Unsupervised Image Classification with a Hierarchical EM Algorithm

Annabelle Chardin and Patrick Pérez
IRISA-INRIA Rennes
Campus Universitaire de Beaulieu
35042 Rennes Cedex, France
Annabelle.Chardin@irisa.fr, Patrick.Perez@irisa.fr

Abstract

This work takes place in the context of hierarchical stochastic models for the resolution of discrete inverse problems from low level vision. Some of these models lie on the nodes of a quad-tree which leads to non-iterative inference procedures. Nevertheless, if they circumvent the algorithmic drawbacks of grid-based models (computational load and/or great dependance on the initialization), they admit modeling shortcomings (cumbersome and somehow artificial). We investigate a new hierarchical stochastic model which takes benefit from both the spatial and the hierarchical prior modelings. The independance graph is based on a tree which has been pollarded with the nodes at the coarsest resolution exhibiting a grid-based interaction structure. For this class of models, we address the critical problem of parameter estimation. To this end, we derive an EM algorithm on the hybrid structure which mixes an exact EM algorithm on each subtrees and a low cost Gibbsian EM algorithm on the coarse spatial grid. Experiments on a synthetic image and on multispectral satellite images are reported.

1. Introduction and background

Many inverse problems from image analysis can be managed by designing an *energy* function $U(x, y)$ which captures the interaction between a large number of unknown variables $x = (x_i)_i$ to be estimated, and the observed variables –the measurements or data–, $y = (y_j)_j$. The manipulation of this function is made tractable by its usual decomposition as a sum of *local* terms involving just a few variables at a time. This kind of problem is encountered in Markov random field(MRF)-based approaches as well as in partial differential equation (PDE)-based approaches. Within the framework of MRF, x and y are random vectors and we have the following relation between the joint distribution and the energy function: $\mathsf{P}(x, y) \propto \exp\{-U(x, y)\}$.

The decomposition property makes the models very flexible, but implies a parameterization of the posterior distribution which has to be known to perform the inference of x. The crucial point here is the estimation of parameters, since it will strongly condition the quality of the inference. Supervised and unsupervised inference methods have been proposed in the literature. In the supervised approaches, the image and the noise model parameters are assumed to be known, whereas in the unsupervised approaches the parameter estimation and the inference of x are conducted at the same time without any human interaction.

1.1. Hierarchical energy-based models

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 in order to conduct the inference of x. 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. This motivates the search for specific models allowing non-iterative or more efficient inference.

So far, the more fruitful approaches in both cases have relied on some notion of *hierarchy*. Hierarchical models or algorithms allow the information to be integrated in a progressive and efficient way (especially in the case of multiresolution data, when images come into a hierarchy of scales) providing gains in terms of both computational efficiency and quality of results.

Model-based hierarchical approaches aim at defining a new global hierarchical model which has nothing to do with any original (spatial) model. It has to be manipulated as a whole, but according to procedures of reduced complexity. These models usually lie on the nodes of a quad-tree (e.g., see Fig. 1(a)) whose leaves fit the pixels of (maximum resolution) images [2, 6, 10, 11, 12]. In this case, the peculiar dependency structure, like in case of Markov chains, allows *non-iterative* inference procedures made of two sweeps: a bottom-up sweep propagating all information to the root, and a top-down one which in turn allows optimal estimate to be obtained at each node given *all the data*.

One of the drawbacks of these tree-based approaches lies in the structural constraints they impose: first of all they

might appear artificial for certain types of problems or data; in any case the relevance of the inferred variables at coarsest levels is not obvious (especially at the root). Second, the complete tree-structure is cumbersome in case of large images. To circumvent this, a hierarchical model based on a "hybrid" structure which combines a spatial grid of reduced size at a coarser level with "sub-trees" appended below it, down to the finest level has been proposed [4].

1.2. EM algorithm

The so-called EM algorithm is the most used method for parameter estimation. This algorithm [7] considers the observed variables y as the "incomplete data" and the couple (x, y) as the "complete data" characterized by the joint distribution $\mathsf{P}(x, y|\theta)$ where θ is a parameter vector to be estimated. The purpose is to find $\hat{\theta}$ which maximizes the likelihood of observed data $\mathsf{P}(y|\theta)$.

The EM procedure is iterative and repeats the two following steps until convergence: the E-step computes the expectation of log joint likelihood, conditionned on observed data and current parameter fit: $Q(\theta|\theta^{(k)}) \triangleq \mathbb{E}[\log \mathsf{P}(X, Y|\theta)|y, \theta^{(k)}]$; the Maximization step then defines the new parameter values as those that maximize this expectation: $\theta^{(k+1)} \triangleq \arg\max_\theta Q(\theta|\theta^{(k)})$. The convergence is guaranteed, but toward an estimate $\hat{\theta}$ that depends very much on the initial guess $\theta^{(0)}$ [13]. As a consequence $\theta^{(0)}$ must be chosen carefully.

This paper investigates an EM-type algorithm on the hybrid structure introduced in [4] and is organized as follows. The section 2 first describes the hybrid hierarchical model and its associated energy function and secondly the EM algorithm derived from it. The section 3 illustrates this procedure for an unsupervised image classification with synthetic and real images.

2. Hierarchical EM algorithm

2.1. Hybrid hierarchical models

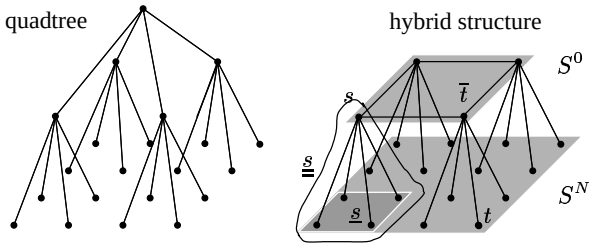


Figure 1. Two hierarchical structures: (a) quadtree with three levels; (b) truncated tree with two levels.

The hierarchical model we use is based on a hybrid structure for which one example is shown in Fig. 1(b) for a single level below the coarsest grid. To describe this graph, we shall introduce some notations.

First, we define the coarsest level S^0 as a rectangular grid with a 1st-order neighborhood. Then each site of S^0 initiates a quadtree, so that the set S^n $(0 < n \leq N)$ made up by the nodes at the level n is $2^n \times 2^n$ times larger than S^0. Now each site s of S^n has four natural correspondents in S^{n+1} (provided that s does not belong to the finest level S^N), its children, forming site set $\underline{s}$, and one natural correspondent in S^{n-1} (provided that s does not belong to the coarsest level S^0), its parent, denoted as $\overline{s}$. Finally, the site set forming the tree rooted at s is denoted $\underline{\underline{s}}$ (Fig. 1(b)). Vectors x and y are now indexed by the nodes of $S \triangleq \bigcup_{n=0}^{N} S^n$.

Given this graphical structure consider an energy function of the following form:

$$U(x, y) \triangleq \sum_{<s,t>\in S^0} v_{st}(x_s, x_t) + \sum_{s\in S\setminus S^0} w_s(x_s, x_{\overline{s}}) + \sum_{s\in S} l_s(x_s, y_s),$$

where $< s, t >$ designates pairs of neighbors in S^0, v_{st} and w_s are local functions capturing respectively the spatial prior and the hierarchical prior (they will usually encourage identity between neighbors and between parents and children, resp.), and l_s expresses the point-wise relation between the observed variable y_s and the unknown one x_s. From a probabilistic point of view, the associated joint distribution of (x, y) is: (with Z a normalizing constant)

$$\begin{aligned}\mathsf{P}(x, y) \triangleq \frac{1}{Z} &\prod_{<s,t>\in S^0} \underbrace{\exp\{-v_{st}(x_s, x_t)\}}_{\triangleq g_{st}(x_s, x_t)} \\ &\times \prod_{s\in S\setminus S^0} \underbrace{\exp\{-w_s(x_s, x_{\overline{s}})\}}_{\triangleq f_s(x_s, x_{\overline{s}})} \prod_{s\in S} \underbrace{\exp\{-l_s(x_s, y_s)\}}_{\triangleq h_s(x_s, y_s)}.\end{aligned} \quad (1)$$

2.2. EM algorithm on the hybrid structure

In the case of a spatial grid ($N = 0$), the computation of $Q(\theta|\theta^{(k)})$ is untractable due to the normalizing constant which depends on θ. Some authors [3, 15] attempt to overcome this difficulty by using the pseudo-likelihood (PL) function $\mathbb{P}(x, y|\theta)$ instead of the likelihood function. It is defined as:

$$\begin{aligned}\mathbb{P}(x, y|\theta) &\triangleq \mathsf{P}(y|x, \theta)\mathbb{P}(x|\theta) \\ &= \prod_s \mathsf{P}(y_s|x_s, \theta)\mathsf{P}(x_s|x_{\mathcal{V}_s}, \theta),\end{aligned}$$

where $\mathcal{V}_s$ represents the set of the neighbors of s and local prior conditionnal distributions can be exactly deduced from (1), according to

$$\mathsf{P}(x_s|x_{\mathcal{V}_s}) = \frac{\prod_{t\in\mathcal{V}_s} g_{st}(x_s, x_t)}{\sum_\lambda \prod_{t\in\mathcal{V}_s} g_{st}(\lambda, x_t)}.$$

Despite the elimination of the normalizing constant problem, the maximization of $Q(\theta|\theta^{(k)})$ still requires the computation of an untractable expectation. This expectation is

approximated by using a Gibbs sampler in the case of the Gibbsian EM algorithm [3].

In the case of a complete tree where S^0 reduces to a single site, a non-iterative two-sweep procedure, similar to Baum-Welch algorithm on a chain [1], can be designed to compute exactly all sitewise and pairwise posterior marginals. Then the EM algorithm can be conducted without any PL and Monte Carlo approximation ([10, 14] for discrete cases, [8, 9] for continuous Gaussian models).

With our hybrid structure, we deal both with non-causal interactions (on S^0) and tree-based interactions on subtrees $\underline{\underline{s}}$, $s \in S^0$. We thus have to introduce a PL function for the non-causal spatial part of the hybrid structure, to avoid the problem of Z. Assuming without loss of generality, that $f_s(i,j) = \mathsf{P}(X_s = i | X_{\overline{s}} = j)$, and $h_s(i,l) = \mathsf{P}(Y_s = l | X_s = i)$, the PL we deal with is:

$$\begin{aligned}\mathbb{P}(x,y|\theta) &\triangleq \prod_{s\in S^0} \mathsf{P}(x_s|x_{\mathcal{V}_s},\theta) \prod_{s\in S\setminus S^0} f_s(x_s,x_{\overline{s}},\theta) \\ &\times \prod_{s\in S} h_s(x_s,y_s,\theta).\end{aligned}$$

We further assume that f_s is independant from s whereas h_s only depends on the level n to which s belongs (these two assumptions can easily be softened or tightened in the following). We then denote: $f(i,j) \triangleq \mathsf{P}(X_s = i | X_{\overline{s}} = j)$, $\forall s \in S \setminus S^0$, and $h^n(i,l) \triangleq \mathsf{P}(Y_s = l | X_s = i)$, $\forall s \in S^n$. Each X_s taking its values in discrete state space Λ, the neighborhood configuration set $\Lambda^{\mathcal{V}_s}$ can always be partitionned in J parts such that $\mathsf{P}(x_s|x_{\mathcal{V}_s},\theta)$ only depends on the "type" $\nu \in \{1 \dots J\}$ to which $x_{\mathcal{V}_s}$ belongs. Then we denote $\mathsf{P}(X_s = i | X_{\mathcal{V}_s} \text{ of type } \nu) \triangleq a(i,\nu)$, $\forall s \in S^0$.

The more general parameterization is given by $\theta \triangleq \{a(i,\nu), f(i,j), h^n(i,l)\}$ under constraints $\sum_i a(i,\nu) = 1$, $\sum_i f(i,j) = 1$ and $\sum_l h^n(i,l) = 1$. The maximization is readily solved by using a Lagrangian based on:

$$\begin{aligned}Q(\theta|\theta^{(k)}) = &\\ &\sum_{i,\nu} [\log a(i,\nu) \sum_{s\in S^0} \underbrace{\mathsf{P}(X_s = i, X_{\mathcal{V}_s} \text{ of type } \nu | y, \theta^{(k)})}_{=\mathbb{E}[n_{i\nu}(X_{S^0})|y,\theta^{(k)}]}] \\ &+ \sum_{i,j} [\log f(i,j) \sum_{s\in S\setminus S^0} \zeta_s^{(k)}(i,j)] \\ &+ \sum_{n,i,l} [\log h^n(i,l) \sum_{s\in S^n : y_s = l} \gamma_s^{(k)}(i)],\end{aligned}$$

where $n_{i\nu}(x_{S^0}) \triangleq \#\{s \in S^0 : x_s = i, x_{\mathcal{V}_s} \text{ of type } \nu\}$, $\gamma_s^{(k)}(i) \triangleq \mathsf{P}(X_s = i | y, \theta^{(k)})$ and $\zeta_s^{(k)}(i,j) \triangleq \mathsf{P}(X_s = i, X_{\overline{s}} = j | y, \theta^{(k)})$, the updating formulae of the parameters are:

$$a^{(k+1)}(i,\nu) = \frac{\mathbb{E}[n_{i\nu}(X_{S^0})|y,\theta^{(k)}]}{\sum_i \mathbb{E}[n_{i\nu}(X_{S^0})|y,\theta^{(k)}]} \quad (2)$$

$$f^{(k+1)}(i,j) = \frac{\sum_{s\in S\setminus S^0} \zeta_s^{(k)}(i,j)}{\sum_{s\in S\setminus S^0} \gamma_{\overline{s}}^{(k)}(j)} \quad (3)$$

$$h^{n\ (k+1)}(i,l) = \frac{\sum_{s\in S^n : y_s = l} \gamma_s^{(k)}(i)}{\sum_{s\in S^n} \gamma_s^{(k)}(i)} \quad (4)$$

In the case of Gaussian data likelihoods with the parameters (μ_i, σ_i), the equation (4) is replaced by:

$$\mu_i^{n(k+1)} = \frac{\sum_{s\in S^n} \gamma_s^{(k)}(i) y_s}{\sum_{s\in S^n} \gamma_s^{(k)}(i)} \quad (5)$$

$$\sigma_i^{n(k+1)} = \left[\frac{\sum_{s\in S^n} \gamma_s^{(k)}(i)(y_s - \mu_i^{n(k+1)})^2}{\sum_{s\in S^n} \gamma_s^{(k)}(i)}\right]^{1/2} \quad (6)$$

The use of these re-estimation equations requires the computation of the expectation $\mathbb{E}[n_{i\nu}(X_{S^0})|y,\theta^{(k)}]$ on the coarse grid, and the computation of the local posterior marginals $\gamma_s^{(k)}(i)$ and of $\zeta_s^{(k)}(i,j)$ on the sub-trees below. The computation of the local posterior marginals is exactly achieved on each node of a complete tree through a non-iterative procedure made of two sweeps [10]. This procedure can be easily extended to the truncated tree [5]. The dowward recursion is now based on $\mathsf{P}(x_s|y) = \sum_{x_{\overline{s}}} \mathsf{P}(x_s|x_{\overline{s}},y)\mathsf{P}(x_{\overline{s}}|y)$, $\forall s \notin S^0$, where $\mathsf{P}(x_s|x_{\overline{s}},y) = \mathsf{P}(x_s|x_{\overline{s}},y_{\underline{\underline{s}}})$ due to separation property. The use of this recursion requires that a previous upward sweep provides $\mathsf{P}(x_s|x_{\overline{s}},y_{\underline{\underline{s}}})$ for $s \notin S^0$ and $\mathsf{P}(x_s|y)$ for $s \in S^0$. The former is achieved by successively summing out the x_s's for all $s \notin S^0$. The recursion is based on:

$$\begin{aligned}&\mathsf{P}(x_s|x_{\overline{s}},y_{\underline{\underline{s}}}) \\ &\propto f_s(x_s,x_{\overline{s}})h_s(x_s,y_s) \times \sum_{x_{\underline{\underline{s}}\setminus\{s\}}} \prod_{t\in\underline{\underline{s}}\setminus\{s\}} f_t(x_t,x_{\overline{t}})h_t(x_t,y_t) \\ &\propto f_s(x_s,x_{\overline{s}})h_s(x_s,y_s) \times \prod_{t\in\underline{s}} \underbrace{\sum_{x_{\underline{t}}} \prod_{k\in\underline{t}} f_k(x_k,x_{\overline{k}})h_k(x_k,y_k)}_{\triangleq \mathbb{F}_t(x_s)},\end{aligned}$$

with $\mathbb{F}_t(x_s) = \sum_{x_t} f_t(x_t,x_{\overline{t}})h_t(x_t,y_t)\prod_{k\in\underline{t}} \mathbb{F}_k(x_t)$. Functions $\mathbb{F}_t(x_s)$ being computed in a bottom-up way, one can then derive:

$$\mathsf{P}(x_s|x_{\overline{s}},y_{\underline{\underline{s}}}) = \frac{f_s(x_s,x_{\overline{s}})h_s(x_s,y_s)\prod_{t\in\underline{s}} \mathbb{F}_t(x_s)}{\sum_\lambda f_s(\lambda,x_{\overline{s}})h_s(\lambda,y_s)\prod_{t\in\underline{s}} \mathbb{F}_t(\lambda)}.$$

Note that the functions $\mathbb{F}_s(x_{\overline{s}})$ depend on $y_{\underline{\underline{s}}}$, even though this is not made explicit by abuse of notation. The upward sweep also provides eventually the probability $\mathsf{P}(x_{S^0}|y) = \sum_{x_{S\setminus S^0}} \mathsf{P}(x|y)$. Because of the non-causal structure on S^0, $\mathsf{P}(x_s|y)$ for $s \in S^0$ has to be approximated with the help of a Gibbs sampling of distribution $\mathsf{P}(x_{S^0}|y)$. This sampling also allows the approximation of the expectation in (2). Now the laws $\gamma_s^{(k)}(i)$ are available as well as $\mathsf{P}(x_s|x_{\overline{s}},y_{\underline{\underline{s}}})$. The computation of $\zeta_s^{(k)}(i,j)$ can be done thanks to the relation: $\mathsf{P}(x_s,x_{\overline{s}}|y,\theta^{(k)}) = \mathsf{P}(x_s|x_{\overline{s}},y,\theta^{(k)})\mathsf{P}(x_{\overline{s}}|y,\theta^{(k)})$.

The whole procedure is shown in Tab. 1. At convergence, an estimate of x is provided by the estimator of the Mode of Posterior Marginals (MPM), as a by-product of the posterior marginal computation: $\forall\ s,\ \hat{x}_s = \arg\max_{x_s} \mathsf{P}(x_s|y)$.

EM algorithm on the hybrid structure

Repeat until convergence

▲ upward sweep

Leaves ($s \in S^N$): $\mathbb{F}_s(x_{\overline{s}}) = \sum_{x_s} f_s(x_s, x_{\overline{s}}) h_s(x_s, y_s)$

Recursion (for $n = N-1 \dots 1$, $s \in S^n$): $\mathbb{F}_s(x_{\overline{s}}) = \sum_{x_s} f_s(x_s, x_{\overline{s}}) h_s(x_s, y_s) \prod_{t \in \underline{s}} \mathbb{F}_t(x_s)$

◀▶ coarse Gibbsian EM:

Repeat until convergence:

draw samples $x_{S^0}(1)$,..., $x_{S^0}(m)$ from $\mathsf{P}(x_{S^0}|y)$

approximations of $\left\{ \begin{array}{lcl} \gamma_s^{(k)}(i) & = & \frac{1}{m-r} \sum_{q=r+1}^{m} \delta[x_s(q), i] \\ a^{(k+1)}(i,\nu) & = & \frac{\sum_{q=r+1}^{m} n_{i\nu}(x_{S^0}(q))}{\sum_i \sum_{q=r+1}^{m} n_{i\nu}(x_{S^0}(q))} \end{array} \right.$

computation of $h^{0(k+1)}(i,l)$ according to (4), or of $\mu_i^{0(k+1)}$ and $\sigma_i^{0(k+1)}$ according to (5-6)

▼ downward sweep

Recursion (for $n = 1 \dots N$, $s \in S^n$): $\left\{ \begin{array}{lcl} \zeta_s^{(k)}(i,j) & = & \gamma_{\overline{s}}^{(k)}(j) \frac{f_s(i,j) h_s(i,y_s)}{\mathbb{F}_s(j)} \prod_{t \in \underline{s}} \mathbb{F}_t(i) \\ \gamma_s^{(k)}(i) & = & \sum_j \zeta_s^{(k)}(i,j) \end{array} \right.$

Computation of $f^{(k+1)}(i,j)$ according to (3)

Computation of $h^{n(k+1)}(i,l)$ according to (4), or of $\mu_i^{n(k+1)}$ and $\sigma_i^{n(k+1)}$ according to (5-6)

Table 1. Synopsis of the EM algorithm on the hybrid structure

3. Unsupervised classification comparisons

To demonstrate the practicability and the relevance of the approach for discrete low-level image analysis, we first reported comparative experiments for unsupervised classification led for $N \in \{0, 3, 4, p\}$. For $N = 0$ the algorithm corresponds to the Gibbsian EM of Chalmond [3], while $N = p$, when the size of S^N is $2^p \times 2^p$, corresponds to the complete tree ($|S^0| = 1$). $N = 3, 4$ correspond to the hybrid structure with four and five levels.

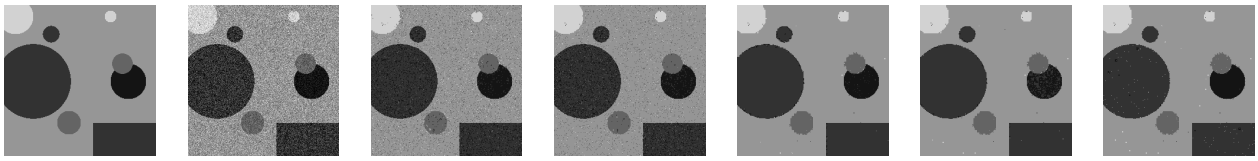
In section 1.1, we mentioned that the initialization of the EM algorithm determined the quality of the results. As a consequence, we had to pay a great attention to it and use the same for all algorithms. To this end, a simple analysis of the finest resolution data histogram was carried out to get starting values for the class parameters which were used in a standard Maximum Likelihood inference procedure. The class parameters were then re-estimated with this classification. In addition, for the Gibbsian EM, the spatial prior parameters were initialized by $\frac{n_{i\nu}(x^0_{S^0})}{\sum_i n_{i\nu}(x^0_{S^0})}$. As for the model parameters on the truncated tree, we initialized the hierarchical prior with the parameterization of Bouman [2]: $f(i,j) \triangleq \alpha\delta(i,j) + \frac{1-\alpha}{M-1}[1-\delta(i,j)]$, where M was the number of classes and with α close to 1. At each node s of S^0, we searched for the label that maximized the contribution of the subtree rooted at s, i.e., the term $\prod_{t \in \underline{s}} \mathbb{F}_t(x_s)$. From this configuration, a first estimate of the spatial prior parameters $a^{(0)}(i,\nu)$ for the coarse Gibbsian EM algorithm could be computed as for the Gibbsian EM algorithm on the lattice.

The EM algorithms were stopped when the variations of the class parameters from an iteration to another became non-significant. In our experiments, data were only available on the finest resolution level, consequently the stopping criterion was the following: $\frac{1}{M}\Big\{ \sum_{i=1}^{M} [(\mu_i^{(k+1)} - \mu_i^{(k)})^2 + (\sigma_i^{(k+1)} - \sigma_i^{(k)})^2] \Big\}^{1/2} < \epsilon$, with $\epsilon = 0.1$ in our experiments.

First, the experiments were carried out on a 256×256 synthetic image involving 5 classes (Fig. 2). We applied an additive Gaussian white noise with a different standard deviation for each class, thus the gray level means and variances $(\mu_i, \sigma_i^2)_{i=1}^5$ were known and could be compared to the ones estimated by the different EM procedures in Tab. 2(a). We reported here the results obtained by the four methods with the number of classes forced to five. The obtained MPM classifications are shown in Fig. 2 and the respective percentages of good classification and computational loads, including cpu times (on a 360 MHz Ultra 60 Sun workstation), can be found in Tab. 2(b).

As can be seen from the results, the hierarchical models provided much better results than the plain spatial Gibbsian EM algorithm and took less cpu time. Increasing the number of retained samples ($m-r$, see Tab. 1) in the plain Gibbsian EM algorithm improved slightly the classification but implied a redhibitory computational load. It should be noticed that the hierarchical algorithms provided almost same results both in terms of quality of the classification and in terms of accuracy of the parameter estimates. The coarse Gibbsian EM took no more than five iterations with 20 retained samples within each iteration. If we computed cpu time needed for one iteration of each hierarchical method, we found 4.3s for the complete-tree-based and the five-level-based algorithms and 4.5s for the last one. Thus the use of a sampling procedure in the hybrid EM algorithm did not seem to imply a significant extra computational load.

The previous algorithms were applied to SPOT satellite images provided by Costel (geography laboratory of the University of Rennes 2) in the context of remote sensing researches within a research project called GSTB ("Groupe-



Ground-truth Noisy image $N = 0^{(1)}$ $N = 0^{(2)}$ $N = 3$ $N = 4$ $N = 8$

Figure 2. Unsupervised classifications with a synthetic image. ((1)=5 samples (2)=100 samples)

	Categories				
	1	2	3	4	5
GT	**20** (15)	**50** (30)	**100** (20)	**150** (30)	**210** (25)
$0^{(1)}$	**16.5** (12.8)	**55.2** (24.2)	**105.3** (19.0)	**152.5** (25.1)	**207.3** (22.0)
$0^{(2)}$	**14.9** (11.8)	**53.6** (23.5)	**103.0** (18.5)	**151.9** (26.1)	**208.0** (22.4)
3	**18.8** (13.1)	**50.2** (28.6)	**99.0** (20.4)	**149.3** (29.8)	**209.2** (23.7)
4	**14.3** (10.3)	**49.1** (28.4)	**99.1** (20.4)	**149.4** (29.7)	**209.1** (23.7)
8	**19.7** (13.9)	**49.9** (27.9)	**98.8** (20.2)	**149.0** (29.6)	**208.7** (23.8)

(a) Estimated class means and standard deviations. (GT=Ground truth)
(The bold-faced number corresponds to the mean and the one in parentheses to the variance.)

Model	good class.	nb iter.	nb samples	cpu time
$0^{(1)}$	88%	29	5	120s
$0^{(2)}$	91%	35	100	25min
3	99%	19	5×20	87s
4	98%	11	5×20	48s
8	99%	12	none	52s

(b) Performances of the different EM algorithms.

Table 2. Comparative results for the different EM algorithms with the synthetic image in Fig. 2. ($0^{(1)}$=Gibbsian EM with 5 samples, $0^{(2)}$=Gibbsian EM with 100 samples, 3=four-level hybrid structure-based EM, 4=five-level hybrid structure-based EM, 8=quad-tree-based EM)

ment Scientifique de Télédétection en Bretagne"). The extracted scene (Fig. 3) was composed of three 512×512 images with different wavelenghts and represented the Bay of Lannion, located in the north-west of France, during December 1996. The goal of this study was to determine the land cover of this area. To reach this aim, the geographers of Costel built a list of eight classification categories: (1) Sea and water, (2) Sand and bare soils, (3) Urban areas, (4) Forests and heath, (5) Temporary meadows, (6) Permanent meadows, (7) Colza, (8) Vegetables. Thanks to both tests on the lands and photointerpretations, they were also able to extract small image portions which are samples of the eight categories on the three SPOT images of the scene. We used them to assess the accuracy of the classifications and compared the parameters (gray level means, variances and possibly correlation coefficients) of each category for each image learned from the samples, to the ones given by the EM algorithms (see Tab. 3(a)). In fact, we fixed the parameters of the last four categories, because they were undistinguishable with automatic process but they were essential for the addressed application.

The model can be easily extended to the case of multispectral data by taking into account the correlation between the three spectral bands. We experimented both uncorrelated and correlated data likelihoods. However, even if the SPOT bands are known to be correlated, considering correlated channels in our example did not improve the classification results significantly. Thus we here only presented results considering the channels as independent.

The algorithms provided quite similar results of a good quality (see Tab. 3 and Fig. 4). About 89% of the pixels of the samples were well classified.

4. Conclusion and extensions

In this paper, we presented an EM algorithm built on a hybrid hierarchical structure which is an interesting compromise between standard spatial models and hierarchical models based on a complete quad-tree. To study this algorithm thoroughly, we should concentrate on how it deals with different initializations of the class parameters, especially when they become very far from the right parameters.

With this structure we now plan to address the issue of automatically estimating the optimal number of levels in the struture. Moreover, we would like to investigate the possibility to truncate the tree in a heterogeneous manner, so that we could obtain pieces of coarse grids at different resolution depending on the available data.

References

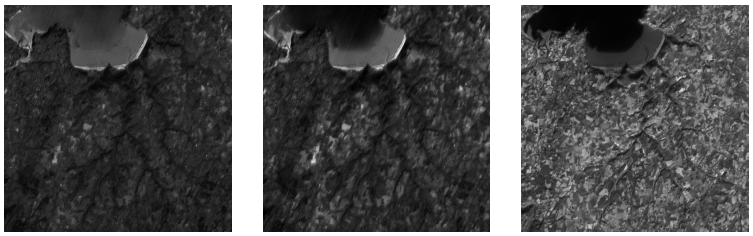
[1] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41:164–171, 1970.

[2] C. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Trans.Im.Proc.*, 3(2):162–177, 1994.

[3] B. Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pat. Recogn.*, 22(6):747–761, 1989.

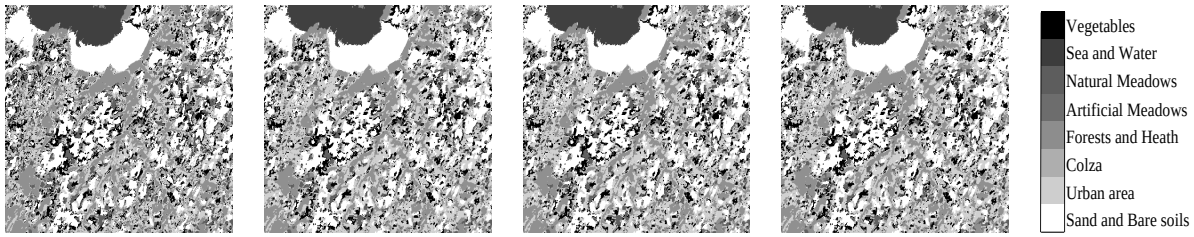
[4] A. Chardin and P. Pérez. Semi-iterative inference with hierarchical models. In *ICIP'98*, pages 630–634, Chicago, USA, October 1998.

[5] A. Chardin and P. Pérez. Mode of posterior marginals with hierarchical models. In *ICIP'99*, Kobe, Japan, October 1999.



Channel XS1 Channel XS2 Channel XS3

Figure 3. 512×512 **SPOT images (courtesy of Costel, University of Rennes 2, and GSTB).**



$N = 0$ $N = 3$ $N = 4$ $N = 8$ legend

Figure 4. Unsupervised classifications with the multispectral satellite in Fig. 3.

[6] K. Chou, S. Golden, and A. Willsky. Multiresolution stochastic models, data fusion and wavelet transforms. *Sig. Proc.*, 34(3):257–282, 1993.

[7] A. Dempster, L. N.M., and D. Rubin. Mixtures densities, maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Stat. Society*, 39(1):1–38, 1977.

[8] V. Digalakis, J. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Trans. Speech and Audio Proc.*, 1(4):431–442, 1993.

[9] A. Kannan, M. Ostendorf, W. Karl, D. Castanon, and R. Fish. ML parameter estimation of a multiscale stochastic process using the EM algorithm. Technical Report ECE-96-009, Boston University, Nov. 1996.

[10] J.-M. Laferté, P. Pérez, and F. Heitz. Discrete Markov image modeling and inference on the quad-tree. *IEEE Trans.Im.Proc.*, Accepted for publication, 1999.

[11] M. Luettgen, W. Karl, and A. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Trans.Im.Proc.*, 3(1):41–64, 1994.

[12] P. Pérez, A. Chardin, and J.-M. Laferté. Noniterative manipulation of discrete energy-based models for image analysis. *under press, Pat. Recogn.*, 1999.

[13] R. Redner and H. F. Walker. Mixtures densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[14] D. Tretter, C. Bouman, K. Khawaja, and A. Maciejewski. A multiscale stochastic image model for automated inspection. *IEEE Trans.Im.Proc.*, 4(12):1641–1654, 1995.

[15] J. Zhang, J. Modestino, and D. Langan. Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. *IEEE Trans.Im.Proc.*, 3(4):404–420, 1994.

		water	bare soils	urban areas	forests
GT	XS1	**32.4** (1.4)	**40.2** (3.6)	**34.1** (1.4)	**28.1** (1.4)
	XS2	**12.2** (0.6)	**24.8** (2.9)	**17.5** (1.1)	**13.6** (1.0)
	XS3	**6.5** (0.6)	**23.9** (4.3)	**26.6** (3.5)	**22.8** (7.3)
N=0	XS1	**34.9** (2.2)	**40.6** (4.1)	**34.1** (1.4)	**30.4** (1.4)
	XS2	**13.7** (1.6)	**25.0** (3.8)	**19.7** (1.5)	**15.6** (1.4)
	XS3	**6.5** (0.6)	**26.7** (8.9)	**28.9** (5.5)	**26.3** (6.3)
N=3	XS1	**35.0** (2.2)	**39.6** (4.2)	**33.9** (1.2)	**30.6** (1.5)
	XS2	**13.8** (1.7)	**24.4** (3.7)	**19.2** (1.3)	**15.8** (1.4)
	XS3	**6.6** (0.6)	**27.3** (8.2)	**31.9** (7.2)	**26.9** (6.6)
N=4	XS1	**35.0** (2.2)	**39.1** (4.2)	**33.8** (1.1)	**30.6** (1.5)
	XS2	**13.8** (1.7)	**23.9** (3.7)	**19.1** (1.2)	**15.8** (1.4)
	XS3	**6.6** (0.6)	**27.5** (8.2)	**31.9** (7.1)	**26.9** (6.5)
N=9	XS1	**35.1** (2.2)	**40.0** (4.2)	**33.9** (1.3)	**30.5** (1.4)
	XS2	**13.8** (1.7)	**24.7** (3.7)	**19.2** (1.4)	**15.7** (1.4)
	XS3	**6.6** (0.6)	**27.0** (8.4)	**31.9** (7.2)	**26.8** (6.6)

(a) Estimated class parameters. (GT=Ground truth)
(means in bold-faced type and variances in parentheses.)

Model	good class.	nb iter.	nb samples	cpu time
$N = 0$	89%	15	5	880s
$N = 3$	89.2%	9	5×20	670s
$N = 4$	89.8%	10	5×20	700s
$N = 9$	89.3%	9	none	660s

(b) Performances of the different EM algorithms

Table 3. Comparative results for the different EM algorithms with the multispectral satellite images in Fig. 3.