Mode of Posterior Marginals with Hierarchical Models

Annabelle Chardin Patrick Pérez
IRISA/INRIA-Rennes
Campus de Beaulieu, F-35042 Rennes cedex, France.
e-mail: {achardin,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. We investigate a new hybrid hierarchical structure: a Markov random field attached to the nodes of a truncated tree. It thus combines causal hierarchical prior on trees with a non-causal spatial prior at the coarsest level. We address the problem of computing posterior marginals with such a prior structure. This is performed using non-iterative two-sweep marginalizations on trees, combined with a low cost Gibbs sampler in between the two sweeps. Posterior marginals are thus obtained in a semi-iterative way. They are then used to infer unknown variables according to the MPM estimator. This is illustrated by experiments on synthetic data and on multispectral satellite images.

1 Background: hierarchical models

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–, $y = (y_j)_j$. Within the framework of Markov random field-based approaches x and y are random vectors and we have the following relation between the joint distribution and the energy function: $\mathsf{P}(x, y) \propto \exp\{-U(x, y)\}$. As far as the estimate of x is concerned, the Bayesian estimation theory provides two standard estimators: the Maximum A Posteriori (MAP) estimator which corresponds to the global minimizer of the energy fonction ($\hat{x} = \arg\max_x \mathsf{P}(x|y) = \arg\min_x U(x, y)$) and the Modes of Posterior Marginals (MPM) estimator ($\forall\ i,\ \hat{x}_i = \arg\max_{x_i} \mathsf{P}(x_i|y)$) which requires computation of marginals by summation over huge sets of configurations.

Although the energy function is usually decomposed as a sum of *local* terms involving just a few variables at a time, exact estimates (with respect to MAP or MPM criterion) are generally not possible. Thus one has to devise deterministic or stochastic iterative algorithms which may converge very slowly. Given this ascertainment, the pressing need for specific models leading to more efficient inference is well understandable.

That is why hierarchical models were introduced. They usually lie on the nodes of a quad-tree whose leaves fit the pixels of (maximum resolution) images [2, 11, 9, 6, 10, 7]. In this case, they allow the design of exact non-iterative inference procedures made of two sweeps. As a consequence, hierarchical models provide gains in terms of the computational efficiency. They also may improve the quality of results. However, their structural constraints might appear artificial for certain types of problems or data and one can especially wonder which significance can be given to the coarsest estimates. Another drawback is that the manipulation of the complete tree-structure can become very cumbersome in case of large images.

To circumvent these shortcomings, we propose here a hierarchical model based on an "hybrid" structure which combines a spatial grid of reduced size at coarser level with "sub-trees" appended below it, down to the finest level (see Fig.1). The study of the MAP estimator on this structure has already been conducted in [4]. MPM estimator, however, is known as more reliable than MAP, and provides posterior marginals which are a key ingredient for EM-type parameter estimation techniques. That is why we investigate the use of the MPM estimator on the hybrid structure.

2 Hybrid hierarchical model and semi-iterative MPM

The hierarchical model we use is based on a hybrid structure composed of a rectangular grid S^0 and of quadtrees rooted at each site of S^0, so that the grid 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 [4]. One example is shown in Fig.1 for a single level below the coarsest grid. We denote $\overline{i}$ the parent of a site i (provided that

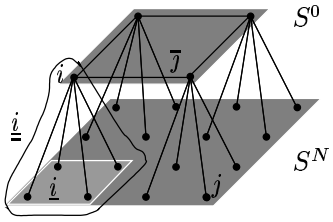


Figure 1: Hybrid hierarchical structure ($N = 1$)

i does not belong to the coarsest level S^0), $\underline{i}$ the set of the children of i (provided that i does not belong to the finest level S^N), and $\underline{\underline{i}}$ the site set forming the tree rooted at i. Vectors x and y are now indexed by the nodes of $S = \bigcup_{n=0}^{N} S^n$. Thus the posterior distribution is assumed to be of the following form:

$$\begin{aligned}\mathsf{P}(x,y) &\propto \exp\{-U(x,y)\} \\ &= \prod_{<i,j>\in S^0} g_{ij}(x_i,x_j) \times \prod_{i\notin S^0} f_i(x_i,x_{\overline{i}}) \\ &\times \prod_{i\in S} h_i(x_i,y_i),\end{aligned}$$

where $< i,j >$ designates pairs of neighbors in S^0 (for a given spatial neighborhood system on S^0), g_{ij} and f_i are local functions capturing respectively the spatial and hierarchical prior (they will usually encourage identity between neighbors and between parents and children, resp.), and h_i expresses the point-wise relation between observed variable y_i and unknown one x_i.

In the case of a complete tree where S^0 reduces to a single site, the exact MPM estimates can be computed on each node [9] using an extension of Baum-Welch algorithm on a chain [1]. In the general case, the downward recursion is now based on the following relation: $\forall i \notin S^0$, $\mathsf{P}(x_i|y) = \sum_{x_{\overline{i}}} \mathsf{P}(x_i|x_{\overline{i}},y)\mathsf{P}(x_{\overline{i}}|y)$, where $\mathsf{P}(x_i|x_{\overline{i}},y) = \mathsf{P}(x_i|x_{\overline{i}},y_{\underline{\underline{i}}})$ due to separation property. The preliminary upward sweep should then provide $\mathsf{P}(x_i|x_{\overline{i}},y_{\underline{\underline{i}}})$, $\forall i \notin S^0$ and $\mathsf{P}(x_i|y)$, $\forall i \in S^0$. This is achieved by successively summing out x_i's for all $i \notin S^0$. This eventually provides the probability $\mathsf{P}(x_{S^0}|y) = \mathsf{P}(x_i,\ i \in S^0|y) = \sum_{x_{S\setminus S^0}} \mathsf{P}(x|y)$. Because of the non-causal structure on S^0, $\mathsf{P}(x_i|y)\ \forall i \in S^0$ have to be approximated with the help of a Gibbs sampling [8] of the distribution $\mathsf{P}(x_{S^0}|y)$.

The whole procedure is shown in Table 1.

Table 1: Semi-iterative MPM estimation on the hybrid model

▲ upward sweep

Leaves ($i \in S^N$)

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

Recursion (for $n = N-1 \ldots 1$, $i \in S^n$)

$\mathbb{F}_i(x_{\overline{i}},y_{\underline{\underline{i}}}) \triangleq \sum_{x_i} h_i(y_i,x_i) f_i(x_i,x_{\overline{i}}) \prod_{j\in\underline{i}} \mathbb{F}_j(x_i,y_{\underline{\underline{j}}})$

◀▶ coarse posterior marginal computation:

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

$$\mathsf{P}(x_{S^0}|y) \propto \prod_{i\in S^0} h_i(x_i,y_i) \times \prod_{<i,j>\in S^0} g_{ij}(x_i,x_j) \times \prod_{i\in S^0}\prod_{j\in\underline{i}} \mathbb{F}_j(x_i,y_{\underline{\underline{j}}})$$

Approximation of

$$\mathsf{P}(x_i|y) \approx \frac{1}{m-k}\sum_{j=k+1}^{m} \delta(x_i(j)-x_i)$$

▼ downward sweep

Recursion (for $n = 1 \ldots N$, $i \in S^n$)

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

MPM at leaves

$\hat{x}_i \triangleq \arg\min_{x_i} \mathsf{P}(x_i|y)$

3 Supervised classification comparisons

We illustrate the relevance of our approach through comparative experiments for supervised classification. For this case study, we chose a Potts-type prior with

$$g_{ij}(x_i,x_j) = \exp\{-2^N\alpha[1-\delta(x_i,x_j)]\},$$

$$f_i(x_i,x_{\overline{i}}) = \exp\{-\beta[1-\delta(x_i,x_{\overline{i}})]\},$$

where δ is the Kroneker delta function, along with Gaussian likelihoods for data only located at the finest level:

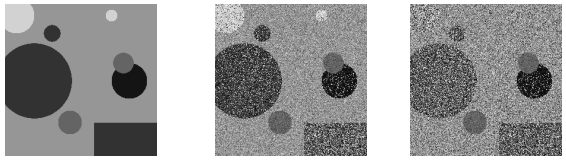
$$h_i(x_i=k,y_i) \triangleq \begin{cases} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right) & \text{if } i \in S^N, \\ 0 & \text{otherwise.}\end{cases}$$

Comparative experiments were led for $N \in \{0,2,3,4,p\}$ (if the size of S^N is $2^p \times 2^p$) with synthetic images and real satellite images. For $N = 0$

this is a standard non-hierarchical model using an iterative Gibbs sampler [8, 3], while $N = p$ corresponds to the complete tree ($|S^0| = 1$) allowing an exact non-iterative estimation of the posterior marginal probabilities [9].

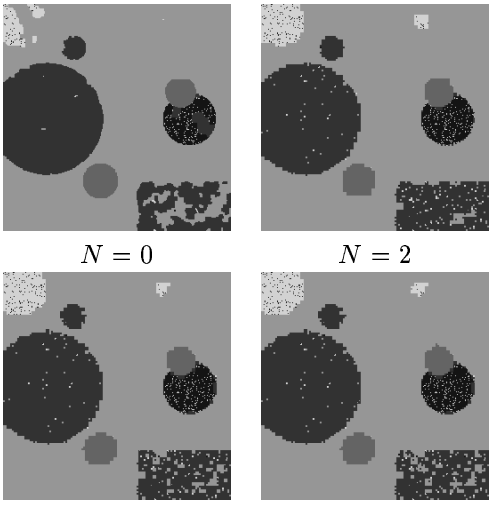
3.1 Synthetic images

We considered the problem of partitioning a 256×256 synthetic image into five classes (see Fig.2). An additive Gaussian white noise corrupts the image with a different standard deviation for each class. Thus the grey level means and variances $(\mu_k, \sigma_k^2)_{k=1}^5$ are known. We chose two levels of noise to study how our algorithms behave while the noise variances become higher.



synthetic image moderate noise larger noise

Figure 2: Synthetic data

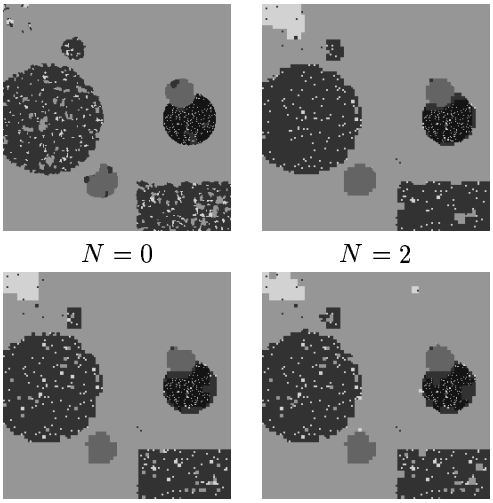


$N = 0$ $N = 2$

$N = 3$ $N = 8$

Model	misclassif.	cpu time
iterative $N = 0$	7%	7min
semi-iterative $N = 2$	3.6%	10.5s
semi-iterative $N = 3$	4.5%	9.3s
non-iterative $N = 8$	4.7%	9s

Figure 3: comparative results for the classification problem with a moderately corrupted image



$N = 0$ $N = 2$

$N = 3$ $N = 8$

Model	misclassif.	cpu time
iterative $N = 0$	10%	7min
semi-iterative $N = 2$	6.2%	10.5s
semi-iterative $N = 3$	7.9%	10s
non-iterative $N = 8$	8%	9s

Figure 4: comparative results for the classification problem with a more corrupted image

The results are shown in Fig.3 and 4 with their respective percentages of misclassification and cpu times in seconds (on a 170 MHz Ultra Sparc I workstation). In addition, based on the posterior marginals, MPM estimator can be complemented with a measure of relative confidence associated to the estimated value at each site (Fig.5). This measure can be, for instance, based on the entropy

$$c_i \triangleq -\sum_{x_i} \mathsf{P}(x_i|y) \log \mathsf{P}(x_i|y) \ ,$$

and allows us to better assess the quality of the obtained estimates and to use them in consequence.

As far as the visual aspect is concerned, the semi-iterative classifications still reveal a blocky aspect (as observed in all tree-based approaches). But these artifacts are less and less pronounced as the number of levels in the hybrid structure decreases.

Nevertheless, the two hierarchical methods provide much better results than the plain spatial MPM and take much fewer cpu times (about forty times fewer). For the Monte Carlo estimate of the posterior marginals, we fit the number of retained samples ($m - k$, see the description of the algorithm in Table 1) to the size of the concerned grid (200 for the

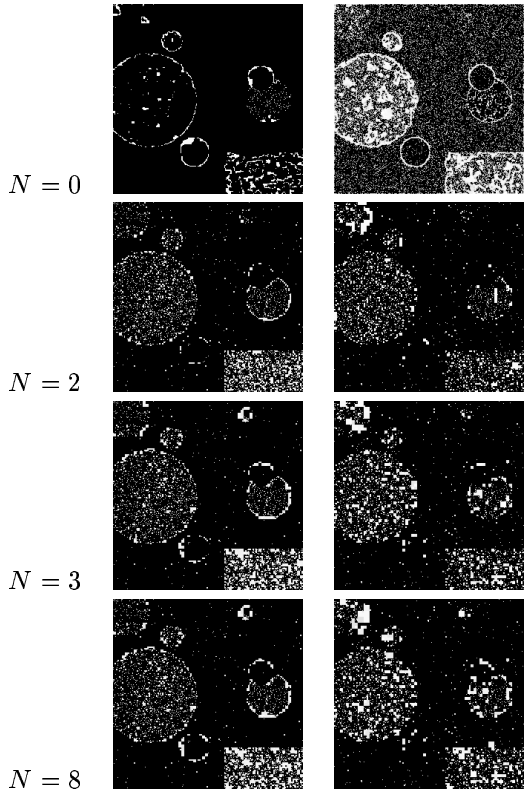


Figure 5: Confidence maps associated to classification of the moderately noisy image (left hand maps) and of the noisier image (right hand maps).

complete grid and 20 for the coarse grid of the hybrid structure). It can also be noticed that the hierarchical methods seem to manage better with the noisier image than the plain non-hierarchical iterative procedure. Thus the more corrupted class (clearer disks) has almost disappeared from the iterative MPM classification and the confidence map of this classification shows a noisier aspect than the others.

Finally, the semi-iterative estimation provides slightly better classification than the non-iterative one, for a comparable computational load but with a reduced memory requirement since the number of sites is lower. Thus the use of a sampling procedure, in order to compute the posterior marginals at the coarsest level, does not seem to imply a significant extra computational load, while improving the results.

3.2 Real satellite images

The previous algorithms were applied to SPOT satellite images provided by the Costel laboratory (University of Rennes 2) in the context of remote sensing researches within a research project called GSTB ("Groupement Scientifique pour le Développement de la Télédétection en Bretagne"). The scene (Fig.6) is composed of three 512×512 images with different wavelenghts in the visible spectrum and represents the Bay of Lannion, located in the north-west of France, in December 1996. The goal of this study was to determine the land cover of this area. So as to reach this aim, the geographers of Costel built a list of eight classification categories: Sea and water, Sand and bare soil, Urban area, Forests and heath, Temporary meadows, Permanent meadows, Colza, Vegetables.

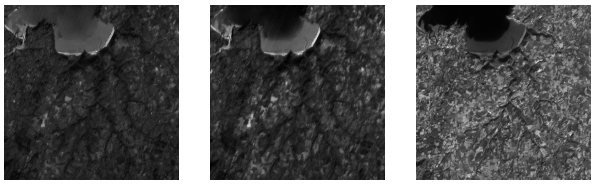


Figure 6: 512×512 Spot images (courtesy of Costel, University of Rennes 2 and GSTB).

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 divided this sample set in two sets: the first one was used to learn the gray level means and variances of each category for each image, so that we could perform supervised classifications, and the remaining samples were used to assess the accuracy of the classifications (the misclassification rates were thus calculated only on these latter samples).

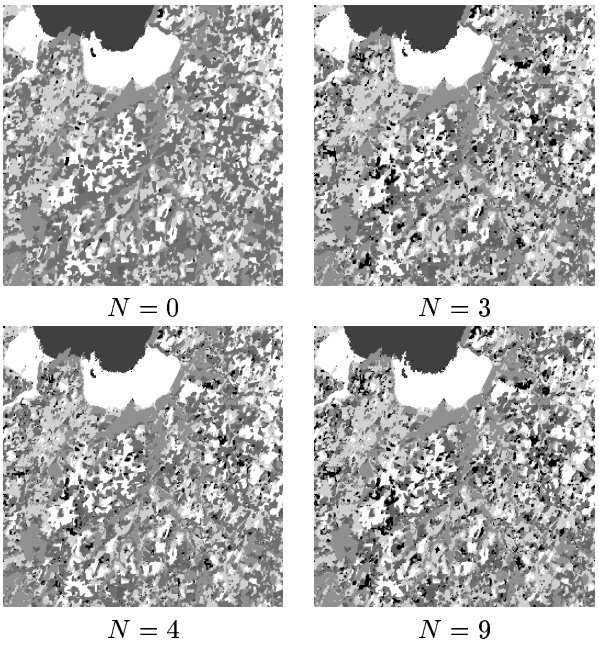
As for the model, we considered the three channels as independant. As a consequence, the form of the relation between the observed variables (here the multispectral scene) and the unknown variables (class label at each pixel) becomes:

$$h_i(x_i = k, y_i^c \; c \in \{1,2,3\}) \triangleq$$
$$\prod_{c=1}^{c=3} \frac{1}{\sqrt{2\pi\sigma_{k_c}^2}} \exp\left(- \frac{(y_i^c - \mu_{k_c})^2}{2\sigma_{k_c}^2}) \right) \quad \text{if } i \in S^N \; ,$$

where $(\mu_{k_c}, \sigma_{k_c})$ are the gray level mean and variance of the class k within channel number $c \in \{1,2,3\}$.

The four algorithms provide quite similar results of a good quality (see Fig.7). About 92% of the pixels of the samples are well classified. Nevertheless, the substancial advantage brought by the hierarchical methods in term of computational load is well put forward. More than one hour is required for the iterative MPM

to be performed, where the three hierarchical inferences only need few minutes (less than 3 minutes). This is all the more appealing that EM-type parameter estimation algorithms require at each iteration a complete computation of posterior marginals, as those involved in the presented MPM technique.



Model	misclassification	cpu time
iterative $N = 0$	8%	1h10min
semi-iterative $N = 3$	7.5%	165s
semi-iterative $N = 4$	7.5%	165s
non-iterative $N = 8$	7.8%	160s

Figure 7: Comparative results for the classification problem with multispectral spot images.

4 Extension

With the MAP estimator [4], we concluded that the hybrid struture was a good compromise for hierarchical image analysis. This is confirmed with the MPM estimator in this paper. As for the continuation of this work, we can now plan to deal with the critical problem of parameter estimation: using marginal computation introduced here, we can design specific EM-like algorithms on the hybrid structure [5], as already done with trees [2, 9] and with spatial grids [3]. We also plan to address the issue of automatically estimating the optimal number of levels in the struture.

References

[1] L.E. 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.*, Vol 41: pp. 164–171, 1970.

[2] C. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Trans. on Image Processing*, 3, No. 2 :162–177, March 1994.

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

[4] A. Chardin and P. Pérez. Semi-iterative inference with hierarchical models. In *Int. Conf. on Image Processing*, pages 630–634, Chicago, USA, October 1998.

[5] A. Chardin and P. Pérez. Unsupervised image classification with a hierarchical em algorithm. In *Proc. Int. Conf. on Computer Vision*, Kerkyra, Greece, September 1999.

[6] K.C. Chou, S.A. Golden, and A.S. Willsky. Multiresolution stochastic models, data fusion and wavelet transforms. *Signal Processing*, Vol. 34, No 3 : pages 257–282, Dec. 1993.

[7] P.W. Fieguth, W.C. Karl, A.S. Willsky, and C. Wunsch. Multiresolution optimal interpolation and statistical analysis of topex/poseidon satellite altimetry. *IEEE Trans. Geoscience and Remote Sensing*, 33(2):280–292, 1995.

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

[9] J-M Laferté, P. Pérez, and F. Heitz. Discrete markov image modeling and inference on the quad-tree. *IEEE Trans. Image Processing*, Accepted for publication, 1999.

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

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