Unsupervised Parallel Image Classification Using a Hierarchical Markovian Model

Zoltan Kato, Josiane Zerubia, Marc Berthod

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Unsupervised Parallel Image Classification Using a Hierarchical Markovian Model

Zoltan Kato, Josiane Zerubia, Marc Berthod

N° 2528

Avril 1995

PROGRAMME 4

ISSN 0249-6399

Unsupervised Parallel Image Classification Using a Hierarchical Markovian Model*

Zoltan Kato, Josiane Zerubia, Marc Berthod

Programme 4 — Robotique, image et vision
Projet PASTIS

Rapport de recherche n°2528 — Avril 1995 — 34 pages

Abstract: This paper deals with the problem of unsupervised classification of images modeled by Markov Random Fields (MRF). If the model parameters are known then we have various methods to solve the segmentation problem (simulated annealing, ICM, etc...). However, when they are not known, the problem becomes more difficult. One has to estimate the hidden label field parameters from the only observable image. Our approach consists of extending a recent iterative method of estimation, called Iterative Conditional Estimation (ICE) to a hierarchical markovian model. The idea resembles the Estimation-Maximization (EM) algorithm as we recursively look at the Maximum a Posteriori (MAP) estimate of the label field given the estimated parameters then we look at the Maximum Likelihood (ML) estimate of the parameters given a tentative labeling obtained at the previous step. We propose unsupervised image classification algorithms using a monogrid or a hierarchical model. The only parameter supposed to be known is the number of regions, all the other parameters are estimated. The presented algorithms have been implemented on a Connection Machine CM200. Comparative tests have been done on noisy synthetic and real images.

Key-words: hierarchical Markovian model, parameter estimation, unsupervised image classification.

(Résumé : tsvp)

*Acknowledgements: this work has been partially funded by CNES, AFIRST and DRED/GdR-TdSI.

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

Classification Non-Supervisée d'Images par Champs de Markov Hiérarchiques

Résumé : Ce rapport est relatif à la classification non-supervisée d'images, modélisées par des champs de Markov. Si les paramètres du modèle sont connus, alors plusieurs méthodes sont disponibles afin de segmenter l'image (recuit simulé, ICM, etc...). Cependant, lorsque ces paramètres sont inconnus, le problème devient plus complexe. Il faut alors estimer les paramètres du champ de Markov caché qui correspond aux étiquettes à partir de la seule observation. L'approche proposée est l'extension d'une méthode d'estimation récente, l'ICE, au cas d'un modèle markovien hiérarchique. L'idée de base est proche de la technique EM puisqu'il s'agit de l'estimation récursive du champ d'étiquette au sens du MAP les paramètres du modèle étant supposés connus, puis d'une estimation des paramètres au sens du maximum de vraisemblance à partir d'un étiquetage obtenu à l'itération précédente. Deux algorithmes (monogrille et hiérarchique) de classification non-supervisée sont décrits, la seule connaissance a priori étant le nombre de classes. Ces techniques ont été mises en œuvre sur une Connection Machine CM200. Des tests comparatifs sont présentés sur des images synthétiques et réelles.

Mots-clé : modèle markovien hiérarchique, estimation des paramètres, classification d'images non-supervisée

Table of Contents

List of Figures

Liste of Tables

1 Introduction

In real life applications, the model parameters are usually unknown, one has to estimate [2, 7] them only from the observable image. From a statistical viewpoint, this means that we want to estimate parameters from random variables whose joint distribution is a mixture of distributions. If we have a realization of the label field then the problem is relatively easy, we have many standard methods to do parameter estimation (Maximum Likelihood, Coding method [4], etc...). Unfortunately, such a realization is not known, so the direct use of such estimation algorithms is impossible. We have to approximate it by some function of the image data, which is the only observable attribute.

Some nowadays used algorithms are iterative [19, 18, 6], subsequently generating a labeling, estimating parameters from it, then generating a new labeling using these parameters, etc ...For such a method, we need a reasonably good initial value for each parameter. Since the classes of a labeling problem are mostly represented by a Gaussian distribution, the initialization of the mean and the variance of each class is very important because of its influence on subsequent labelings and hence on the final estimates. On the other hand, it is a classical problem, namely the determination of the modes of a Gaussian mixture without any a priori information. There are many approaches in this domain: Method of moments [10], Prony's Method [9] or geometrical analysis of the histogram [20], for instance.

Herein, we will present new parameter estimation methods applied to monogrid and hierarchical MRF models. The proposed algorithms have been tested on image segmentation problems. Comparative test have been done on noisy synthetic and real satellite images.

2 The Parameter Estimation Problem

Let us briefly review some notations. $\mathcal{F} = \{F_s : s \in \mathcal{S}\}$ denotes a set of image data on the sites (or pixels) $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$. Furthermore, each of these sites may take a label from $\Lambda = \{0, 1, \ldots, L-1\}$. The configuration space Ω is the set of all global discrete labelings $\omega = (\omega_{s_1}, \ldots, \omega_{s_N}), \omega_s \in \Lambda$. The label process is denoted by $\mathcal{X}$.

In parameter estimation problems, $\mathcal{F}$ is also called the *observed image* and $\mathcal{X}$ denotes the *unobserved* image attributes (labels). Furthermore, we are given n parameters forming a vector Θ which appears in the MRF model:

$$\Theta = \begin{pmatrix} \vartheta_1 \\ \vdots \\ \vartheta_n \end{pmatrix} \tag{1}$$

Usually, Θ is considered to be known. Therefore, one is looking for the labeling which maximizes the a posteriori distribution:

$$\widehat{\omega} = \arg\max_{\omega \in \Omega} P_\Theta(\omega \mid \mathcal{F}, \Theta). \tag{2}$$

where $\widehat{\omega}$ is the MAP estimate of the label field, given $\mathcal{F}$, under the model P_Θ (in the followings, the index Θ will be omitted). If both Θ and ω are unknown, the maximization problem in Equ. 2 becomes [11, 16]:

$$(\widehat{\omega}, \widehat{\Theta}) = \arg\max_{\omega, \Theta} P(\omega, \mathcal{F} \mid \Theta). \tag{3}$$

The pair $(\widehat{\omega}, \widehat{\Theta})$ is the global maximum of the joint probability $P(\omega, \mathcal{F} \mid \Theta)$. If we regard Θ as a random variable, the above maximization is an ordinary MAP estimation in the following way: Let us suppose, that Θ is restricted to a finite volume domain $\mathcal{D}_\Theta$ and suppose that Θ is uniform on $\mathcal{D}_\Theta$ (that is $P(\Theta)$ is constant). Then, we get [11]:

$$\begin{aligned} \arg\max_{\omega, \Theta} P(\omega, \Theta \mid \mathcal{F}) &= \arg\max_{\omega, \Theta} \frac{P(\omega, \mathcal{F} \mid \Theta) P(\Theta)}{P(\mathcal{F})} && (4) \\ &= \arg\max_{\omega, \Theta} \frac{P(\omega, \mathcal{F} \mid \Theta)}{\int_{\mathcal{D}_\Theta} \sum_{\omega \in \Omega} P(\omega, \mathcal{F} \mid \Theta) d\Theta} && (5) \\ &= \arg\max_{\omega, \Theta} P(\omega, \mathcal{F} \mid \Theta). && (6) \end{aligned}$$

However, this maximization is very difficult, having no direct solution. Even Simulated Annealing (SA) is not implementable because the local characteristics with respect to the parameters Θ cannot be computed from $P(\omega, \mathcal{F} \mid \Theta)$. One possible solution is to adopt the following criterion instead [11, 16]:

$$\begin{aligned} \widehat{\omega} &= \arg\max_{\omega} P(\omega, \mathcal{F} \mid \widehat{\Theta}) && (7) \\ \widehat{\Theta} &= \arg\max_{\Theta} P(\widehat{\omega}, \mathcal{F} \mid \Theta) && (8) \end{aligned}$$

Clearly, Equ. 7 is equivalent to Equ. 3 for $\Theta = \widehat{\Theta}$ and Equ. 8 is equivalent to Equ. 3 with $\omega = \widehat{\omega}$. Furthermore, Equ. 7 is equivalent to the MAP estimate of ω in the case of known parameters:

$$\arg\max_{\omega} P(\omega, \mathcal{F} \mid \widehat{\Theta}) = \arg\max_{\omega} P(\omega \mid \mathcal{F}, \widehat{\Theta}) P(\mathcal{F} \mid \widehat{\Theta}) = \arg\max_{\omega} P(\omega \mid \mathcal{F}, \widehat{\Theta}). \quad (9)$$

3 Parameter Estimation from Incomplete Data

In real life applications, labeled samples are usually not available. We have to estimate the parameters from an *unlabeled* sample. In statistics, the problem is known as the *incomplete data* problem. A broadly applicable algorithm has been proposed by Dempster *et al.* [8], called *Expectation – Maximization* (EM). The algorithm aims at determining the ML estimate of the parameters Θ by making use of the estimation of the missing data (i.e. the label field $\mathcal{X}$). Hereafter, we are also describing a few other estimation methods [11, 18, 19] available when dealing with incomplete data.

3.1 Adaptive Simulated Annealing (ASA)

Another EM-like algorithm has been proposed by Geman in [11], which is called *Adaptive Simulated Annealing* (ASA). The algorithm was adapted to image segmentation problems in [16], where the convergence of ASA has also been proved. The **ASA algorithm** is very similar to the SEM, it may be seen as a special case where the S-step is implemented by a Simulated Annealing:

- step 1. Set $k = 0$ and initialize $\widehat{\Theta}^0$.
- step 2. Do n iterations ($n \geq 1$) of SA sampling from $P(\omega \mid \mathcal{F}, \widehat{\Theta}^k)$. The resulting labeling is denoted by $\widehat{\omega}^{k+1}$.
- step 3. Update the current estimate of the parameters, $\widehat{\Theta}^{k+1}$ to the ML estimate based on the current labeling $\widehat{\omega}^{k+1}$.
- step 4. Goto step 2. with $k = k + 1$ until $\widehat{\Theta}$ stabilizes.

If ML estimate is not tractable, which is often the case when dealing with MRF models, one can use an approximation (Maximum Pseudo Likelihood–MPL, for instance). We remark that a similar algorithm has been reported in [4]. It uses ICM instead of SA in step 2.

3.2 Iterative Conditional Estimation (ICE)

Finally, we present a solution to the incomplete data problem originally proposed by Pieczynski *et al.* [19, 5, 1]. Let us consider an estimator $\mathcal{E}_\Theta(\mathcal{F},\omega)$ of Θ (ML, for instance). Since realizations of the label field are unknown, the direct use of $\mathcal{E}_\Theta(\mathcal{F},\omega)$ is impossible, we have to approximate it. The best approximation, in the mean-square sense, is the *conditional expectation*. Since $E\{\mathcal{E}_\Theta \mid \mathcal{F},\omega\}$ depends on the parameters Θ, we need a parameter $\widehat{\Theta}^k$ previously defined by some way. This defines an iterative procedure, called **ICE** [19, 5]:

- step 1. Set $k=0$ and initialize $\widehat{\Theta}^0$.
- step 2. Generate n realizations (n is fixed a priori) $\widehat{\omega}^i (1 \leq i \leq n)$ of the label field based on $\widehat{\Theta}^k$.
- step 3. Based on the sample $\widehat{\omega}^i (1 \leq i \leq n)$, $\widehat{\Theta}^{k+1}$ is obtained as the conditional expectation:

  $$\widehat{\Theta}^{k+1} = E\{\mathcal{E}_\Theta \mid \mathcal{X} = \omega\} \approx \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathcal{E}_\Theta(\mathcal{F}, \widehat{\omega}^i).$$

- step 4. Goto step 2. until $\widehat{\Theta}$ stabilizes.

Compared to the EM algorithm, ICE results in a better estimate of the parameters and the convergence is faster [5, 1].

4 Gaussian Mixture Identification (GMI)

In image labeling problems, the classes are often modeled by a Gaussian distribution. What we observe is a *mixture* of Gaussian distributions and we have to determine the modes of this mixture corresponding to the classes. The problem is to estimate the parameters of these modes from an *unlabeled* sample. Herein, we are interested in non-iterative methods.

In [20], a geometrical analysis of the mixture density function is used: First, let us consider a n-dimensional normal density function $f(\vec{x})$ and review the relationships between the geometrical characteristics of the concave domain of $f(\vec{x})$ and its mean vector $\vec{\mu}$ and covariance matrix Σ. Considering a zero-mean density,

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{x}\right). \tag{10}$$