A multiresolution EM algorithm for unsupervised image classification

J.-M. Laferté†, F. Heitz‡ & P. Pérez†

† IRISA - INRIA Rennes - Université de Rennes I
Campus universitaire de Beaulieu - 35042 Rennes Cedex, France
Email {laferte,perez}@irisa.fr

‡ ENSPS / LSIIT, URA CNRS 1871,
Bd Sébastien Brant, 67400 Illkirch, France.
Email heitzfa@enspsmail.u-strasbg.fr

Abstract

We take benefit from a causal Markov model defined on a quadtree to derive a multiresolution EM algorithm for unsupervised image classification. This algorithm is an efficient alternative to expensive or approximate EM algorithms associated with Markov Random Fields. We show on synthetic and real images that our algorithm also provides good or even better results than those obtained by spatial MRF models.

1. Introduction

Hidden Markov Models (HMM) have been successfully introduced in many important issues in image processing, computer vision or pattern recognition. The mathematical framework is a statistical one: entities (labels) of interest are described by statistical models (Markov Chains or Markov Random Fields) and Bayesian estimation theory is used to extract the relevant information from noisy observations. By defining comprehensive, global non-linear models, HMMs have led to significant improvements with respect to local methods.

A key issue, when handling these statistical approaches, is the estimation of the model parameters. To this end, several statistical methods have been proposed in the last decade, leading to supervised or unsupervised estimation schemes. Unsupervised or "data-driven" methods are the most attractive, since they do not require any human interaction. They are generally based on "restoration-maximization" techniques in which the labels to be extracted and the model parameters are estimated alternately. Iterative algorithms, such as the classical EM algorithm [9, 11], belong to this class and are used to produce maximum likelihood estimates of the model parameters. When dealing with Markov chains one can derive an exact EM algorithm known as the Baum-Welch algorithm [1]. Unfortunately, this is not the case for non-causal Markov models generally used in image processing such as Markov Random Fields (MRF). In this case approximations, such as the Gibbsian EM algorithm have to be considered [5]. These algorithms remain however painfully slow and are quite sensitive to initialization.

In this paper we consider a different modeling framework (Causal Markov Random Fields defined on a quadtree) which enables a maximum likelihood estimation of the model parameters to be computed in an unsupervised image classification problem. Markov models on quadtrees have been introduced recently [3, 10, 8] to represent various features in low level image analysis (regions, optical flow fields, etc.). They provide fast and efficient estimation algorithms, thanks to the causality property of the model [8]. The approach is thus less expensive than a standard spatial Markov modeling approach, and, as we will see, gives better qualitative results than the (approximate) Gibbsian EM algorithm or other stochastic variants of the EM algorithm.

The remainder of this paper is organized as follows: in section 2 we recall briefly the main features of the EM algorithm and its variants, especially in the Markovian framework. In section 3, we introduce the hierarchical statistical representation under concern: the causal Markov model on the quadtree and the multires-

olution EM algorithm associated with it. We show that the quadtree structure avoids the main approximations one has to resort to, when non causal spatial models such as MRFs are considered. In section 4 the multiresolution EM algorithm (defined on the quadtree model) is compared to the Gibbsian EM algorithm (associated to a 4-neighborhood MRF) and to another stochastic variant of the same algorithm. Results on synthetic and real data show that the multiresolution EM algorithm is an attractive alternative to the EM Gibbsian scheme.

2. The EM algorithm:background

2.1. The standard EM algorithm

Let X designate a hidden label field: $X = \{X_s, s \in S\}$ defined on lattice S. Let $Y = \{Y_s, s \in S\}$ be the associated observation field. (X, Y) called the "complete data" is characterized by the joint distribution $P(X, Y|\Phi)$ depending on a parameter vector Φ to be estimated. The EM algorithm is directed at finding a value of Φ which maximizes the likelihood:

$$\hat{\Phi} \stackrel{\triangle}{=} \arg\max_{\Phi} P(y|\Phi).$$

The EM algorithm iteratively computes the maximum likelihood estimate by repeating, until convergence, the two following steps:

- **E-step** (*expectation*) Computation of $Q(\Phi|\Phi^{(n)}) = \mathbb{E}\left[\log P(X, Y|\Phi)|y, \Phi^{(n)}\right]$.
- **M-step** (*maximization*) Update $\Phi^{(n+1)}$ such that $\Phi^{(n+1)} = \arg\max_{\Phi} Q(\Phi|\Phi^{(n)})$

The sequence $(\Phi^{(i)})_i$ converges toward a stationary point that can be either a saddle point or a local maximum of the likelihood (see [9, 11] for more details).

2.2. Stochastic variants of the EM algorithm

The main limitations of the EM algorithm are the following:

- the final result strongly depends on the initialization;
- the algorithm can get stuck in a local maximum or in a saddle point;
- it is generally painfully slow;
- the maximization step may be intractable;
- the computation of the expectation may not be feasible.

To overcome these limitations different variants of the EM algorithm have been developed. Among them is the stochastic EM family (an overview of stochastic EM algorithms can be found in [4]). Stochastic versions of the EM algorithm are less trapped by saddle points or local maxima and thus partially overcome the initialization problem.

There are other motivations for using stochastic steps in the EM algorithm. In the MCEM (Monte-Carlo EM) algorithm for instance, Wei and Tanner [12] propose a Monte-Carlo method to estimate the expectation Q by a Monte-Carlo method when it can not be expressed in a closed form.

2.3. The EM algorithm and Markov models

If (X, Y) is a MRF, then the distribution of (X, Y) is Gibbsian (according to a theorem by Hammersley and Clifford)[7]:

$$P(X, Y|\Phi) = \frac{1}{Z(\Phi)} e^{-U(X,Y,\Phi)}$$

where $Z(\Phi)$ is a normalizing constant that depends on Φ. This constant generally makes the computation of Q untractable. The computation of $Z(\Phi)$ is indeed generally not tractable apart from small size or factored problems. The solution adopted by some authors [5, 13] is to substitute the likelihood function for the so-called "pseudo-likelihood" function defined as follows:

$$\mathbb{P}(X, Y|\Phi) \stackrel{\triangle}{=} P(Y|X, \Phi)\mathbb{P}(X|\Phi)$$

with $\mathbb{P}(X|\Phi) \stackrel{\triangle}{=} \prod_s P(X_s|X_{\partial_s}, \Phi)$, where ∂_s stands for the set of neighbors of s. Thus the global normalizing constant disappears. Unfortunately the maximization step involves the computation of expectations which are very expensive. Zhang *et al.* [13] avoid these computations thanks to new approximations and derive a ICM-like algorithm. Chalmond [5] estimates these expectations by using the expensive Gibbs sampler (justifying the name of the associated algorithm : "Gibbsian EM algorithm").

Although they provide interesting results, these algorithms have two main drawbacks:

- These algorithms cannot be considered as EM-like algorithms, despite their name. Indeed by using a pseudo likelihood function they do not converge toward a global (or even a local) maximum of the likelihood.
- A major drawback of the Gibbsian EM algorithm is the huge computational complexity of the Gibbs sampler used to sample the non-causal MRF. As

we know, the Gibbs sampler is an iterative algorithm consisting in stochastic relaxations [7]. It generally takes a long time before one can assume that the "thermal equilibrium" is reached.

The modeling approach proposed in the next section allows to overcome these three major problems by introducing a causal Markov model in scale which enables very efficient parameter estimation methods to be developed. Let us notice that some authors use simple Markov chains to model an image (Pickard Fields [6], 1-D Markov chains using Peano sweeps [2] ...). Although these models induce very efficient algorithms (based on the well known Baum-Welch algorithm [1] which is the exact EM version for Markov chains), they are obviously less general than MRFs.

3. A Multiresolution EM algorithm

3.1. Hidden Markov models on the quadtree

Let us briefly recall the hierarchical model introduced by the authors in [8]. X and Y are random processes indexed by the nodes of a quadtree. Let S denote the set of these nodes (see figure 1). Let $s\,\bar{\gamma}$ designate the unique "father" of node s. The restriction of a set E to a scale n is denoted by E^n (thus $X^n \triangleq \{X_s, s \in S^n\}$). Level $n = 0$ corresponds to the finest resolution, whereas scale $n = R$ is the coarsest one. S^R is the root that will be denoted r in the sequel (see figure 1): $S^R = \{r\}$.
X_s represents the label at site s while Y_s represents the observation at the same site. In the sequel we focus on a standard classification task, that is X_s takes its values in a finite set $\Lambda = \{1, \cdots, M\}$ where M stands for the number of classes. In order to have observations at all sites, we fill in missing data by creating a low-pass pyramid from the original image (using Gaussian filtering followed by a subsampling). Thus if we consider the simplest case where we only have a single image y^0 at the finest resolution [1] we build a pyramid of observations $y = \{y^n, n = 0, .., R\}$.

We make the following assumptions about the processes X and Y [8]:

- The fundamental hypothesis consists in assuming the stochastic process X to be Markovian in scale:

$$P(X^n|X^i, i > n, \Phi) = P(X^n|X^{n+1}, \Phi) \qquad (1)$$

[1] If we have multispectral data, Y_s becomes a vector of observations.

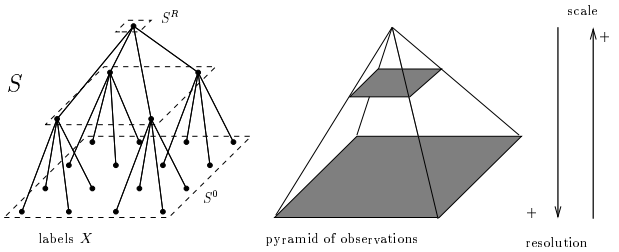

Figure 1: *Hierarchical structure: the quadtree*

- The scale-to-scale transitions are assumed to decompose as follows:

$$P(X^n|X^{n+1}, \Phi) = \prod_{s \in S^n} g^n(X_s|X_{s\bar{\gamma}}, \Phi) \qquad (2)$$

- We also assume that the likelihood of observations is of the following form:

$$P(Y|X, \Phi) = \prod_n \prod_{s \in S^n} f^n(Y_s|X_s, \Phi) \qquad (3)$$

Under assumptions (1), (2) and (3) (X, Y) is Markovian on the quadtree.

Finally we assume that the *a priori* distribution of the root is uniform, *i.e.* $P(x_r|\Phi) = \frac{1}{M}$.

We also introduce the following notations:

$$\begin{aligned}
P_{i,j} &\triangleq P(X_s = i|X_{s\bar{\gamma}} = j), \forall s \in S \setminus \{r\};\\
P_i(k) &\triangleq P(Y_s = k|X_s = i), \forall s \in S^0;\\
n_{i,j}(x) &\triangleq \sharp\{s \in S : X_s = i, X_{s\bar{\gamma}} = j\};\\
m_{i,k}(x) &\triangleq \sharp\{s \in S : X_s = i, Y_s = k\};\\
\gamma_s^{(n)}(i) &\triangleq P(x_s = i|y, \Phi^{(n)}).
\end{aligned}$$

For the classification task (in the Gaussian case), the set of parameters is $\Phi \triangleq \{M, P_{i,j}, \mu_i^n, \Gamma_i^n, i, j \in \Lambda, n = 0, \ldots, R\}$ where μ_i^n designates the mean of class i at scale n and Γ_i^n is the corresponding covariance matrix. If we only have a single image at the finest resolution [2] we set $\mu_i^n = \mu_i^0$ and $\Gamma_i^n = \Gamma_i^0, \forall i, \forall n$. In this case $\Phi \triangleq \{M, P_{i,j}, \mu_i^0, \Gamma_i^0, i, j \in \Lambda\}$.

3.2. The multiresolution EM algorithm

Now we describe the algorithm and emphasize why this modeling approach is more appealing than the spatial MRF modeling technique.

[2] Otherwise consider the case where we have an image at level 0 and an image at level 2 for instance. Then we set for $n = 0, 1, \mu_i^n = \mu_i^0, \forall i$ and we set $\mu_i^n = (\mu_i^0, \mu_i^2)^T, \forall i, \forall n \geq 2$.

- First, we do not need to use the pseudo-likelihood (PL) function. More precisely for the quadtree model, the PL function is equal to the likelihood function, *i.e.* :

$$P(x,y|\Phi) = P(x_r) \prod_{s \neq r} P(x_s|x_{\bar{s\gamma}},\Phi) \prod_s P(y_s|x_s,\Phi).$$

Thus a standard EM may be used (contrary to the case of a spatial MRF model). We have:

$$\begin{aligned} Q(\Phi|\Phi^{(n)}) &= \sum_{i,k} \mathbb{E}(m_{i,k}|y,\Phi^{(n)}) \log P_i(k) + \\ &\quad \sum_{i,j} \mathbb{E}(n_{i,j}|y,\Phi^{(n)}) \log P_{i,j} \end{aligned}$$

The maximization of Q is classically solved by using a Lagrangian. Finally the re-estimation formulae become [5]:

$$\begin{cases} P_{i,j}^{(n+1)} &= \frac{\mathbb{E}_{(n_{i,j}|y,\Phi^{(n)})}}{\sum_i \mathbb{E}_{(n_{i,j}|y,\Phi^{(n)})}}, i,j \in \Lambda \\ \mu_i^{(n+1)} &= \frac{\sum_s \gamma_s^{(n)}(i)\, y_s}{\sum_s \gamma_s^{(n)}(i)} \\ \Gamma_i^{(n+1)} &= \frac{\sum_s \gamma_s^{(n)}(i)\,(y_s-\mu_i^{(n)})\,(y_s-\mu_i^{(n)})^T}{\sum_s \gamma_s^{(n)}(i)} \end{cases}$$

- The model is causal, contrary to the single resolution spatial MRF model. Thus, it induces non iterative sampling methods. As a consequence the sampling of (X,Y) is far less expensive than with the Gibbs sampler. At each iteration we draw T samples from the following posterior distribution $P(x_s|y_s,x_{\bar{s\gamma}},\Phi^{(n)})$, using non-iterative coarse-to-fine sweeps over the quadtree. The following approximations are used in the re-estimation formulae:

$$\begin{aligned} \mathbb{E}(n_{i,j}|y,\Phi^{(n)}) &\simeq \frac{1}{T}\sum_{t=1}^{T} n_{i,j}(x(t)) \\ \gamma_s^{(n)}(i) &\simeq \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{x_s(t)=i} \end{aligned}$$

where $x(t)$ represents the t^{th} sample from $P(X|y,\Phi^{(n)})$.

- The labels are simultaneously updated according to the "Maximum of Posterior Marginals" criterion:

$$x_s^{(n+1)} = \arg\max_i \gamma_s^{(n+1)}(i), \forall s.$$

- The number of classes is estimated as follows: at the beginning it is set equal to an upper bound (generally quite large), then at each iteration, classes with size lower than a threshold are removed. The criterion chosen to end the algorithm is:

$$\max_{i\in\Lambda} \|(\mu_i,\Gamma_i)^{(k+1)} - (\mu_i,\Gamma_i)^{(k)}\| < \epsilon$$

with $\|(\mu,\Gamma)\| \triangleq |\mu| + |\Gamma|$ in the scalar case and with $\epsilon = 1.0$ in our experiments.

4. Experimental results

We present experimental results on synthetic and real-world images. The application we address is a standard classification problem which consists in partitioning the images into a certain number of classes. We compare the multiresolution EM algorithm with the Gibbsian EM algorithm (and with the MCEM for a mixture-of-Gaussian model [12]). A standard MRF with a 4-neighborhood structure has been used in the case of the Gibbsian EM algorithm.

4.1. Synthetic image

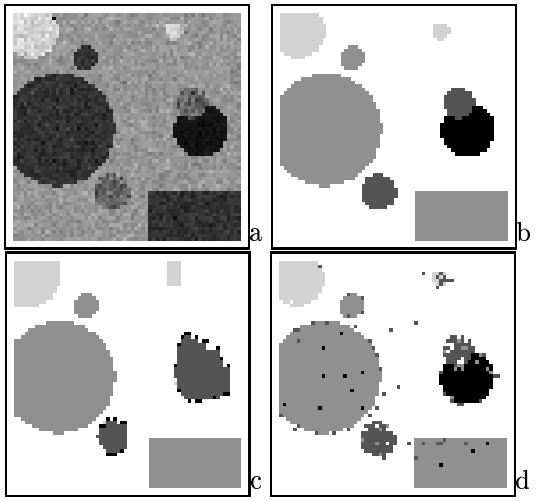

Figure 2: *Results obtained using different versions of the EM algorithm on a synthetic image. The number of classes is* 5. *a)* 64×64 *image, b) Ground-truth, c) classification map after* 341×50 *(that is* $T = 50$*) iterations of the Gibbsian EM algorithm, d) classification map after* 228×10 *iterations of the multiresolution EM algorithm.*

We begin by a synthetic image showing different noisy patterns on an uniform background (figure 2). The multiresolution EM algorithm succeeds in finding the good number of classes. The parameters estimated by the multiresolution EM algorithm and by the Gibbsian EM are compared in table 1 in the case of a fixed

		Class				
		1	2	3	4	5
GT	Mean	20	50	100	150	210
MREM	Mean	19	49	117	149	210
GEM	Mean	60	49	109	148	197
GT	Var	25	100	400	144	225
MREM	Var	28	102	1896	134	195
GEM	Var	2255	129	2150	277	1441

Table 1: *Parameters estimated by two different versions of the EM algorithm (the image is presented figure 2). GT = Ground-truth, MREM = Multi-Resolution EM and GEM = Gibbsian EM.*

number of classes. The classification maps appear on figure 2. The multiresolution model is faster and gives better results than the spatial MRF model. In particular the Gibbsian EM algorithm does not succeed in discriminating two different classes corresponding to circles (fig. 2c).

4.2. Muscle cells image

We have compared the three versions of the EM algorithm on a real-world image representing muscle cells: the MCEM (with 10 iterations to compute the MC approximations of expectation Q at each step), the Gibbsian EM algorithm (with 50 iterations for the Gibbs sampler, of which 40 to reach the "thermal equilibrium") and the multiresolution EM algorithm presented in this paper (with 10 samples at each step, *i.e.* $T = 10$). The results are presented in fig. 3. To compare these algorithms, they have been run in the same conditions (the number of classes is not estimated here and is set equal to 4, and the initialization is obtained by a coarse analysis of the histogram). It appears that the Gibbsian EM algorithm does not succeed in escaping from the (relatively) bad initialization and does not discriminate the background (in white on fig. 3) and a class of cells. This is not the case of the multiresolution EM algorithm. Lastly the MCEM with the mixture model leads to more noisy results since it does not include contextual information.

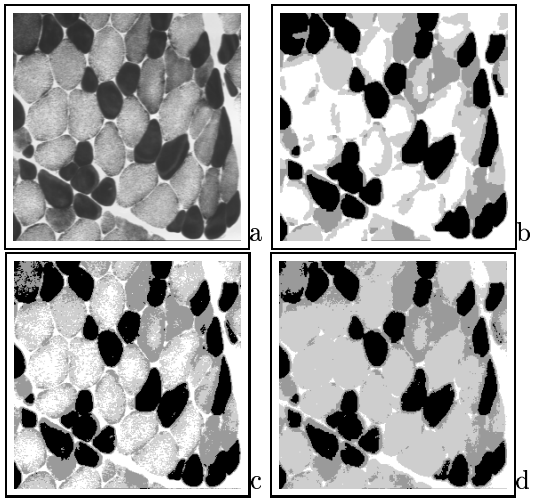

Figure 3: *Results of different versions of the EM algorithm on the "muscle" image. The number of classes is not estimated here and is set equal to* 4. *a)* 256×256 *image, b) result after* 41×50 *iterations of the Gibbsian EM algorithm, c) result after* 52×10 *iterations of a simple MCEM algorithm, d) result after* 228×10 *iterations of the multiresolution EM algorithm.*

References

[1] L. E. BAUM and T.PETRIE and G. SOULES and N. WEISS. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.*, Vol 41: pp. 164–171, 1970.

[2] B. BENMILOUD and W. PIECZYNSKI. Parameter Estimation in Hidden Markov Chains and Segmentation of Images. (in french) *Traitement du Signal*, Vol. 12, No 5: pages 433–454, 1995.

[3] C. BOUMAN and M. SHAPIRO. A multiscale random field model for bayesian image segmentation. *IEEE Trans. on Image Processing*, Vol. 3, No. 2 : pages 162–177, March 1994.

[4] G. CELEUX, D CHAUVEAU, and J. DIEBOLT. On stochastic versions of the EM algorithm. Technical Report 2514, INRIA, March 1995.

[5] B. CHALMOND. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, Vol. 22, No 6: pages 747–761, 1989.

[6] P.A. DEVIJVER. Image segmentation using causal markov random fields models. In *Int. Conf. Pattern Rec.*, pages 153–158, Atlantic City, June 1990.

[7] S. GEMAN and D. GEMAN. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No 6: pp. 721–741, November 1984.

[8] J.-M. LAFERTÉ, F. HEITZ, P. PÉREZ, and E. FABRE. Hierarchical statistical models for the fusion of multiresolution image data. In *Proc. Int. Conf. on Computer Vision*, pages 908–913, Cambridge, USA, 1995.

[9] N. M. LAIRD A. P. DEMSTER and D. B. RUBIN. Mixtures densities, maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Stat. Society*, Vol. 39, No 1: pages 1–38, 1977.

[10] M. LUETTGEN, W. KARL, A. WILLSKY, and R. TENNEY. Multiscale representations of Markov Random Fields. *IEEE Trans. Signal Processing*, Vol. 41, No 12: pages 3377–3395, Dec. 1993.

[11] R. A. REDNER and H. F. WALKER. Mixtures densities, maximum likelihood and the EM algorithm. *SIAM Review*, Vol. 26, No 2: pages 195–239, April 1984.

[12] G.C.G. WEI and M.A. TANNER. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85: pages 699–704, 1990.

[13] J. ZHANG, J. W. MODESTINO, and D. A. LANGAN. Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. *IEEE Trans. on Image Processing*, Vol. 3, No 4: pages 404–420, july 1994.