A hierarchical Markov random field model and multi-temperature annealing for parallel image classification

Zoltan Kato, Marc Berthod, Josiane Zerubia

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Hierarchical Markov Random Field Model and Multi-Temperature Annealing for Parallel Image Classification

Zoltan KATO
Marc BERTHOD
Josiane ZERUBIA

N° 1938—version 2

Août 1993, version revisée Octobre 1994

PROGRAMME 4

Robotique,
image
et vision

ISSN 0249-6399

Champs de Markov hierarchiques et Recuit Multi-Température. Application à la classification d'image par algorithmes parallèles.

Zoltan KATO, Marc BERTHOD and Josiane ZERUBIA
INRIA - 2004 Route des Lucioles - BP 93
06902 Sophia Antipolis Cedex - FRANCE
Tel (33) 93 65 78 57 - Fax (33) 93 65 76 43
email: kato@sophia.inria.fr, berthod@sophia.inria.fr,
zerubia@sophia.inria.fr

Résumé

Dans ce rapport, nous nous intéressons à la classification d'image par algorithmes de relaxation multi-échelle mis en œuvre de façon massivement parallèle. Les techniques multi-grille sont bien connues pour améliorer nettement les taux de convergence ainsi que la qualité des résultats des techniques itératives de relaxation. Tout d'abord, nous présentons un modèle multi-échelle classique qui consiste à travailler sur une pyramide des étiquettes mais à conserver tout le champ d'observation. Le calcul des fonctions de potentiel aux grilles grossiéres est obtenu très simplement. L'optimisation est d'abord réalisée à une échelle grossiére grâce à une algorithme parallèle de relaxation, puis le niveau plus fin suivant est initialisé par la projection du résultat obtenu à l'échelle plus grossiére. Dans un deuxième temps, nous proposons un modèle Markovien hiérarchique construit à partir du modèle précédent. Nous introduisons des nouvelles interactions entre les niveaux voisins de la pyramide. Ceci permet de travailler avec des cliques dont les sites sont assez eloignés à un coût raisonable. Ce modèle conduit à un algorithme de relaxation utilisant un nouveau type de recuit: le Recuit Multi-Température. Il s'agit d'associer de hautes températures aux niveaux les plus grossiers, étant ainsi moins sensibles aux minima locaux. Nous avons prouvé la convergence de cet algorithme vers un optimum global en généralisant le théorème de Geman et Geman.

Mots Clefs

champs de Markov, multi-échelle, modèle hierarchique, algorithmes de relaxation, classification d'image supervisée.

Remerciements

Ce travail a reçu un soutien financier partiel du CNES, de l'AFIRST et du GdR TdSI (DRED/CNRS).

A Hierarchical Markov Random Field Model and Multi-Temperature Annealing for Parallel Image Classification

Zoltan KATO, Marc BERTHOD and Josiane ZERUBIA
INRIA - 2004 Route des Lucioles - BP 93
06902 Sophia Antipolis Cedex - FRANCE
Tel (33) 93 65 78 57 - Fax (33) 93 65 76 43
email: kato@sophia.inria.fr, berthod@sophia.inria.fr,
zerubia@sophia.inria.fr

Abstract

In this report, we are interested in massively parallel multiscale relaxation algorithms applied to image classification. It is well known that multigrid methods can improve significantly the convergence rate and the quality of the final results of iterative relaxation techniques. First, we present a classical multiscale model which consists of a label pyramid and a whole observation field. The potential functions of coarser grids are derived by simple computations. The optimization problem is first solved at the higher scale by a parallel relaxation algorithm, then the next lower scale is initialized by a projection of the result. Second, we propose a hierarchical Markov Random Field model based on this classical model. We introduce new interactions between neighbor levels in the pyramid. It can also be seen as a way to incorporate cliques with far apart sites for a reasonable price. This model results in a relaxation algorithm with a new annealing scheme: The Multi-Temperature Annealing (MTA) scheme, which consists of associating higher temperatures to higher levels, in order to be less sensitive to local minima at coarser grids. The convergence to the global optimum is proved by a generalisation of the annealing theorem of Geman and Geman.

Key Words

Markov Random Fields, multiscale, hierarchical model, relaxation algorithms, supervised image classification.

Acknowledgments

This work has been partially funded by CNES (French Aerospace Agency), AFIRST (French-Israeli Collaboration) and GdR TdSI (DRED/CNRS).

Contents

List of Figures

List of Tables

Notations

$\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ . . . . . Set of sites or pixels

$\mathcal{G} = \{\mathcal{G}_s \mid s \in \mathcal{S}\}$ . . . . . Neighborhood system over $\mathcal{S}$

$\mathcal{G}_s$ . . . . . Neighborhood of s

$C \subseteq \mathcal{S}$ . . . . . A clique

$\mathcal{C}$ . . . . . Set of cliques

C_s . . . . . Set of cliques containing s

$\deg(\mathcal{C})$ . . . . . Degree of the cliques

$\mathcal{L}$ . . . . . Lattice defined on $\mathcal{S}$ $(s = (i, j))$

$\mathcal{G}^n$ . . . . . Homogeneous neighborhood system of order n

$\mathcal{X} = \{X_s : s \in \mathcal{S}\}$ . . . . . MRF over $\mathcal{S}$

$\Lambda = \{0, 1, \ldots, L-1\}$ . . . . . Common state space for X_s

Ω . . . . . Set of all possible configurations

$\omega = (\omega_{s_1}, \ldots, \omega_{s_N}) : \omega_{s_i} \in \Lambda, 1 \leq i \leq N$ . . . . . Any configuration of $\mathcal{X}$

$\pi(\omega)$ . . . . . Gibbs distribution on Ω

Z . . . . . Partition function

T . . . . . Temperature

$U(\omega)$ . . . . . Energy function

$V_C(\omega)$ . . . . . Potential of clique C

$\mathcal{F} = \{f_s : s \in \mathcal{S}\}$ . . . . . Observed data

$\{n_k, k = 1, 2, \ldots\}$ . . . . . Updating order of the sites

$W = w^n$ . . . . . Width of the lattice $\mathcal{L}$

$H = h^m$ . . . . . Height of the lattice $\mathcal{L}$

$M = \inf(n, m)$ . . . . . The highest level of the pyramid (we have $M + 1$ levels)

$\mathcal{B}^i = \{b_1^i, \ldots, b_{N_i}^i\}$ . . . . . i^{th} scale

b_k^i . . . . . k^{th} block at the i^{th} scale

$N_i = N/(wh)^i$ . . . . . Number of blocks at the i^{th} scale

$\omega_k^i \in \Lambda$ . . . . . Common label of the block b_k^i

Ω_i . . . . . Configuration space at the i^{th} scale

C_j^i . . . . . A clique of order j at scale i

$\mathcal{C}^i$ . . . . . Set of cliques at scale i

$\mathcal{D}_{C_j^i}$ . . . . . Set of cliques included in the clique C_j^i at scale i

$\mathcal{A}_j^i$ . . . . . Set of cliques included in any clique of order j at scale i

$V_{C_j^i}^{\mathcal{B}^i}$ . . . . . Potential of the clique C_j^i at scale i

$\mathcal{S}^i$ . . . . . Grid at the i^{th} level of the pyramid

$\Xi_i = \{\xi_s^i : s \in \mathcal{S}^i, \xi_s^i \in \Lambda\}$ . . . . . Configuration space of the i^{th} level of the pyramid

Φ^i . . . . . Isomorphism between $\mathcal{S}^i$ and $\mathcal{B}^i$

$U^i(\xi^i)$ . . . . . Energy function at level i

$V_{C^i}^i(\xi^i)$ . . . . . Clique-potentials at level i

$\bar{\mathcal{S}} = \{\bar{s}_1, \ldots, \bar{s}_{\bar{N}}\}$ . . . . . Set of sites of the pyramid

$\bar{N}$ . . . . . Number of sites of the pyramid

$\bar{\Omega}$ . . . . . Configuration space of the pyramid

$\bar{\omega}$ . . . . . A configuration of the pyramid

Ψ . . . . . Projection-function between two neighbor levels

$\bar{\mathcal{G}}$ . . . . . Neighborhood system on the pyramid

$\mathcal{G}_i$ . . . . . Neighborhood system at level i

$\bar{\mathcal{C}}$ . . . . . Set of cliques over the pyramid

$\mathcal{C}^*$ . . . . . Set of the cliques located between two neighbor levels

$\bar{\mathcal{X}}$ . . . . . MRF over the pyramid

$\bar{U}(\bar{\omega})$ . . . . . Energy function on the pyramid

$\bar{V}_{\bar{C}}(\omega)$ . . . . . Clique-potentials on the pyramid

$U^*(\bar{\omega})$ . . . . . Energy over the cliques $\mathcal{C}^*$

$P_{\omega,\eta}(k-1,k)$ . . . . . Probability of the k^{th} transition $\omega \to \eta$

$X(k)$ $(k = 1, 2, \ldots)$ . . . . Markov chain generated by the Simulated Annealing algorithm

$P_{\omega,\eta}(T)$ . . . . . Transition matrix

$G_{\omega,\eta}(T)$ . . . . . Generation matrix

$A_{\omega,\eta}(T)$ . . . . . Acceptance matrix

Ω_{opt} . . . . . Set of globally optimal configurations

$T(k,C)$ . . . . . Temperature function depending on the iteration k and on the cliques C

$\pi_{T(k,C)}(\omega)$ . . . . . Gibbs distribution with temperature T(k,C)

$\oslash$ . . . . . The operation $\sum_{C \in \mathcal{C}} \frac{V_C(\omega)}{T(k,C)}$

π_0 . . . . . Uniform distribution on Ω_{opt}

U^{sup} . . . . . Maximum value of the energy function $U(\omega)$

U^{inf} . . . . . Minimum value of the energy function $U(\omega)$

Δ . . . . . Difference between the maximum and minimum of $U(\omega)$

T_k^{inf} . . . . . Minimum of $T(k, C)$ at the k^{th} iteration

μ_λ . . . . . Mean value of class $\lambda \in \Lambda$

σ_λ . . . . . Deviation of class $\lambda \in \Lambda$

β . . . . . Second order clique-potential at the finest level

γ . . . . . Second order clique-potential between neighbor levels

$P(k, \omega | l, \eta)$ . . . . . The same as the transition probability $P(X(k) = \omega | X(l) = \eta)$

$P(k, \omega | l, \mu)$ . . . . . The same as $\sum_\eta P(X(k) = \omega | X(l) = \eta)\mu(\eta)$

$P(k, \cdot | l, \mu)$ . . . . . The "$\cdot$" means any configuration here

$\|\mu - \nu\|$ . . . . . The L^1 norm of two distributions on Ω

T_k^{inf} . . . . . Minimum of $T(k, C)$ at the k^{th} iteration

1 Introduction

Markov Random Fields (MRF) have become more and more popular during the last few years in image processing [1, 7, 10, 12, 14, 30]. A good reason for that is that such a modelization is the one which requires the less a priori information on the world model. On the other hand, the local behavior of MRF permits to develop highly parallel algorithms in the resolution of the combinatorial optimization problem associated with such a model.

In this report, we are interested in massively parallel multiscale relaxation algorithms applied to image classification [8, 9, 16, 26, 28]. It is well known that multigrid methods can improve significantly the convergence rate and the quality of the final results of iterative relaxation techniques.

There are many approaches in multigrid image segmentation. F. Marques et al. [26] propose a hierarchical Compound Gauss-Markov Random Field model with a label pyramid and an observation pyramid. Bouman [8, 9] proposes a multiscale MRF model, where each scale is causally dependent on the coarser grid field above it. This model yields to a non-iterative segmentation algorithm and direct methods of parameter estimation. The basis of our approach is a consistent multiscale MRF model proposed by F. Heitz et al. in [16, 28] for motion analysis. This model consists of a label pyramid and a whole observation field. The original energy function can be decomposed as a sum of potential functions which are defined on neighbor blocks and only depend on the labels associated with these blocks and on the observation field. Using this decomposition, the parameters of coarser grids can be computed very easily. This model results in a multigrid relaxation scheme which replaces the original optimization problem by a sequence of more tractable problems. Using a top down strategy in the label pyramid, the optimization problem is first solved at a higher level, then the lower grid is initialized with the previous result by a simple projection. This algorithm is very efficient in the case of deterministic relaxation (for instance ICM [4, 18]) which gets stuck in a local minimum near the starting configuration. In the case of stochastic relaxation (for instance Simulated Annealing [13, 24, 27]), which are far less dependent on the initial configuration, the results are only slightly better, but the method is still interesting with respect to computer time, especially on a sequential machine. After a brief introduction to the theory of Markov Random Fields (Section 2), we give a general description of this model and the relaxation scheme associated with it in the Section 3.

Then, we propose a new hierarchical MRF model defined on the whole label pyramid (Section 4). In this model, we have introduced a new interaction scheme between neighboring levels in the pyramid yielding a better communication between the grids. It can also be seen as a way to incorporate cliques with far apart sites for a reasonable price. This model gives a relaxation algorithm with a new annealing scheme which can be run in parallel on the entire pyramid. The basic idea of this annealing scheme, which we propose to call Multi-Temperature Annealing (MTA) is the following: to the higher levels, we associate higher temperatures which enable the algorithm to be less sensitive to local minima. However at a finer resolution, the relaxation is performed at a lower temperature (at the bottom level, it is closed to 0). The complete convergence study of the relaxation algorithm in the case of a homogeneous, inhomogeneous and Multi-Temperature Annealing schedule can be found in Section 5. In the multi-temperature case, our annealing theorem is a generalisation of the well known theorem of Geman and Geman [13] and the proof can be found in Appendix A.

In Section 6, we apply these models to supervised image classification. Using a first order MRF model to take into account the context and a Gaussian representation of the classes, we define the energy function for the monogrid, multiscale and hierarchical models.

Finally, experiments are shown in Section 7 with the Gibbs sampler [13] and the Iterated Conditional Mode [4, 18] using the three models for each algorithm (monogrid, multiscale and hierarchical). These methods have been implemented in parallel on a Connection Machine CM200 [17].

2 Markov Random Fields

First, we briefly give an introduction to the theory of Markov Random Fields (MRF) [1, 29], then we describe a general image model used in the following sections. Finally, we recall a few classical relaxation algorithms used for the optimization of the cost function of the model.

2.1 Neighborhood Systems

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ be a set of sites.

Definition 2.1 (Neighborhood system) *$\mathcal{G} = \{\mathcal{G}_s \mid s \in \mathcal{S}\}$ is a neighborhood system for $\mathcal{S}$ if*

1. $s \notin \mathcal{G}_s$
2. $s \in \mathcal{G}_r \Leftrightarrow r \in \mathcal{G}_s$

Definition 2.2 (Clique) *A subset $C \subseteq \mathcal{S}$ is a clique if every pair of distinct sites in C are neighbors. $\mathcal{C}$ denotes the set of cliques and $\deg(\mathcal{C}) = \max_{C \in \mathcal{C}} \mid C \mid$.*

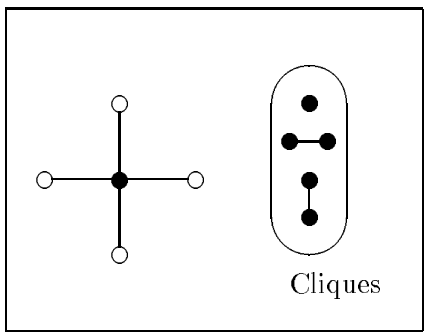

Figure 1: *First order neighborhood-system with cliques*

The most commonly used neighborhood systems are the homogeneous systems. In this case, we consider $\mathcal{S}$ as a lattice $\mathcal{L}$ and define these neighborhoods as

$$\begin{aligned} \mathcal{G}^n &= \{\mathcal{G}^n_{(i,j)} : (i,j) \in \mathcal{L}\}, \\ \mathcal{G}^n_{(i,j)} &= \{(k,l) \in \mathcal{L} : (k-i)^2 + (l-j)^2 \leq n\}. \end{aligned}$$

Obviously, sites near the boundary have fewer neighbors than interior ones. Furthermore, $\mathcal{G}^0 \equiv \mathcal{S}$ and for all $n \geq 0 : \mathcal{G}^n \subset \mathcal{G}^{n+1}$. Figure 1 shows a first-order neighborhood corresponding to $n = 1$. The cliques are $\{(i,j)\}, \{(i,j),(i,j+1)\}, \{(i,j),(i+1,j)\}$.

2.2 Gibbs Distribution and MRF's

Let $\mathcal{X} = \{X_s : s \in \mathcal{S}\}$ denotes any family of random variables so that $\forall s \in \mathcal{S} : X_s \in \Lambda$, where $\Lambda = \{0, 1, \ldots, L-1\}$ is a common state space. Let $\Omega = \{\omega = (\omega_{s_1}, \ldots, \omega_{s_N}) : \omega_{s_i} \in \Lambda, 1 \leq i \leq N\}$ be the set of all possible configurations.

Definition 2.3 (Markov Random Field) *$\mathcal{X}$ is a Markov Random Field (MRF) with respect to $\mathcal{G}$ if*

1. *for all $\omega \in \Omega$: $P(\mathcal{X} = \omega) > 0$,*
2. *for every $s \in \mathcal{S}$ and $\omega \in \Omega$:*
   $P(X_s = \omega_s \mid X_r = \omega_r, r \neq s) = P(X_s = \omega_s \mid X_r = \omega_r, r \in \mathcal{G}_s)$.

The functions in 2. are called the *local characteristics* of the MRF, and the probability distribution $P(\mathcal{X} = \omega)$ of any process satisfying 1. is uniquely determined by these conditional probabilities. However, it is extremely difficult to determine these characteristics in practice.

Definition 2.4 (Gibbs distribution) *A Gibbs distribution relative to the neighborhood system $\mathcal{G}$ is a probability measure π on Ω with the following representation:*

$$\pi(\omega) = \frac{1}{Z} \exp\left(\frac{-U(\omega)}{T}\right), \tag{1}$$

where Z is the normalizing constant or partition function*:*

$$Z = \sum_{\omega} \exp\left(\frac{-U(\omega)}{T}\right),$$

T is a constant called the temperature *and the* energy function *U is of the form*

$$U(\omega) = \sum_{C \in \mathcal{C}} V_C(\omega). \tag{2}$$

Each V_C is a function defined on Ω depending only on those elements ω_s of ω for which $s \in C$. Such a function is called a potential.

One of the most important theorem is probably the *Hammersley-Clifford theorem* [1] which points out the relation between MRF and Gibbs distribution:

Theorem 2.1 (Hammersley-Clifford) *$\mathcal{X}$ is a MRF with respect to the neighborhood system $\mathcal{G}$ if and only if $\pi(\omega) = P(\mathcal{X} = \omega)$ is a Gibbs distribution with respect to $\mathcal{G}$.*

The main benefit of this equivalence is that it provides us a simple way to specify MRF's, namely specifying potentials instead of local characteristics (see definition 2.3), which is usually very difficult.

2.3 A General Markov Image Model

We now look at the image labeling model. Image labeling is a general framework to solve low level vision tasks, such as image classification, edge detection, etc... To each pixel of the image, we assign a label. The meaning of the labels depends on the task that we want to solve. For image classification, for example, a label means a class; for edge detection, it means the presence or the direction of an edge; etc... Thus, we have the following general problem:

We are given a set of pixels (an image) $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ with some neighborhood system $\mathcal{G} = \{\mathcal{G}_s : s \in \mathcal{S}\}$ and $\mathcal{F} = \{f_s : s \in \mathcal{S}\}$ a set of image data (or observations). Each of these pixels may take a label from $\Lambda = \{0, 1, \ldots, L-1\}$. The configuration space Ω is the set of all global discrete labeling $\omega = (\omega_{s_1}, \ldots, \omega_{s_N}), \omega_s \in \Lambda$. We assume that $\mathcal{X}$ is a MRF relative to $\mathcal{G}$ with a corresponding energy function U_2 and potentials $\{V_C\}$:

$$\begin{aligned} P(\mathcal{X} = \omega) &= \frac{1}{Z} \exp\left(\frac{-U_2(\omega)}{T}\right) \\ U_2(\omega) &= \sum_{C \in \mathcal{C}} V_C(\omega) \end{aligned}$$

Now, we will construct a Bayesian estimator to find the optimal labeling, that is the labeling which maximizes the posterior distribution $P(\mathcal{X} = \omega \mid \mathcal{F})$ of the label field:

$$P(\mathcal{X} = \omega \mid \mathcal{F}) = \frac{P(\mathcal{F} \mid \mathcal{X} = \omega) P(\mathcal{X} = \omega)}{P(\mathcal{F})} \tag{3}$$

Since $P(\mathcal{F})$ is constant, the MAP estimator of the label field is given by:

$$\max_{\omega \in \Omega} P(\mathcal{X} = \omega \mid \mathcal{F}) = \max_{\omega \in \Omega} P(\mathcal{F} \mid \mathcal{X} = \omega) P(\mathcal{X} = \omega). \tag{4}$$

If we assume that the observed image $\mathcal{F}$ is affected at site s only by the pixel s itself (i.e. the image is not blurred), one can prove, that $P(\mathcal{F} \mid \mathcal{X} = \omega)$ is a Gibbs distribution over $\mathcal{G}^0 \equiv \mathcal{S}$ with an energy function U_1 and potentials $V_{\{s\}}$ (a blurred image model is studied, for example, in [13]). Thus, the posterior distribution is also a MRF over $\mathcal{G}$ with the following energy function:

$$\begin{aligned} U(\omega) &= U_1(\omega) + U_2(\omega) \text{ where} && (5) \\ U_1(\omega) &= \sum_{s \in \mathcal{S}} V_{\{s\}}(\omega_s) \text{ and} && (6) \\ U_2(\omega) &= \sum_{C \in \mathcal{C}} V_C(\omega_C) && (7) \end{aligned}$$

Using this function, the MAP estimator is given by:

$$\begin{aligned} \hat{\omega} &= \arg\max_{\omega \in \Omega} P(\mathcal{X} = \omega \mid \mathcal{F}) = \arg\max_{\omega \in \Omega} \frac{1}{Z} \exp\left(\frac{-U(\omega)}{T}\right) \\ &= \arg\min_{\omega \in \Omega} U(\omega). \end{aligned} \tag{8}$$