Bayesian Image Segmentation Using Wavelet-Based Priors

Mário A. T. Figueiredo

Instituto de Telecomunicações, and
Department of Electrical and Computer Engineering, *Instituto Superior Técnico*
1049-001 Lisboa, **Portugal**, mario.figueiredo@lx.it.pt

Abstract

This paper introduces a formulation which allows using wavelet-based priors for image segmentation. This formulation can be used in supervised, unsupervised, or semi-supervised modes, and with any probabilistic observation model (intensity, multispectral, texture). Our main goal is to exploit the well-known ability of wavelet-based priors to model piece-wise smoothness (which underlies state-of-the-art methods for denoising, coding, and restoration) and the availability of fast algorithms for wavelet-based processing.

The main obstacle to using wavelet-based priors for segmentation is that they're aimed at representing real values, rather than discrete labels, as needed for segmentation. This difficulty is sidestepped by the introduction of real-valued hidden fields, to which the labels are probabilistically related. These hidden fields, being unconstrained and real-valued, can be given any type of spatial prior, such as one based on wavelets. Under this model, Bayesian MAP segmentation is carried out by a (generalized) EM algorithm. Experiments on synthetic and real data testify for the adequacy of the approach.

1. Introduction

Image segmentation has been one of the core problems in computer vision. Although remarkable success has been obtained in specific domains with clear goals (*e.g.*, medical imaging), a general purpose segmentation criterion is an elusive concept and many different approaches, formulations, and tools have been proposed. Most methods work by combining evidence from the observed image (via image features/cues) with some form of prior (or regularization) that embodies the concept of "*a priori* probable" segmentation. All the research on segmentation has concentrated on one (or both) of the following fronts:
(a) Development of image features, and feature models, as relevant and informative as possible for segmentation. Classical examples for texture segmentation include Gabor features [10], wavelets-based features [4], [21], Markov random field models [5], [6]. See [19], for a survey of the vast literature on texture features. Some recent work combines intensity, texture, and contour features, with the goal of mimicking human image segmentation [16]. There are many other features developed for specific domains, such as color images, medical images, or remote sensing images.
(b) Development of methods that enforce some form of spatial coherence, *i.e.*, that integrate local cues into a globally coherent segmentation. The recent graph-based methods [20], [22], [24], achieve this by formulating image segmentation as a graph partitioning problem. In Bayesian approaches, spatial coherence is usually imposed by a Markov random field (MRF) prior (see [13] and references therein).

This paper falls in the second work front: it describes a new way of including (spatial) priors in image segmentation, and illustrates it by showing how wavelet-based models (so successful in image denoising, coding, and restoration [14]) can be used in this context. The proposed formulation, which is based on seeing segmentation as spatially regularized logistic regression, is general and can be used in supervised, unsupervised, or semi-supervised modes, as well as with generative or discriminative features.

The key difficulty in using wavelet-based priors (or other priors for real-valued fields/images) for segmentation, is that these priors (unlike MRFs) are not suited for the categorical-type variables (the region labels) involved in image segmentation. This issue is sidestepped by using an approach which is common in machine learning: introduction of real-valued hidden fields, to which the labels are probabilistically related. This approach is used, *e.g.*, in the very successful methods for regression and classification known as "Gaussian processes" [23]. These hidden fields, being real-valued, can be given any type of spatial prior. In this paper, wavelet-based priors are adopted, aiming at exploiting their well-known ability to encode preference for piece-wise smoothness and the availability of fast algorithms for wavelet-based image processing. We show how the approach can be used in supervised, unsupervised, and semi-supervised modes, by presenting expectation-maximization (EM) algorithms for the three cases. In the supervised case, the resulting segmentation criterion consists in minimizing a convex function, thus initialization problems do not arise, unlike in MRF-based methods.

2. Formulation

2.1 Images and Segmentations

Let $\mathcal{L} = \{(n,m),\ n = 1,...,N,\ m = 1,...,M\}$ be a $2D$ lattice of sites/pixels on which observed images, and their segmentations, are defined. An observed image $\mathbf{x}$ is a set of (maybe vector valued) observations, indexed by the lattice $\mathcal{L}$, that is $\mathbf{x} = \{x_i \in I\!R^d,\ i \in \mathcal{L}\}$. A segmentation $\mathcal{R} = \{R_k \subseteq \mathcal{L},\ k = 0,...,K-1\}$ is a partition of $\mathcal{L}$ into K regions, in an exhaustive and mutually exclusive way:

$$\bigcup_{k=0}^{K-1} R_k = \mathcal{L} \quad \text{and} \quad \left(R_j \bigcap R_k = \emptyset\right) \Leftarrow (j \neq k).$$

In the sequel, it will be convenient to represent partitions by a set of binary labels $\mathbf{y} = \{y_i = [y_i^{(0)}, ..., y_i^{(K-1)}],\ i \in \mathcal{L}\}$, where $y_i^{(k)} \in \{0,1\}$, such that $(y_i^{(k)} = 1) \Leftrightarrow i \in R_k$.

2.2. Observation Model

Given a segmentation $\mathcal{R}$, we assume that the observations are independently distributed, that is,

$$p(\mathbf{x}|\mathcal{R}) = p(\mathbf{x}|\mathbf{y}) = \prod_{k=0}^{K-1} \prod_{i \in R_k} p(x_i|\boldsymbol{\phi}_k), \qquad (1)$$

where the $p(\cdot|\boldsymbol{\phi}_k)$ are region-specific distributions, and $\boldsymbol{\phi}_k$ the corresponding parameters. This type of model may be used for intensity-based segmentation, for texture-based segmentation (each x_i would then be a d-dimensional vector with d local texture features), or for segmentation of multi-spectral images (such as color or remote sensing images, with each $x_i \in I\!R^d$, where d is the number of spectral bands). The densities $p(\cdot|\boldsymbol{\phi}_k)$ can simply be Gaussians, or any other arbitrarily complex models, such as finite mixtures or kernel-based density estimates. Initially, we will focus on supervised segmentation with generative models, *i.e.*, we assume full knowledge of all $p(\cdot|\boldsymbol{\phi}_k)$. We will later show how to relax this assumption.

The goal of segmentation is, of course, to estimate $\mathbf{y}$, having observed $\mathbf{x}$. Clearly, the maximum likelihood (ML) estimate, $\widehat{\mathbf{y}}_{\text{ML}} = \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})$, can be obtained pixel-by-pixel, due to the independence assumption. However, pixel-wise segmentations are well known to lack spatial coherence [25]. To overcome this, a standard approach is to adopt an MRF prior $p(\mathbf{y})$, expressing the *a priori* preference for segmentations in which neighboring sites belong to the same region (see [13] for details and references). Under this prior, it is then common to adopt the *maximum a posteriori* (MAP) criterion, $\widehat{\mathbf{y}} = \arg\max_{\mathbf{y}}[\log p(\mathbf{y}) + \log p(\mathbf{x}|\mathbf{y})]$ (although there are other criteria). Due to the discrete nature of $\mathbf{y}$, the MAP criterion leads to a hard combinatorial optimization problem, to which much research has been devoted [13]. A recent breakthrough in MRF-based approaches is the adoption of fast algorithms based on graph cuts to solve this type of combinatorial problems [25].

2.3. Logistic Model

To keep the notation simple, consider first the binary case ($K = 2$, thus $y_i = [y_i^{(0)},\ y_i^{(1)}]$). Instead of writing directly a prior for $\mathbf{y}$ (the discrete labels), consider a "hidden image" $\mathbf{z} = \{z_i \in I\!R,\ i \in \mathcal{L}\}$ to which $\mathbf{y}$ is related via

$$p(y_i^{(1)} = 1|z_i) = \left(1 + e^{-z_i}\right)^{-1} \equiv \sigma(z_i), \qquad (2)$$

where $\sigma(\cdot)$ is the *logistic* function [9], and $p(y_i^{(0)} = 1|z_i) = 1 - \sigma(z_i)$. This formulation is close, in spirit, to the hidden Markov measure fields proposed in [15]; however, our hidden field $\mathbf{z}$ is real-valued, and totally unconstrained, thus much easier to model and manipulate than measure fields.

For K regions, K hidden images $\{\mathbf{z}^{(0)}, ..., \mathbf{z}^{(K-1)}\}$ are needed, where $\mathbf{z}^{(k)} = \{z_i^{(k)},\ i \in \mathcal{L}\}$. The region probabilities are given by a *multinomial logistic* model [1],

$$p(y_i^{(k)} = 1|\mathbf{z}_i) = e^{z_i^{(k)}} \left(\sum_{j=0}^{K-1} e^{z_i^{(j)}}\right)^{-1},\ k = 0, ..., K-1, \qquad (3)$$

where $\mathbf{z}_i = \{z_i^{(0)}, ..., z_i^{(K-1)}\}$. Since these probabilities are normalized, $\sum_{k=0}^{K-1} p(y_i^{(k)} = 1|\mathbf{z}_i) = 1$, we can set $\mathbf{z}^{(0)}$ to be identically zero, without loss of generality [1]. Notice that $\mathbf{z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(K-1)}\}$ is not under any type of constraint; any assignment of real values to its elements leads to valid probabilities for each site of $\mathbf{y}$.

2.4. Prior

It is now formally simple to write a prior for $\mathbf{z}$, due to its unconstrained real-valued nature. Although there are other possibilities, we will consider here wavelet-based priors, aiming at exploiting their ability to represent piece-wise smooth images. More precisely, it is known that piece-wise smooth images have sparse representations on wavelet bases [14], a fact which underlies their excellent performance in denoising and compression. Piece-wise smooth hidden images will translate into segmentations in which pixels in each class tend to form connected regions.

In a wavelet-based model, each image $\mathbf{z}^{(k)}$ is represented in terms of a wavelet expansion

$$\mathbf{z}^{(k)} = \mathbf{W}\boldsymbol{\theta}^{(k)}, \quad k = 1, ..., K-1, \tag{4}$$

where $\mathbf{W}$ is a matrix where each column is a wavelet basis function and $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K-1)}\}$ is a set of coefficients. For an orthogonal wavelet basis, $\mathbf{W}$ is a $|\mathcal{L}| \times |\mathcal{L}|$ orthogonal matrix, *i.e.*, $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$; in over-complete representations (*e.g.*, shift-invariant), $\mathbf{W}$ has more columns than lines (thus is no longer orthogonal) [14].

Since $\mathbf{z}^{(k)}$ is a deterministic function of $\boldsymbol{\theta}^{(k)}$, we write

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i\in\mathcal{L}} \prod_{k=0}^{K-1} [p(y_i^{(k)} = 1|\mathbf{z}_i(\boldsymbol{\theta}))]^{y_i^{(k)}}, \tag{5}$$

where $\mathbf{z}_i(\boldsymbol{\theta}) = \{(\mathbf{W}\boldsymbol{\theta}^{(1)})_i, ..., (\mathbf{W}\boldsymbol{\theta}^{(K-1)})_i\}$.

A prior $p(\boldsymbol{\theta})$ on the coefficients induces a prior $p(\mathbf{z})$ on the hidden images, via the (deterministic) relationship (4), and consequently on the segmentations. The role of $p(\boldsymbol{\theta})$ is to favor sparseness of the representation, which corresponds to favoring piece-wise smoothness in $\mathbf{z}$. Classical choices for $p(\boldsymbol{\theta})$ are independent (heavy-tailed) generalized Gaussians [18], of which the Laplacians,

$$p(\boldsymbol{\theta}) = \prod_{k=1}^{K-1} \prod_j (\lambda/2) \exp\{-\lambda\, |\theta_j^{(k)}|\}, \tag{6}$$

are a particular case (Laplacians are the heaviest-tailed densities which are still log-concave). This $p(\boldsymbol{\theta})$ corresponds to a strongly non-Gaussian and non-Markovian prior $p(\mathbf{z})$. However, unlike with MRF priors, simple exact deterministic algorithms can be used, supported on the available fast algorithms for forward and inverse wavelet transforms [14].

Summarizing, our complete model includes: **(a)** the prior $p(\boldsymbol{\theta})$, given by (6); **(b)** the probabilistic model $p(\mathbf{y}|\boldsymbol{\theta})$, given by (5); **(c)** the probabilistic model $p(\mathbf{x}|\mathbf{y})$, given by (1).

3. Estimation Criterion and Algorithm

3.1. Marginal MAP and the EM Algorithm

In our formulation, $\mathbf{x}$ is observed but, of course, $\mathbf{y}$ and $\boldsymbol{\theta}$ are not. The *a posteriori* probability is thus

$$p(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})\; p(\mathbf{y}|\boldsymbol{\theta})\; p(\boldsymbol{\theta}). \tag{7}$$

Among several possible criteria, we consider the *marginal maximum a posteriori* (MMAP), given by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\boldsymbol{\theta}) \;\sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}|\boldsymbol{\theta}) \right\} \tag{8}$$

where $\sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})p(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ is the marginal likelihood (over all possible segmentations). From $\widehat{\boldsymbol{\theta}}$, the posterior probability that each site belongs to each region is given by $p(\mathbf{y}|\widehat{\boldsymbol{\theta}})$; hard segmentations can be obtained by assigning each site to the region of highest posterior probability.

Although the maximization in (8) can not be done directly, due to the combinatorial nature of $p(\mathbf{x}|\boldsymbol{\theta})$, the following observations suggest using the EM algorithm [17], by treating $\mathbf{y}$ as missing data:

- If $\mathbf{y}$ was observed, estimating $\boldsymbol{\theta}$ would reduce to standard logistic regression under prior $p(\boldsymbol{\theta})$, that is, one could solve $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}}[\log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})]$.
- The complete log-likelihood $\log p(\mathbf{y}|\boldsymbol{\theta})$ is the standard logistic log-likelihood (see [1]), which is linear with respect to the $y_i^{(k)}$ variables:

$$\begin{aligned}\log p(\mathbf{y}|\boldsymbol{\theta}) &= \sum_{i\in\mathcal{L}} \sum_{k=0}^{K} y_i^{(k)} \log \frac{e^{(\mathbf{W}\boldsymbol{\theta}^{(k)})_i}}{\sum_{j=0}^{K-1} e^{(\mathbf{W}\boldsymbol{\theta}^{(j)})_i}} \\ &= \sum_{i\in\mathcal{L}} \sum_{k=0}^{K-1} y_i^{(k)} (\mathbf{W}\boldsymbol{\theta}^{(k)})_i - \sum_{i\in\mathcal{L}} \log \sum_{j=0}^{K-1} e^{(\mathbf{W}\boldsymbol{\theta}^{(j)})_i}. \end{aligned} \tag{9}$$

3.2. The E-step

The fact that the complete log-likelihood is linear w.r.t. the missing variables is crucial: the E-step reduces to computing the conditional expectation of the missing variables, which are then plugged into the complete log-likelihood [17]. As in finite mixtures [17], each missing variable $y_i^{(k)}$ is binary, thus its expectation, herein denoted as $\widehat{y}_i^{(k)}$, equals its probability of being equal to one:

$$\widehat{y}_i^{(k)} = p\left[y_i^{(k)} \!=\! 1|\widehat{\boldsymbol{\theta}}, \mathbf{x}\right] = \frac{p(x_i|\boldsymbol{\phi}_k)\, p\left[y_i^{(k)} = 1|\mathbf{z}_i(\widehat{\boldsymbol{\theta}})\right]}{\displaystyle\sum_{j=0}^{K-1} p(x_i|\boldsymbol{\phi}_j)\, p\left[y_i^{(j)} = 1|\mathbf{z}_i(\widehat{\boldsymbol{\theta}})\right]}. \tag{10}$$

Notice that this is essentially the same as the E-step for finite mixtures [17], with $p(y_i^{(k)} = 1|\mathbf{z}_i(\widehat{\boldsymbol{\theta}}))$ playing the role of mixing probabilities, and with fixed component densities $p(x|\boldsymbol{\phi}_k)$ (recall we're temporarily assuming known $\boldsymbol{\phi}_k$). The computational cost of this E-step is $O(K|\mathcal{L}|)$.

3.3. The M-step

Given the expected values of the missing variables, $\widehat{\mathbf{y}} = \{\widehat{\mathbf{y}}^{(0)}, ..., \widehat{\mathbf{y}}^{(K-1)}\}$, where $\widehat{\mathbf{y}}^{(k)} = \{\widehat{y}_i^{(k)},\ i \in \mathcal{L}\}$, the EM algorithm proceeds by plugging $\widehat{\mathbf{y}}$ into the complete log-likelihood and maximizing it w.r.t the $\boldsymbol{\theta}$, that is,

$$\widehat{\boldsymbol{\theta}}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) + l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \right\}, \tag{11}$$

where $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ is obtained by inserting $\widehat{\mathbf{y}}$ in the place of $\mathbf{y}$ in (9). Solving (11) is thus equivalent to performing multinomial logistic regression, with the usual hard (binary) training labels $y_i^{(k)} \in \{0,1\}$ replaced by "soft" labels $\widehat{y}_i^{(k)} \in [0,1]$, and under a prior $p(\boldsymbol{\theta})$.

The usual approach to ML logistic regression (*i.e.*, for maximizing just $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ w.r.t. $\boldsymbol{\theta}$) is the Newton-Raphson algorithm [9], known in this context as *iteratively reweighted least squares* (IRLS). However, IRLS can't be used for solving (11) with heavy-tailed priors (*e.g.*, (6)), which are not differentiable (at the origin). Alternatively, we adopt the bound optimization approach [12], introduced for logistic regression in [1] and [2] (see also [11]).

Let us temporarily ignore the prior $p(\boldsymbol{\theta})$ and consider $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ as the objective function, simply denoted as $l(\boldsymbol{\theta})$ for notational economy. In the bound optimization approach, this maximization is achieved by iteratively maximizing a so-called "surrogate" function (with t denoting an iteration counter)

$$\widehat{\boldsymbol{\theta}}_{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(t)}). \tag{12}$$

The condition that $l(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(t)})$ attains its minimum for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{(t)}$ is sufficient to guarantee monotonicity, *i.e.*, $l(\widehat{\boldsymbol{\theta}}_{(t+1)}) \geq l(\widehat{\boldsymbol{\theta}}_{(t)})$ [12].

In [1], the following surrogate function for multinomial logistic regression was introduced:

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(t)}) &= l(\widehat{\boldsymbol{\theta}}_{(t)}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{(t)})^T \mathbf{g}(\widehat{\boldsymbol{\theta}}_{(t)}) \\ &\quad - \frac{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{(t)})^T \mathbf{B} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{(t)})}{2}, \end{aligned}$$

where $\mathbf{B}$ is a positive definite matrix, which provides a lower bounds for the (negative definite) Hessian $\mathcal{H}(\boldsymbol{\theta})$ of $l(\boldsymbol{\theta})$, *i.e.*, $-\mathcal{H}(\boldsymbol{\theta}) \preceq \mathbf{B}$ (where $\mathbf{P} \preceq \mathbf{Q}$ means that $\mathbf{Q} - \mathbf{P}$ is positive semi-definite), and $\mathbf{g}(\boldsymbol{\theta}')$ denotes the gradient of $l(\boldsymbol{\theta})$ computed at $\boldsymbol{\theta}'$. Matrix $\mathbf{B}$ is given by (see [1])

$$\mathbf{B} = \frac{1}{2}\left[\mathbf{I}_{K-1} - (\mathbf{1}_{K-1}\,\mathbf{1}_{K-1}^T)/K\right] \otimes \left(\mathbf{W}^T\mathbf{W}\right), \tag{13}$$

where $\mathbf{I}_a$ is an $a \times a$ identity matrix, $\mathbf{1}_a = [1, ..., 1]^T$ is an a-dimensional vector of ones, and $\otimes$ the Kroenecker product.

The following Lemma (proved in [8]) provides a simpler (though looser) bound, applicable when using orthogonal or redundant wavelet representations, *i.e.*, when the columns of $\mathbf{W}$ are a *C-tight frame* (see Appendix for definition).

Lemma 1 *Let the columns of* $\mathbf{W}$ *be a* C*-tight frame. Then, matrix* $\mathbf{B}$*, defined in (13) (and consequently also* $-\mathcal{H}(\boldsymbol{\theta})$*) is upper bounded as follows:*

$$\mathbf{B} \preceq \mathbf{I}\,\xi_K, \quad \textit{where} \;\; \xi_K = \begin{cases} C/2 & \Leftarrow \;\; K > 2 \\ C/4 & \Leftarrow \;\; K = 2. \end{cases} \tag{14}$$

It is thus possible to replace $\mathbf{B}$ by $\mathbf{I}\;\xi_K$ and still have a valid surrogate bound. Simple manipulation, using the fact that one is free to add to the surrogate any terms independent of $\boldsymbol{\theta}$ (thus irrelevant for the maximization) leads to

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(t)}) = -\frac{\xi_K}{2}\;\|\boldsymbol{\theta} - \mathbf{v}_{(t)}\|_2^2, \tag{15}$$

where

$$\mathbf{v}_{(t)} = \widehat{\boldsymbol{\theta}}_{(t)} + \frac{\mathbf{g}(\widehat{\boldsymbol{\theta}}_{(t)})}{\xi_K}. \tag{16}$$

The gradient of the logistic log-likelihood function is

$$\mathbf{g}(\widehat{\boldsymbol{\theta}}_{(t)}) = \begin{bmatrix} \mathbf{W}^T\mathbf{d}_{(t)}^{(1)} \\ \vdots \\ \mathbf{W}^T\mathbf{d}_{(t)}^{(K-1)} \end{bmatrix}, \tag{17}$$

with $\mathbf{d}_{(t)}^{(k)} = [\widehat{y}_1^{(k)} - \widehat{p}_1^{(k)}, ..., \widehat{y}_{|\mathcal{L}|}^{(k)} - \widehat{p}_{|\mathcal{L}|}^{(k)}]^T$, where

$$\widehat{p}_i^{(k)} = p\left[y_i^{(k)} = 1|\mathbf{z}_i(\widehat{\boldsymbol{\theta}}_{(t)})\right] \tag{18}$$

are the current class probability estimates at each site. Let us define the images $\widehat{\mathbf{p}}^{(k)} = \{\widehat{p}_i^{(k)},\; i \in \mathcal{L}\}$.

Since a bound on $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ is, of course, also a bound on $l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) + \log p(\boldsymbol{\theta})$, the following update equation results:

$$\widehat{\boldsymbol{\theta}}_{(t+1)} = \arg\max_{\boldsymbol{\theta}} \left\{-\xi_K\|\boldsymbol{\theta} - \mathbf{v}_{(t)}\|_2^2 + 2\log p(\boldsymbol{\theta})\right\}. \tag{19}$$

This is simply the MAP estimate of $\boldsymbol{\theta}$, under prior $p(\boldsymbol{\theta})$, for a Gaussian white noise observation model with variance $1/\xi_K$. For example, under independent Laplacian priors, as given in (6), the update equation (19) is

$$\widehat{\boldsymbol{\theta}}_{(t+1)} = \text{soft}\left[\mathbf{v}_{(t)}, (\lambda\,\xi_K)^{-1}\right] \tag{20}$$

where $\text{soft}(\mathbf{u}, t) = \text{sign}(\mathbf{u})\max\{0, |\mathbf{u}| - t\}$ denotes the component-wise *soft-threshold* function (see [14], [18]).

In summary, the iterative procedure defined by (20) and (16) is used to solve the maximization (11) required by the M-step. Running this a finite number of times is not guaranteed to solve (11), but to obtain a new estimate that improves $\log p(\boldsymbol{\theta}) + l(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$. The resulting scheme is thus a generalized EM (GEM) algorithm [17], with the same monotonicity properties as standard EM.

3.4. Summary of the Algorithm

Inputs: Observed image $\mathbf{x}$, number of classes K, observation models $p(\cdot|\phi_k)$, wavelet basis $\mathbf{W}$, parameter λ, stopping threshold ε, number of inner iterations r.

Output: Parameter estimates $\widehat{\boldsymbol{\theta}}^{(k)}$, for $k = 1, ..., K-1$.

Initialization: For $k = 1, ..., K-1$, set $\widehat{\boldsymbol{\theta}}^{(k)} = \mathbf{0}$.

Step 1: For $k = 1, ..., K-1$, compute $\mathbf{z}^{(k)} = \mathbf{W}\widehat{\boldsymbol{\theta}}^{(k)}$.

Step 2: For $k = 1, ..., K-1$, compute $\widehat{\mathbf{y}}^{(k)}$ using (10).

Step 3: Store the current estimate: $\widehat{\boldsymbol{\theta}}_{\text{old}} = \widehat{\boldsymbol{\theta}}$.

Step 4: Repeat r times:

- **Step 4.a:** For $k = 1, ..., K-1$, let $\mathbf{d}^{(k)} = \widehat{\mathbf{y}}^{(k)} - \widehat{\mathbf{p}}^{(k)}$.
- **Step 4.b:** For $k = 1, ..., K-1$, compute the "wavelet transforms" $\mathbf{g}^{(k)} = \mathbf{W}^T\mathbf{d}^{(k)}$ (see (17)).
- **Step 4.c:** For $k = 1, ..., K-1$, let $\mathbf{v}^{(k)} = \widehat{\boldsymbol{\theta}}^{(k)} + \mathbf{g}^{(k)}/\xi_K$ (see (16)).
- **Step 4.d:** For $k = 1, ..., K-1$, update estimates using (20), which yields a new $\widehat{\boldsymbol{\theta}}^{(k)}$.
- **Step 4.e:** Go back to **Step 4.a**.

Step 5: If $\max_k \|\widehat{\boldsymbol{\theta}}_{\text{old}}^{(k)} - \widehat{\boldsymbol{\theta}}^{(k)}\|_\infty < \varepsilon$, then stop; otherwise, go back to **Step 1**.

It is important to notice that, in this supervised mode, with a Laplacian prior, the objective function being maximized in concave (since the logistic log-likelihood is also concave) and so there are no initialization problems. This is the reason behind the choice of this simple initialization procedure.

4. Extensions

4.1. Unsupervised/Semi-supervised Segmentation

The model and algorithm above described can be extended to the unsupervised case, *i.e.*, with the parameters $\boldsymbol{\phi}_k$ considered unknown. In this case, the full posterior is

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})\, p(\mathbf{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, p(\boldsymbol{\phi}) \tag{21}$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_0, ..., \boldsymbol{\phi}_{K-1}\}$. In this paper, we will assume a flat prior on $\boldsymbol{\phi}$, although other alternatives could be considered at little additional cost. Let us adopt again the MMAP criterion, now jointly w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. The following observations can now be added to those made in Section 3.1:

- If $\mathbf{y}$ was observed, estimating $\boldsymbol{\phi}$ could simply be achieved by the ML criterion, by maximizing the complete log-likelihood $\log p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$ w.r.t. $\boldsymbol{\phi}$.
- The complete log-likelihood is linear w.r.t. all $y_i^{(k)}$:

$$\log p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) = \sum_{i\in\mathcal{L}} \sum_{k=0}^{K-1} y_i^{(k)} \log p(x_i|\boldsymbol{\phi}_k).$$

The algorithm presented in Section 3.4 can thus be modified by inserting an extra step, between steps 3 and 4:

Step 3.5: Update the observation model parameters according to the following weighted ML criterion:

$$\widehat{\boldsymbol{\phi}}_k = \arg\max_{\boldsymbol{\phi}_k} \sum_{i\in\mathcal{L}} \widehat{y}_i^{(k)} \log p(x_i|\boldsymbol{\phi}_k).$$

For example, if the $p(\cdot|\boldsymbol{\phi}_k)$ are Gaussians, this update equations coincide with those of EM for Gaussian mixtures.

In the semi-supervised case, one is given a subset of pixels for which the true label is known. In this case, the EM algorithm for the unsupervised case is applied, but holding the labels of the pre-classified pixels at their known values.

Of course, in the unsupervised or semi-supervised modes, the log-marginal-posterior is no longer convex, and the results will depend critically on the initialization.

4.2. Discriminative Features

The formulation above presented (in fact, most of the work on probabilistic segmentation) adopts a generative perspective: each $p(\cdot|\boldsymbol{\phi}_k)$ is assumed to model the probabilistic data generation mechanism in each class. However, discriminative methods (*e.g.*, logistic regression, Gaussian processes, trees, support vector machines, boosting) are seen as the current state-of-the-art in classification [9].

Observe that all our EM algorithm requires, in the E-step defined in (10), is the posterior class probabilities, given the pixel values and the current parameter estimates $\widehat{\boldsymbol{\theta}}^{(k)}$. These parameter estimates work by forcing some prior class probabilities in (10). Consider a probabilistic discriminative classifier, *i.e.*, one that, for each pixel x_i, provides estimates of the posterior class probabilities $p(y_i^{(k)} = 1|x_i)$, $k = 0, ..., K-1$. Let us assume that this classifier was trained on balanced data, *i.e.*, using the same amount of data from each class. It can thus be assumed that these posterior class probabilities verify $p(y_i^{(k)} = 1|x_i) \propto p(x_i|y_i^{(k)} = 1)$, as can be easily verified by plugging uniform class priors $p(y_i^{(k)} = 1) = 1/K$ in Bayes rule. It is then possible to "bias" these classes, with given prior probabilities $q(y_i^{(k)} = 1)$, for $k = 0, ..., K-1$, by computing

$$p_{\text{biased}}(y_i^{(k)} = 1|x_i) = \frac{p(y_i^{(k)} = 1|x_i)\; q(y_i^{(k)} = 1)}{\sum_{j=0}^{K-1} p(y_i^{(j)} = 1|x_i)\; q(y_i^{(j)} = 1)}.$$

This procedure allows using a pre-trained probabilistic discriminative classifier in our EM algorithm, by using the "biased" probabilities in the E-step. We have not yet performed experiments with this discriminative approach.

5. Experiments

This section reports experimental results illustrating the behavior of the proposed algorithm, namely the adequacy of wavelet-based priors for segmentation.

The first experiment (reported in Fig. 1) is based on a simple synthetic segmentation problem, with known class models. Each of the four regions follows a Gaussian distribution with standard deviation 0.5 and means 1, 2, 3, and 4. This is very similar to what would be obtained by an MRF-based method; however, it must be stressed that the algorithm herein proposed is optimal, deterministic, and has computational complexity which grows linearly with the image size. In this example, undecimated Haar wavelets were used, $r = 4$, and $\varepsilon = 0.001$. Instead of fixing some value for lambda, the Bayes-shrink method [3] (which estimates a different threshold for each level of the wavelet decomposition) was used; this avoids having to specify or estimate parameters of the wavelet prior and, in all experiments, lead to very good results. This result illustrates the ability of the proposed wavelet-based prior to regularize image segmentation, producing well defined boundaries.

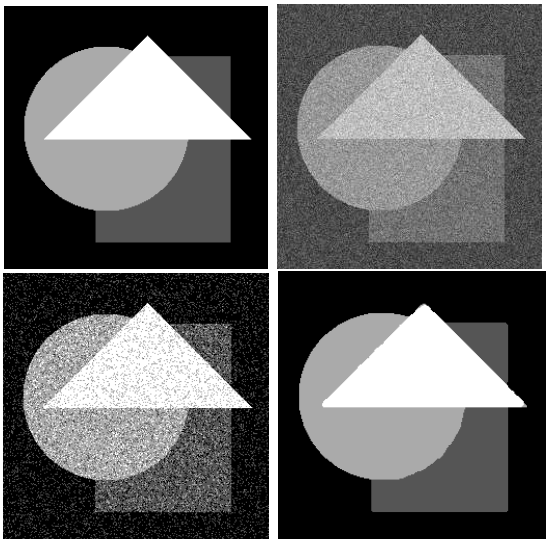

Figure 1. Top row: true regions and observed image. Bottom row: ML segmentation and that obtained by our algorithm.

The previous experiment was repeated in unsupervised mode, using a threshold-based segmentation to initialize the algorithm. The result obtained is visually very similar to the one in Fig. 1, and it's not show it here for the sake of space. The parameter estimates are within 1% of the true values.

For real images, the results depend strongly on the features and feature models used, which are not the focus of this paper. Only two examples of color image segmentation, using Gaussian color models, will be shown. In Fig. 2, the goal is to segment the image into three regions: clothe, skin, and background. Fig. 3 shows a figure-ground segmentation problem. These results were obtained by using the proposed algorithm in unsupervised mode, initialized with ML segmentations resulting from fitting Gaussian mixtures to the observed (RGB) pixels.

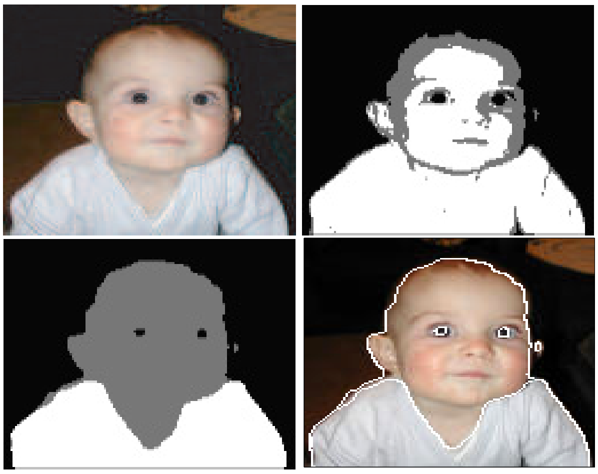

Figure 2. Top: observed image; ML segmentation. Bottom: segmentation obtained by our algorithm; corresponding boundaries.

6. Summary and Future Work

A new formulation of image segmentation as spatially-regularized logistic regression was introduced. This approach allows using priors for continuous-valued fields in image segmentation; here, it was used with wavelet-based priors. An EM algorithm was derived for supervised segmentation; it was shown how this algorithm is extended to handle unsupervised and semi-supervised problems, as well as discriminative features. Preliminary experiments show that the proposed approach has promising performance.

Future research will include a thorough experimental evaluation of the method, namely in comparison with graph-based and MRF-based methods. We are currently developing criteria for selecting the number of classes/regions, following the approach in [7].

Appendix: Frames and Tight Frames

Recall (see, *e.g.*, [14]) that a set $\{\mathbf{w}_1, ..., \mathbf{w}_p\}$ of p vectors in $\mathbb{R}^q$ is a *frame* if there exist two real constants A and B, with $0 < A \leq B$, such that for any vector $\mathbf{v} \in \mathbb{R}^q$,

$$A\|\mathbf{v}\|^2 \leq \sum_{i=1}^{p} |\mathbf{w}_i^T \mathbf{v}|^2 \leq B\|\mathbf{v}\|^2.$$

In a C-*tight* frame, $A = B = C$ (the inequalities become equalities). Examples of tight frames are orthonormal bases (1-thight) and unions of R orthonormal bases (R-tight).

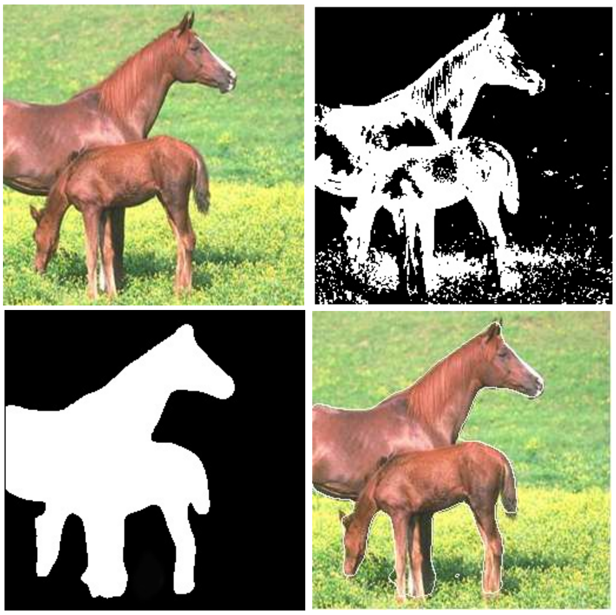

Figure 3. Top: observed image; ML segmentation. Bottom: segmentation obtained by our algorithm; corresponding boundaries.

References

[1] D. Böhning. "Multinomial logistic regression algorithm." *Annals Inst. Stat. Math.*, vol. 44, pp. 197–200, 1992.

[2] D. Böhning and B. Lindsay. "Monotonicity of quadratic-approximation algorithms." *Annals Inst. Stat. Math.*, vol. 40, pp. 641–663, 1988.

[3] G. Chang, B. Yu and M. Vetterli. "Adaptive wavelet thresholding for image denoising and compression." *IEEE Trans. Image Proc.*, vol. 9, pp. 1532–1546, 2000.

[4] H. Choi and R. Baraniuk. "Multiscale image segmentation using wavelet-domain hidden Markov models." *IEEE Trans. Image Proc.*, vol. 10, pp. 1309–1321, 2001.

[5] G. Cross and A. Jain. "Markov random field texture models." *IEEE-TPAMI*, vol. 5, pp. 25–39, 1983.

[6] H. Derin and H. Elliot. "Modelling and segmentation of noisy and textured images using Gibbsian random fields." *IEEE-TPAMI*, vol. 9 , pp. 39–55, 1987.

[7] M. Figueiredo and A.K.Jain. "Unsupervised learning of finite mixture models." *IEEE-TPAMI*, vol. 24, pp. 381-396, 2002.

[8] M. Figueiredo. "Wavelet-based logistic regression". In preparation. 2005.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Springer Verlag, New York, 2001.

[10] A. Jain and F. Farrokhnia. "Unsupervised texture segmentation using Gabor filters." *Pattern Recognition*, vol. 24, pp. 1167–1186, 1991.

[11] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. "Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds", *IEEE-TPAMI*, vol. 27, no. 6, 2005.

[12] K. Lange, D. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Jour. Comp. Graph. Stat.*, vol. 9, pp. 1–59, 2000.

[13] S. Z. Li, *Markov Random Field Modelling in Computer Vision*. Springer Verlag, 2001.

[14] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.

[15] J. Marroquin, E. Santana, and S. Botello. "Hidden Markov measure field models for image segmentation." *IEEE-TPAMI*, vol. 25, pp. 1380–1387, 2003.

[16] D. Martin, C. Fowlkes, and J. Malik. "Learning to detect natural image boundaries using local brightness, color and texture cues." *IEEE-TPAMI*, vol. 26, pp. 530–549, 2004.

[17] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997.

[18] P. Moulin and J. Liu. "Analysis of multiresolution image denoising schemes using generalized-Gaussian and complexity priors," *IEEE Trans. Info. Th.*, vol. 45, pp. 909–919, 1999.

[19] T. Randen and J. Husoy. "Filtering for texture classification: a comparative study." *IEEE-TPAMI*, vol. 21, pp. 291–310, 1999.

[20] J. Shi and J. Malik, "Normalized cuts and image segmentation." *IEEE-TPAMI*, vol. 22, pp. 888–905, 2000.

[21] M. Unser, "Texture classification and segmentation using wavelet frames." *IEEE Trans. Image Proc.*, vol. 4, pp. 1549–1560, 1995.

[22] Y. Weiss, "Segmentation using eigenvectors: a unifying view." *Proc. ICCV'99*, pp. 975–982, 1999.

[23] C. Williams and D. Barber. "Bayesian classification with Gaussian priors." *IEEE-TPAMI*, vol. 20, pp. 1342–1351, 1998.

[24] Z. Wu and R. Leahy, "Optimal graph theoretic approach to data clustering: theory and its application to image segmentation." *IEEE-TPAMI*, vol. 15, pp. 1101–1113, 1993.

[25] R. Zabih and V. Kolmogorov, "Spatially coherent clustering with graph cuts." *Proc. IEEE-CVPR*, vol. II, pp. 437–444, 2004.