangulated graphs (i.e., not necessarily with $\tilde{\mathsf{n}}(i)$ reducing to an unique parent node): it concerns a discrete energy-base model with triangulated independence graph and $\exp\{-U(x,y)\} = \prod_i f_i(x_i, x_{\tilde{\mathsf{n}}(i)}) h_i(y_i, x_i)$, with $\sum_{x_i} h_i(y_i, x_i) = m_i$. Note however that for sake of simplicity, downward recursions (indicated with a black symbol) that require summations over possible values of past neighborhood, i.e., w.r.t. $x_{\tilde{\mathsf{n}}(i)}$, are only written down for a tree, when $\tilde{\mathsf{n}}(i)$ reduces to $\{\bar{\imath}\}$. Apart from providing a practical summary of the different noniterative computations on these models, this table allows to emphasize the profound similarity of the procedures.

5. Experimental results

To demonstrate the practicability and the relevance of the causal models we have presented for low level image

Table 1
Generic upward sweeps (summations on the prior model, summations on the posterior model, and maximization on the posterior model) and downward sweeps (for computing various marginals, sampling from them, and inferring) for discrete energy-base model with triangulated independence graph and $\exp\{-U(x,y)\} = \prod_i f_i(x_i, x_{\tilde{\mathsf{n}}(i)}) h_i(y_i, x_i)$, with $\sum_{x_i} h_i(y_i, x_i) = m_i$

▲Upward sweep▲

leaves	$F_i(x_{\tilde{\mathsf{n}}(i)}) = \sum_{x_i} f_i$	$\mathbb{F}_i(x_{\tilde{\mathsf{n}}(i)}, y_i) = \sum_{x_i} h_i f_i$	$\mathscr{F}_i(x_{\tilde{\mathsf{n}}(i)}, y_i) = \max_{x_i} h_i f_i$
recursion	$F_i(x_{\tilde{\mathsf{n}}(i)}) = \sum_{x_i} f_i \prod_{j \in \underline{i}} F_j(x_{\tilde{\mathsf{n}}(j)})$	$\mathbb{F}_i(x_{\tilde{\mathsf{n}}(i)}, y_{\underline{\underline{i}}}^{+}) = \sum_{x_i} h_i f_i \prod_{j \in \underline{i}} \mathbb{F}_j(x_{\tilde{\mathsf{n}}(j)}, y_{\underline{\underline{j}}}^{+})$	$\mathscr{F}_i(x_{\tilde{\mathsf{n}}(i)}, y_{\underline{\underline{j}}}^{+}) = \max_{x_i} h_i f_i \prod_{j \in \underline{i}} \mathscr{F}_j(x_{\tilde{\mathsf{n}}(j)}, y_{\underline{\underline{i}}}^{+})$
root	$F_1 = \sum_{x_1} f_1 \prod_{j \in \underline{1}} F_j(x_1)$	$\mathbb{F}_1(y) = \sum_{x_1} h_1 f_1 \prod_{j \in \underline{1}} \mathbb{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$	$\mathscr{F}_1(y) = \max_{x_1} h_1 f_1 \prod_{j \in \underline{1}} \mathscr{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$

▼Downward sweep▼

	Prior sampling □	Posterior sampling ◇	Posterior marginals $(\tilde{\mathsf{n}}(i) = \{\bar{\imath}\})$◆	MAP inference
	Prior marginals $(\tilde{\mathsf{n}}(i) = \{\bar{\imath}\})$■	Semi MPM inference ○	MPM inference $(\tilde{\mathsf{n}}(i) = \{\bar{\imath}\})$●	
root	□ ■ $\mathsf{P}(x_1) = \frac{1}{F_1} \prod_{j \in \underline{1}} F_j(x_1)$	◇◆$\mathsf{P}(x_1|y) = \frac{h_1 f_1}{\mathbb{F}_1(y)} \prod_{j \in \underline{1}} \mathbb{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$		
		○●$\hat{x}_1 = \arg\max_{x_1} h_1 f_1 \prod_{j \in \underline{1}} \mathbb{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$		$\hat{x}_1 = \arg\max_{x_1} h_1 f_1 \prod_{j \in \underline{1}} \mathscr{F}_j(x_1, y_{\underline{\underline{j}}}^{+})$
recursion	□ $\mathsf{P}(x_i|x_{\tilde{\mathsf{n}}(i)}) = \frac{f_i}{F_i} \prod_{j \in \underline{i}} F_j$	◇$\mathsf{P}(x_i|x_{\tilde{\mathsf{n}}(i)}, y_{\underline{\underline{i}}}^{+}) = \frac{h_i f_i}{\mathbb{F}_i} \prod_{j \in \underline{i}} \mathbb{F}_j$	◆$\mathsf{P}(x_i|y) = \sum_{x_{\bar{\imath}}} \mathsf{P}(x_{\bar{\imath}}|y) \frac{h_i f_i}{\mathbb{F}_i} \prod_{j \in \underline{i}} \mathbb{F}_j$	
	■ $\mathsf{P}(x_i) = \sum_{x_{\bar{\imath}}} \mathsf{P}(x_{\bar{\imath}}) \frac{f_i}{F_i} \prod_{j \in \underline{i}} F_j$	○ $\hat{x}_i = \arg\max_{x_i} h_i f_i \prod_{j \in \underline{i}} \mathbb{F}_j$ with $x_{\tilde{\mathsf{n}}(i)} = \hat{x}_{\tilde{\mathsf{n}}(i)}$	●$\hat{x}_i = \arg\max_{x_i} \sum_{x_{\bar{\imath}}} \times \mathsf{P}(x_{\bar{\imath}}|y) h_i f_i \prod_{j \in \underline{i}} \mathbb{F}_j$	$\hat{x}_i = \arg\max_{x_i} h_i f_i \prod_{j \in \underline{i}} \mathscr{F}_j$ with $x_{\tilde{\mathsf{n}}(i)} = \hat{x}_{\tilde{\mathsf{n}}(i)}$

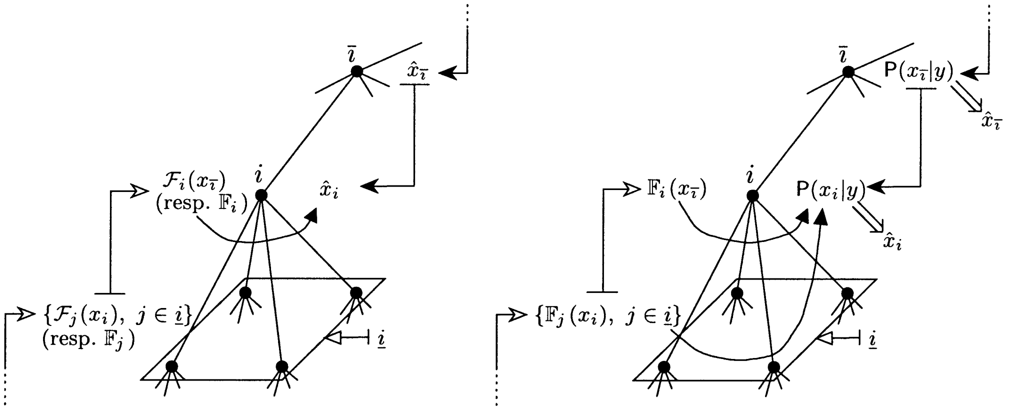


Fig. 5. Downward and backward steps for (a) MAP (resp. semi-mpm), and (b) MPM inference, on a (quad)tree.

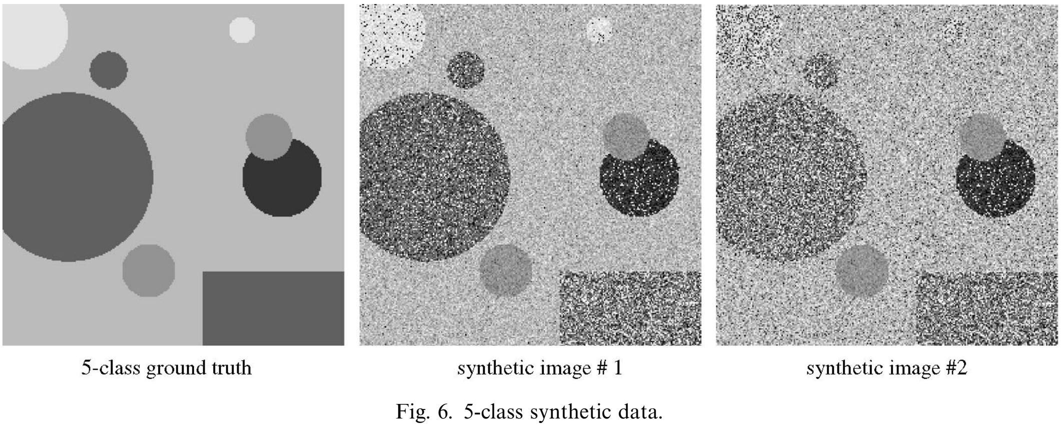
analysis problems, we report experimental results of classification with a model based on the standard quadtree as for its prior independence graph, with the leaves fitting the pixels of the image y to be classified. The energy function is composed of a Potts-like prior term encouraging likeness of children with parents, along with a Gaussian data term:

$$U(x,y) = \sum_{i>1} \beta[1-\delta(x_i, x_{\bar{\imath}})] + \sum_{i:\underline{i}=\emptyset} \frac{(y_i-\mu_{x_i})^2}{2\sigma^2_{x_i}} + \log(\sigma_{x_i}), \qquad (15)$$

where x is a tree labeling with $x_i \in \{1, \dots, M\}$, β is a positive parameter, and $\{(\mu_k, \sigma_k^2)\}_{k=1}^{M}$ are the mean and variances of the M classes.

First experiments were carried out on 256×256 synthetic images involving five classes with known means and variances (Fig. 6). The variances were set to a higher level in the second image (standard deviations range from 15 to 40 in the first image, and from 15 to 70 in the second one). We compared the three noniterative inference procedures on the quadtree, and the iterative ICM algorithm running on the spatial counterpart of energy (15). The obtained classifications are shown in Fig. 7 while Table 2 indicates the corresponding rates of good classification and CPU times in seconds.

On both images, the three noniterative estimators have provided very close classifications which are better than those obtained by iterative estimation with the grid-based model (and noniterative estimations are less degraded that the iterative one for image #2), while taking two to three times less cpu time. Their noniterative nature results in a *fixed computational complexity per site* (e.g., they exhibit an $\mathcal{O}(n)$ complexity). We experimentally determined, using MATLAB implementations, that MAP, semi-MPM, and MPM inferences are achieved with respectively around 79, 94 and 107 floating point operations (flops) per site, when x_i's can take two possible



5-class ground truth

synthetic image # 1

synthetic image #2

Fig. 6. 5-class synthetic data.

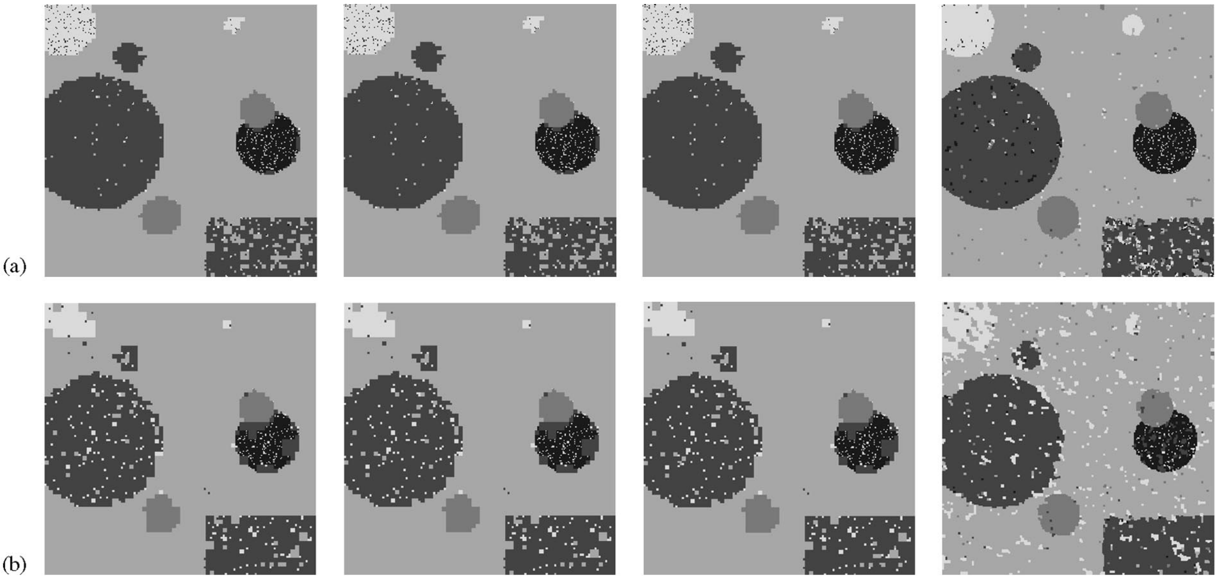


Fig. 7. From left to right: MAP, semi-MPM, and MPM estimates on the quadtree, ICM iterative estimate on the pixel grid, for the classification (a) of synthetic image #1, (b) of synthetic image #2.

Table 2
Comparative percentages of misclassification, and CPU times in seconds on synthetic images

	quadtree			*2d grid*
	MAP	*s*MPM	MPM	ICM
Image #1	4.79% (3.5 s)	4.73% (5.7 s)	4.73% (8.4 s)	5.30% (10.1 s)
Image #2	8.01% (3.5 s)	7.96% (5.7 s)	7.97% (8.4 s)	9.65% (10.5 s)

values. With a similar implementation, standard ICM estimation on bidimensional grid [1] costs around 52 flops/site, whereas the overall procedure is *iterative* with no guarantee on the required number of iterations.

Among the three noniterative estimators, the MPM estimator is the more time consuming due to the larger amount of calculations required in the downward sweep. However, this extra cost (for similar estimates) might worth the pain since the obtained knowledge of posterior marginals $\mathsf{P}(x_i|y)$ allows to assess for each site the degree of *confidence* that can be associated to the estimated value, e.g., through the entropy $-\sum_{x_i}\mathsf{P}(x_i|y)\log\mathsf{P}(x_i|y)$ of the marginal. Fig. 8 shows such "confidence maps". These confidence measures, reminiscent of error covariance matrices of Gaussian models on trees [24], can be useful for a better appreciation and use of obtained estimates.

Visually, the classifications provided by the three noniterative estimators exhibit a "blocky" aspect, reminding the underlying prior quadtree structure. The amount of such artifact depends on the relative location of spatial patterns with respect to the block partition induced on the pixel grid by the quadtree. Also, these artifacts are more apparent in the processing of more noisy images, where the role of quadtree-based prior has to be enforced to get rid of noise. In the prospect of parameter estimation, this is not a serious problem, provided that the overall estimate is good (i.e., the percentage of misclassification is low). However, if the visual rendering of the estimate is at the heart of the concerned application, a *single* ICM smoothing sweep suffices to remove the "blockyness" at reasonable cost.

There is an other source of concern lying in the huge number of successive summations/multiplications usually involved in functions computed through upward sweeps. If no attention is paid to that aspect, one will often end up with quantities either too small, or too large to be handled by computers. To prevent the algorithms from being trapped in these tricky situations, it might be necessary to devise a rescaling of the quantities of interest (namely F_i's, $\mathbb{F}_i$'s, or $\mathcal{F}_i$'s). A simple way to proceed, consists in normalizing these functions such that

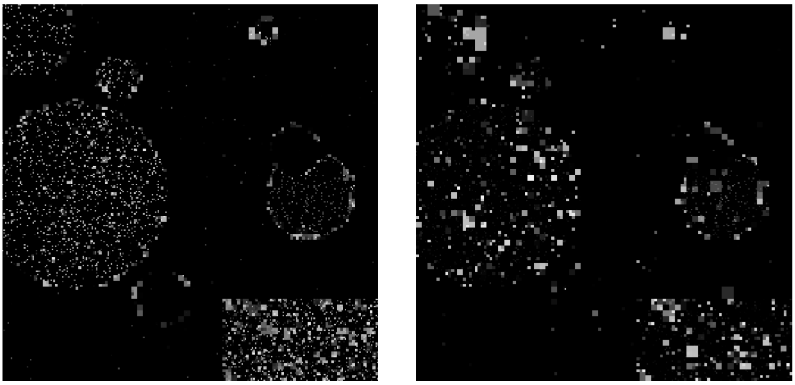


Fig. 8. "Confidence maps" associated to noniterative MPM classification of synthetic images #1 and #2 by the entropy of posterior marginals at leaves (the darker, the less entropy and the higher confidence).

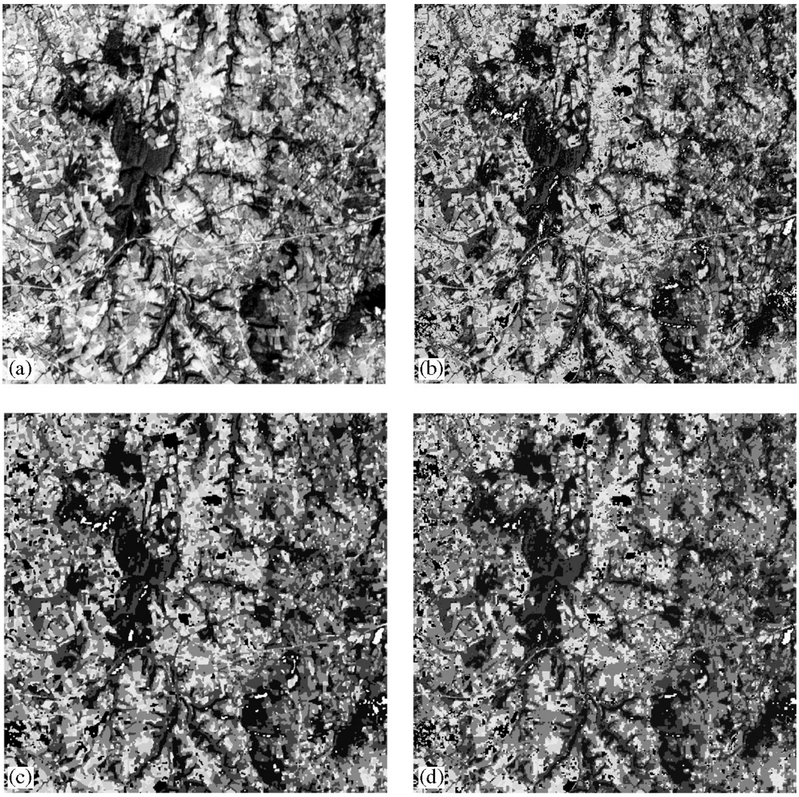


Fig. 9. (a) 512 × 512 Spot image (courtesy of Costel, University of Rennes 2 and GSTB); (b) direct MV classification; (c) ICM iterative classification on the pixel grid; (d) MAP noniterative classification on the quadtree.

summing out $x_{\tilde{n}(i)}$ yields 1. For instance $\mathbb{F}_i = \sum_{x_i} h_i f_i \prod_{j \in \underline{i}} \mathbb{F}_j / (\sum_{x_{n(i)}, x_i} h_i f_i \prod_{j \in \underline{i}} \mathbb{F}_j)$. It is easy to see that these normalizations have no incidence whatsoever on the procedures we have described.

Finally we consider the supervised classification of a 512 × 512 Spot image (Fig. 9a) provided by the Costel laboratory (University of Rennes 2), into 8 classes with physical meanings (mainly the types of culture). Maximum likelihood classification (often used in remote sensing applications) is poor (Fig. 9b), but provides a simple and sensible initial configuration for the iterative grid-based classification whose final result is obtained after 65 s (Fig. 9c). In less time, the three tree-based noniterative estimators have provided close results of good quality. See for instance in Fig. 9d the MAP classification, obtained within 40 s.

6. Conclusion

In this paper, we intended to provide a comprehensive and unified picture of models on "causal graphs". We presented in detail the manipulation of such discrete models, with emphasis on (a) the use of graph theoretic concepts as tools to devise models and get insight into algorithmic procedures; (b) the profound unity which underlies the different procedures whether they compute probabilities, draw samples, or infer estimates.

In particular, we presented three generic exact non-iterative inference algorithms devoted to models exhibiting a triangulated independence graph. The first algorithm allows to compute the MAP estimate (and can be considered apart from any probabilistic framework as performing global energy minimization). The second one, whose aim is intrinsically probabilistic, allows to compute local posterior marginals which can be used to get the MPM estimate or to estimate parameters within an EM-like algorithm [20]. The third one mixes, to some extent, the characteristics of the two others. On simple quadtrees, these two-sweep procedures provide a hierarchical framework suitable for discrete image analysis problems such as detection, segmentation or classification. Apart from providing a lower cost alternative to iterative inference schemes, these tree-based models are good candidates for handling multiresolution data, as advocated in [21,27].

Acknowledgements

Authors gratefully acknowledge their debt to Eric Fabre for enlightening and stimulating discussions.

References

[1] J. Besag, Spatial interaction and the statistical analysis of lattice systems, J. Royal Statist. Soc. B 36 (1974) 192–236.

[2] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Machine Intell. 6 (6) (1984) 721–741.

[3] B. ter Haar Romeny (Ed.), Geometry-driven Diffusion in Computer Vision, Kluwer Academic Publishers, Dordrecht, 1995.

[4] G. Celeux, D. Chauveau, J. Diebolt, On stochastic versions of the EM algorithm, Technical Report 2514, INRIA, March 1995.

[5] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc. B 39 (1977) 1–38, with discussion.

[6] B. Chalmond, An iterative Gibbsian technique for reconstruction of M-ary images, Pattern Recognition 22 (6) (1989) 747–761.

[7] K. Abend, T.J. Harley, L.N. Kanal, Classification of binary random patterns, IEEE Trans. Inform. Theory 11 (1965) 538–544.

[8] H. Derin, P.A. Kelly, Discrete-index Markov-type random processes, Proc. IEEE 77 (10) (1989) 1485–1509.

[9] J. Goutsias, Mutually compatible Gibbs random fields, IEEE Trans. Inform. Theory 35 (6) (1989) 1233–1249.

[10] J. Goutsias, Unilateral approximation of Gibbs random field images, Graph. Mod. Image Proc. 53 (1991) 240–257.

[11] A. Habibi, Two-dimensional Bayesian estimate of images, Proc. IEEE 60 (1972) 878–883.

[12] J. Moura, N. Balram, Recursive structure of noncausal Gauss–Markov random fields, IEEE Trans. Inform. Theory 38 (2) (1992) 335–354.

[13] D. Pickard, A curious binary lattice, J. Appl. Probab. 14 (1977) 717–731.

[14] D. Pickard, Unilateral Markov fields, Adv. Appl. Probab. 12 (1980) 655–671.

[15] J. Woods, C. Radewan, Kalman filtering in two dimensions, IEEE Trans. Inform. Theory 23 (1977) 473–481.

[16] G.D. Forney, The Viterbi algorithm, Proc. IEEE 61 (3) (1973) 268–278.

[17] L. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains, IEEE Ann. Math. Statist. 41 (1970) 164–171.

[18] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–285.

[19] C. Bouman, M. Shapiro, A multiscale image model for Bayesian image segmentation, IEEE Trans. Image Process. 3 (2) (1994) 162–177.

[20] J.-M. Laferté, F. Heitz, P. Pérez, A multiresolution EM algorithm for unsupervised image classification using a quadtree model, in: Proc. Int. Conf. on Pattern Recognition, Vienna, Austria, August 1996.

[21] J.-M. Laferté, F. Heitz, P. Pérez, E. Fabre, Hierarchical statistical models for the fusion of multiresolution image data, in: Proc. Int. Conf. Computer Vision, Cambridge, June 1995.

[22] K. Chou, A. Willsky, A. Benveniste, Multiscale recursive estimation, data fusion, and regularization, IEEE Trans. Autom. Control 39 (3) (1994) 464–477.

[23] K. Chou, A. Willsky, R. Nikoukhah, Multiscale systems, Kalman filters, and Riccati equations, IEEE Trans. Autom. Control 39 (3) (1994) 479–491.

[24] M. Luettgen, W. Karl, A. Willsky, Efficient multiscale regularization with applications to the computation of optical flow, IEEE Trans. Image Process. 3 (1) (1994) 41–64.

[25] M. Luettgen, A. Willsky, Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination, IEEE Trans. Image Processing 4 (2) (1995) 194–207.

[26] P. Fieguth, Application of multiscale estimation to large scale multidimensional imaging and remote sensing problems, Ph.D. thesis, MIT Dept. of EECS, June 1995.

[27] M. Daniel, A.S. Willsky, A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing, Proc. IEEE 85 (1) (1997) 164–180.

[28] R. Kindermann, J.L. Snell, Markov Random Fields and their Applications, vol. 1, Amer. Math. Soc., Providence, RI, 1980.

[29] P. Pérez, F. Heitz, Restriction of a Markov random field on a graph and multiresolution statistical image modeling, IEEE Trans. Inform. Theory 42 (1) (1996) 180–190.

[30] S. Lauritzen, Graphical Models, Oxford Science Publications, Oxford, 1996.

[31] J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley, Chichester, 1990.

[32] J. Woods, Two dimensional discrete Markovian fields, IEEE Trans. Inform. Theory 18 (1972) 232–240.

[33] S. Lauritzen, A. Dawid, B. Larsen, H.-G. Leimer, Independence properties of directed Markov fields, Networks 20 (1990) 491–505.

[34] P.A. Devijver, Real-time modeling of image sequences based on hidden Markov mesh random field models, Technical Report M-307, Philips Research Lab., June 1989.

[35] J.-M. Laferté, F. Heitz, P.Pérez, E. Fabre, Hierarchical statistical models for the fusion of multiresolution data, in: Proc. SPIE Conf. on Neural, Morphological, and Stochastic Methods in Image and Signal Processing, San Diego, USA, July 1995.

About the Author—PATRICK PÉREZ was born in 1968. He graduated from École Centrale Paris, France, in 1990. He received the Ph.D. degree in Signal Processing and Telecom. from the University of Rennes, France, in 1993. He now holds a full-time research position at the Inria center in Rennes. His research interests include statistical and/or hierarchical models for large inverse problems in image analysis.

About the Author—ANNABELLE CHARDIN was born in 1973. She graduated from École Nationale Supérieure de Physique de Marseille. She is completing her Ph.D. degree in Signal Processing and Telecom. from the University of Rennes, France.

About the Author—JEAN-MARC LAFERTÉ was born in 1968. He received the Ph.D. degree in Computer Science from the University of Rennes, France, in 1996. He now holds an assistant professor position at the computer science department of the University of Rennes.