Bayesian Image Segmentation Using Gaussian Field Priors

Mário A.T. Figueiredo

Instituto de Telecomunicações,
and Department of Electrical and Computer Engineering,
Instituto Superior Técnico, 1049-001 Lisboa, Portugal
Phone: +351 218418464, Fax: +351 218418472
mario.figueiredo@lx.it.pt

Abstract. The goal of segmentation is to partition an image into a finite set of regions, homogeneous in some (e.g., statistical) sense, thus being an intrinsically discrete problem. Bayesian approaches to segmentation use priors to impose spatial coherence; the discrete nature of segmentation demands priors defined on discrete-valued fields, thus leading to difficult combinatorial problems.

This paper presents a formulation which allows using continuous priors, namely Gaussian fields, for image segmentation. Our approach completely avoids the combinatorial nature of standard Bayesian approaches to segmentation. Moreover, it's completely general, *i.e.*, it can be used in supervised, unsupervised, or semi-supervised modes, with any probabilistic observation model (intensity, multispectral, or texture features).

To use continuous priors for image segmentation, we adopt a formulation which is common in Bayesian machine learning: introduction of hidden fields to which the region labels are probabilistically related. Since these hidden fields are real-valued, we can adopt any type of spatial prior for continuous-valued fields, such as Gaussian priors. We show how, under this model, Bayesian MAP segmentation is carried out by a (generalized) EM algorithm. Experiments on synthetic and real data shows that the proposed approach performs very well at a low computational cost.

1 Introduction

Image segmentation has been one of the most studied problems in computer vision. Although remarkably successful approaches have been proposed for specific domains in which the goals are well defined (*e.g.*, segmentation of magnetic resonance images, segmentation of remote sensing images), a general purpose segmentation criterion remains an elusive concept. In the past couple of decades, many different approaches, formulations, and tools have been proposed.

Most segmentation methods work by combining cues from the observed image (via image features) with some form of regularization (or prior, in Bayesian terms), embodying the concept of "acceptable" (or "*a priori* probable") segmentation. Arguably, all the work on image segmentation can be classified as belonging to on one (or even both) of the following two research fronts:

A. Rangarajan et al. (Eds.): EMMCVPR 2005, LNCS 3757, pp. 74–89, 2005.

(a) Development of image features, and feature models, which are as informative as possible for the segmentation goal. Some of the most recent proposals combine intensity, texture, and contour-based features, with the specific goal of mimicking human image segmentation [26]. Another recent approach combining several types of features is reported in [27]. Classical examples for texture-based segmentation include Gabor features [16], wavelet-based features [29], co-occurrence matrices [11], features derived from Markov random field local texture models [7], [8]. It's possible to perform segmentation using nonparametric statistical measures of texture similarity by resorting to pairwise clustering techniques [14]. The literature on texture features and models is vast; [25] provides a reasonably recent survey. There are many other examples of features developed for specific domains, such as color segmentation, segmentation of medical images, or segmentation of remote sensing images.

(b) Development of methods that impose some form of spatial regularity to the segmentation, *i.e.*, that integrate local cues (from features) into a globally coherent segmentation. The recent graph-based methods [28], [30], [32], achieve this by formulating image segmentation as the partitioning of a graph. Spatial coherence may also be achieved by constraining the class of image partitions which are considered by the segmentation algorithm (*e.g.*, [13] and [24] consider hierarchies of polygonal and quad-tree-like partitions, respectively) or by imposing some prior on the length or the smoothness of the region boundaries [34]; see recent work and many references in [17], which also advances research front **(a)**. In a probabilistic Bayesian approach, as adopted in this paper, the preference for some form of spatial regularity is usually formulated via a Markov random field (MRF) prior (see [20], for a comprehensive set of references).

This paper belongs to research front **(b)**: it describes a new way of introducing spatial priors for Bayesian image segmentation. The proposed approach uses priors on real-valued fields/images, rather than MRF priors for discrete labels, thus removing any combinatorial nature from the problem. Our formulation, is very general in that it can be used in supervised, unsupervised, or semi-supervised manners, as well as with generative or discriminative features.

To open the door to the use of priors on real-valued fields/images for image segmentation, we adopt an approach which is used in Bayesian machine learning: introduction of a (collection of) real-valued hidden field(s), to which the region labels are probabilistically related; these hidden field(s), being real-valued, can then be given any type of spatial prior, *e.g.*, it can be modelled as a (collection of) Gaussian field(s). This approach is used in the very successful approach to Bayesian learning of classifiers known as "Gaussian processes" [31]. In this paper, Gaussian field priors are adopted as a means of encoding a preference for spatially coherent segmentations. We show how the proposed approach can be used in supervised, unsupervised, and semi-supervised modes, by deriving (generalized) expectation-maximization (EM) algorithms for the three cases. In the supervised case, the resulting segmentation criterion consists in minimizing a convex cost function, thus initialization problems do not arise. If the underlying

Gaussian process prior is stationary, the M-step can be implemented in a very fast way using FFT-based processing in the Fourier domain. This is, arguably, one of the key advantages of the proposed approach.

Finally, we should mention that our formulation is close, in spirit, to the "hidden Markov measure fields" proposed in [22]; however, our hidden fields are real valued, and totally unconstrained, thus much easier to model and manipulate than measure fields. Recently, we have used a similar formulation to allow the use of wavelet-based spatial priors for image segmentation [9].

In the next section, we introduce notation and the proposed formulation. In Section 3, we present our segmentation criterion and derive the EM algorithm for implementing it. Section 4 describes the extensions to unsupervised, semi-supervised and discriminative segmentation. Finally, experiments are presented in Section 5, and Section 6 concludes the paper.

2 Formulation

2.1 Images and Segmentations

Let $\mathcal{L} = \{(n,m),\ n = 1,...,N,\ m = 1,...,M\}$ be a $2D$ lattice of $|\mathcal{L}| = MN$ 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 indicator images $\mathbf{y}^{(k)} = \{y_i^{(k)},\ i \in \mathcal{L}\}$, for $k = 0,...,K-1$, where $y_i^{(k)} \in \{0,1\}$, such that $(y_i^{(k)} = 1) \Leftrightarrow (i \in R_k)$. We denote as $\mathbf{y}$ the set of all these binary images, $\mathbf{y} = \{\mathbf{y}^{(0)},...,\mathbf{y}^{(K-1)}\}$, and as $\mathbf{y}_i$ the set of all $y_i^{(k)}$ for a given site i, that is, $\mathbf{y}_i = \{y_i^{(0)},...,y_i^{(K-1)}\}$. Of course, $\mathbf{y}$ and $\mathcal{R}$ carry exactly the same information.

2.2 Observation Model

Given a segmentation $\mathbf{y}$, we follow the standard assumption that the observed "pixels" are (conditionally) independently distributed,

$$p(\mathbf{x}|\mathbf{y}) = \prod_{k=0}^{K-1} \prod_{i \in R_k} p(x_i|\boldsymbol{\phi}^{(k)}) = \prod_{i \in \mathcal{L}} \prod_{k=0}^{K-1} \left[p(x_i|\boldsymbol{\phi}^{(k)})\right]^{y_i^{(k)}}, \tag{1}$$

where the $p(\cdot|\boldsymbol{\phi}^{(k)})$ are region-specific distributions. This type of model may be used for intensity-based segmentation, for texture-based segmentation (each x_i

is then a d-dimensional vector containing the values of d local texture features), or for segmentation of multi-spectral images (such as color images, or remote sensing images, with each x_i being in this case a d-dimensional vector, where d is the number of spectral bands). The region-specific densities $p(\cdot|\boldsymbol{\phi}^{(k)})$ can be simple Gaussians, or any other arbitrarily complex models, such as finite mixtures, kernel-based density representations, or even histograms. When the $p(\cdot|\boldsymbol{\phi}^{(k)})$ are fully known *a priori*, we are in the context of supervised segmentation with generative models. This is the case we will focus on first; later, it will be shown how the approach can be extended to unsupervised and semi-supervised scenarios, and to "discriminative features".

The goal of segmentation is, of course, to estimate $\mathbf{y}$, having observed $\mathbf{x}$. The maximum likelihood (ML) estimate, $\widehat{\mathbf{y}}_{\mathrm{ML}} = \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})$, can clearly be obtained pixel-by-pixel, due to the independence assumption. However, it's well known that pixel-wise segmentations may lack spatial coherence [20], [33]. To overcome this, one of the standard approaches consists in adopting an MRF prior $p(\mathbf{y})$, expressing the *a priori* preference for segmentations in which neighboring sites belong to the same region (see [20] for details and references). Given this prior, it is then most common to adopt the *maximum a posteriori* (MAP) criterion, $\widehat{\mathbf{y}}_{\mathrm{MAP}} = \arg\max_{\mathbf{y}}[\log p(\mathbf{y}) + \log p(\mathbf{x}|\mathbf{y})]$ (although there are other criteria). Due to the discrete nature of $\mathbf{y}$, finding $\widehat{\mathbf{y}}_{\mathrm{MAP}}$ involves a combinatorial optimization problem, to which much research has been devoted [20]. A recent breakthrough in MRF-type approaches (to segmentation [33] and other vision problems [5]) is the adoption of fast algorithms based on graph cuts[1].

2.3 Logistic Model

To keep the notation initially simple, consider the binary case ($K = 2$, thus each $\mathbf{y}_i = [y_i^{(0)}, y_i^{(1)}]$). Instead of designing a prior for $\mathbf{y}$ (the field of discrete labels), we consider a "hidden" (or latent) image $\mathbf{z} = \{z_i \in I\!R,\ i \in \mathcal{L}\}$, such that

$$p(\mathbf{y}|\mathbf{z}) = \prod_i p(\mathbf{y}_i|z_i) \quad \text{with} \quad p(y_i^{(1)} = 1|z_i) = \frac{e^{z_i}}{1 + e^{z_i}} \equiv \sigma(z_i), \tag{2}$$

where $\sigma(\cdot)$ is called the *logistic* function and, obviously, $p(y_i^{(0)} = 1|z_i) = 1 - \sigma(z_i)$.

In general, for K regions, we need K hidden images $\mathbf{z} = \{\mathbf{z}^{(0)}, ..., \mathbf{z}^{(K-1)}\}$, where $\mathbf{z}^{(k)} = \{z_i^{(k)} \in I\!R,\ i \in \mathcal{L}\}$. The region label probabilities are obtained via a multinomial logistic model (also known as a "soft-max"),

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

where $\mathbf{z}_i = \{z_i^{(0)}, ..., z_i^{(K-1)}\}$. Since these probabilities verify the normalization condition $\sum_{k=0}^{K-1} p(y_i^{(k)} = 1|\mathbf{z}_i) = 1$, one of the hidden images can be set to

[1] See `http://www.cs.cornell.edu/~rdz/graphcuts.html` for details and references.

zero; without loss of generality, we set $\mathbf{z}^{(0)} = \mathbf{0}$ (see, *e.g.*, [3]). 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 Gaussian Random Field Prior

It is now formally simple to write priors for $\mathbf{z}$, due to its unconstrained real-valued nature. Among the several possibilities, we will focus here on what is arguably the simplest choice: a Gauss-Markov random field (GMRF) prior defined on the lattice $\mathcal{L}$.

The goal of the prior on $\mathbf{z}$ is to express preference for segmentations such that neighboring sites have high probability of belonging to the same region. This is achieved by encouraging neighboring values of each $\mathbf{z}^{(k)}$ to be close to each other. A GMRF prior that embodies this preference is

$$p(\mathbf{z}) \propto \exp\left\{ -\frac{1}{4} \sum_{i \sim j} \sum_{k=1}^{K-1} w_{i,j} \left(z_i^{(k)} - z_j^{(k)} \right)^2 \right\}, \tag{4}$$

where $i \sim j$ denotes that sites i and j are neighbors (in some neighborhood system defined in $\mathcal{L}$), and the $w_{i,j}$ are (non-negative) weights. It is clear that (4) models the set of hidden fields $\mathbf{z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(K-1)}\}$ as *a priori* independent, *i.e.*,

$$p(\mathbf{z}) = \prod_{k=1}^{K-1} p(\mathbf{z}^{(k)}) \tag{5}$$

with

$$p(\mathbf{z}^{(k)}) \propto \exp\left\{ -\frac{1}{4} \sum_{i,j} w_{i,j} \left(z_i^{(k)} - z_j^{(k)} \right)^2 \right\}, \tag{6}$$

where the sum is now over all i, j because we encode the neighborhood structure in the $w_{i,j}$ by letting $w_{i,j} = 0$ when i and j are not neighbors. Let now $\mathbf{z}^{(k)} = [z_1^{(k)}, ..., z_{|\mathcal{L}|}^{(k)}]^T \in I\!R^{|\mathcal{L}|}$ denote an $|\mathcal{L}|$-vector obtained by stacking all the $z_i^{(k)}$ variables (for a given k) in standard lexicographical order. Also, let $\mathbf{W}$ be the $|\mathcal{L}| \times |\mathcal{L}|$ matrix with the $w_{i,j}$ weights. With this notation, we can write

$$p(\mathbf{z}^{(k)}) \propto \exp\left\{ -\frac{1}{2} (\mathbf{z}^{(k)})^T \boldsymbol{\Delta} (\mathbf{z}^{(k)}) \right\}. \tag{7}$$

where

$$\boldsymbol{\Delta} = \text{diag}\left\{ \sum_{j=1}^{|\mathcal{L}|} w_{1,j}, ..., \sum_{j=1}^{|\mathcal{L}|} w_{|\mathcal{L}|,j} \right\} - \mathbf{W} \tag{8}$$

is called the *graph-Laplacian matrix* [6]; in our case, the graph nodes are the sites of the lattice $\mathcal{L}$ and the edge weights are given by $w_{i,j}$ (with $w_{i,j} = 0$ denoting absence of edge between nodes i and j). Notice that $\boldsymbol{\Delta}$ has (at least) one zero eigenvalue since $\boldsymbol{\Delta}[1, 1, ..., 1]^T = \mathbf{0}$; thus, $p(\mathbf{z}^{(k)})$ is an improper prior (it can't be normalized [2]), but this is will not be a problem for MAP estimation. In the GMRF literature, $\boldsymbol{\Delta}$ is also called the *potential matrix* [1].

3 Estimation Criterion and Algorithm

3.1 Marginal MAP Criterion

Let us summarize our model: we have the observed field $\mathbf{x}$, and unobserved fields $\mathbf{y}$ and $\mathbf{z}$. These fields are probabilistically related by $p(\mathbf{x}|\mathbf{y})$, given by (1), $p(\mathbf{y}|\mathbf{z})$, given by (2) - (3), and a prior $p(\mathbf{z}) = p(\mathbf{z}^{(1)}) \cdots p(\mathbf{z}^{(K-1)})$ with each $p(\mathbf{z}^{(k)})$ given by (7). Given $\mathbf{x}$, the posterior probability of $\mathbf{y}$ and $\mathbf{z}$ is thus

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}|\mathbf{z})\, p(\mathbf{z}). \tag{9}$$

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

$$\widehat{\mathbf{z}} = \arg\max_{\mathbf{z}} \{p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})\} = \arg\max_{\mathbf{z}} \left\{ p(\mathbf{z}) \sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}|\mathbf{z}) \right\} \tag{10}$$

where $p(\mathbf{x}|\mathbf{z}) = \sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})\, p(\mathbf{y}|\mathbf{z})$ is the marginal likelihood obtained by summing over (the huge set of) all possible segmentations.

The estimate $\widehat{\mathbf{z}}$ is a probabilistic segmentation in the sense that it provides the probability that each pixel belongs to each region, via the logistic model (3). To obtain a hard segmentation, one can simply choose the *a posteriori* most probable class $\widehat{k}_i$ at each site i which is

$$\widehat{k}_i = \arg\max_{k} \{p(y_i^{(k)} = 1|\mathbf{z}_i)\}. \tag{11}$$

Clearly, the maximization in (10) can not be done directly, due to the combinatorial nature of $p(\mathbf{x}|\mathbf{z})$. In the next subsections, we will derive an EM algorithm for this purpose.

3.2 Why the EM Algorithm?

The following observations clearly suggest using the EM algorithm [23], treating $\mathbf{y}$ as missing data, to solve (10):

- If $\mathbf{y}$ was observed, estimating $\mathbf{z}$ would reduce to standard logistic regression under prior $p(\mathbf{z})$, that is, one could solve $\widehat{\mathbf{z}} = \arg\max_{\mathbf{z}} [\log p(\mathbf{y}|\mathbf{z}) + \log p(\mathbf{z})]$.
- The so-called complete log-likelihood $\log p(\mathbf{y}|\mathbf{z})$ (based on which we could estimate $\mathbf{z}$ if $\mathbf{y}$ was observed) is linear with respect to the hidden $y_i^{(k)}$ variables. In fact, $\log p(\mathbf{y}|\mathbf{z})$ is the standard logistic regression log-likelihood with an identity design matrix (see, e.g., [3], [12], [18]):

$$\log p(\mathbf{y}|\mathbf{z}) = \sum_i \sum_{k=0}^{K} y_i^{(k)} \log \frac{e^{z_i^{(k)}}}{\sum_{j=0}^{K-1} e^{z_i^{(k)}}} = \sum_i \left(\sum_{k=0}^{K} y_i^{(k)} z_i^{(k)} - \log \sum_{k=0}^{K} e^{z_i^{(k)}} \right). \tag{12}$$

The EM algorithm proceeds by iteratively applying the following two steps [23]:

E-step: Compute the expected value of the complete log-likelihood, given the current estimate $\widehat{\mathbf{z}}$ and the observations $\mathbf{x}$: $Q(\mathbf{z}|\widehat{\mathbf{z}}) = E_{\mathbf{y}}[\log p(\mathbf{y}|\mathbf{z})|\widehat{\mathbf{z}}, \mathbf{x}]$.
M-step: Update the estimate: $\widehat{\mathbf{z}} \leftarrow \widehat{\mathbf{z}}_{\text{new}} = \arg\max_{\mathbf{z}} \{Q(\mathbf{z}|\widehat{\mathbf{z}}) + \log p(\mathbf{z})\}$.

3.3 The E-Step

The fact that the complete log-likelihood is linear w.r.t. the missing variables is very important for EM: the E-step reduces to computing the expectation of the missing variables, with these expectations then plugged into the complete log-likelihood [23]. Moreover, as in finite mixtures [10], the missing $y_i^{(k)}$ are binary, thus their expected values are equal to their probabilities of being equal to one, which can be obtained via Bayes law:

$$\widehat{y}_i^{(k)} \equiv E[y_i^{(k)}|\widehat{\mathbf{z}}_i, \mathbf{x}] = p(y_i^{(k)} = 1|\widehat{\mathbf{z}}_i, \mathbf{x}_i) = \frac{p(x_i|\boldsymbol{\phi}^{(k)})\, p(y_i^{(k)} = 1|\widehat{\mathbf{z}}_i)}{\sum_{j=0}^{K-1} p(x_i|\boldsymbol{\phi}^{(j)})\, p(y_i^{(j)} = 1|\widehat{\mathbf{z}}_i)}. \quad (13)$$

Notice that this is essentially the same as the E-step for finite mixtures [10], with site-specific mixing probabilities given by $p(y_i^{(k)} = 1|\widehat{\mathbf{z}}_i)$ and with fixed component densities $p(x|\boldsymbol{\phi}^{(k)})$ (recall that we're temporarily assuming that all the $\boldsymbol{\phi}^{(k)}$ are known). Finally, $Q(\mathbf{z}|\widehat{\mathbf{z}})$ is obtained by plugging the $\widehat{y}_i^{(k)}$ (which depend on $\widehat{\mathbf{z}}$ via (13)) into the logistic log-likelihood (12):

$$Q(\mathbf{z}|\widehat{\mathbf{z}}) = \sum_i \left(\sum_{k=0}^{K} \widehat{y}_i^{(k)} z_i^{(k)} - \log \sum_{k=0}^{K} e^{z_i^{(k)}} \right). \quad (14)$$

Notice that $Q(\mathbf{z}|\widehat{\mathbf{z}})$ is formally a standard logistic regression log-likelihood, but with the usual hard (binary) training labels $y_i^{(k)} \in \{0, 1\}$ replaced by "soft" labels $\widehat{y}_i^{(k)} \in [0, 1]$.

3.4 Solving the M-Step

Our M-step, $\widehat{\mathbf{z}}_{\text{new}} = \arg\max_{\mathbf{z}} \{Q(\mathbf{z}|\widehat{\mathbf{z}}) + \log p(\mathbf{z})\}$, consists in solving a logistic regression problem with identity design matrix, given soft labels $\widehat{y}_i^{(k)}$, and under a prior $p(\mathbf{z})$. It is well known that this problem does not have a closed form solution and has to be solved by an iterative algorithm [3]. The standard choice for maximum likelihood logistic regression (*i.e.*, for maximizing only $Q(\mathbf{z}|\widehat{\mathbf{z}})$ w.r.t. $\mathbf{z}$) is Newton's algorithm [12]. However, as shown below, we will obtain a much simpler method by adopting the bound optimization approach [19], introduced for logistic regression in [3] and [4] (see also [18]).

Let us temporarily ignore the log-prior $\log p(\mathbf{z})$ and consider only $Q(\mathbf{z}|\widehat{\mathbf{z}})$, simply denoted as $q(\mathbf{z})$ for notational economy. In the bound optimization approach, the maximization of $q(\mathbf{z})$ is achieved by iterating the two following steps

$$\widehat{\mathbf{z}}_{\text{new}} = \arg\max_{\mathbf{z}} l(\mathbf{z}|\widehat{\mathbf{z}}), \qquad \widehat{\mathbf{z}} \leftarrow \widehat{\mathbf{z}}_{\text{new}}, \quad (15)$$

where $l(\mathbf{z}|\widehat{\mathbf{z}})$ is a so-called "surrogate" function verifying the following condition: $q(\mathbf{z}) - l(\mathbf{z}|\widehat{\mathbf{z}})$ attains its minimum for $\mathbf{z} = \widehat{\mathbf{z}}$ (see [19]). This condition is sufficient to guarantee that this iteration monotonically increases $q(\mathbf{z})$, *i.e.*, $q(\widehat{\mathbf{z}}_{\text{new}}) \geq q(\widehat{\mathbf{z}})$.

Thus, by running iteration (15) one or more times, after each application of the E-step (equations (13)-(14)), the resulting procedure is a generalized EM (GEM) algorithm [23].

It is important to notice that, in the supervised mode, the objective function being maximized in concave (since the logistic log-likelihood and the logarithm of the GMRF prior are both concave) and so there are no initialization problems.

From this point on, we assume that $\mathbf{z}$ is organized into a $((K-1)|\mathcal{L}|)$-vector by stacking the several $\mathbf{z}^{(k)}$ vectors, *i.e.*, $\mathbf{z} = [(\mathbf{z}^{(1)})^T, ..., (\mathbf{z}^{(K-1)})^T]^T$. In [3], the following surrogate for logistic regression was introduced:

$$l(\mathbf{z}|\widehat{\mathbf{z}}) = q(\widehat{\mathbf{z}}) + (\mathbf{z} - \widehat{\mathbf{z}})^T \mathbf{g}(\widehat{\mathbf{z}}) - \frac{(\mathbf{z} - \widehat{\mathbf{z}})^T \mathbf{B} (\mathbf{z} - \widehat{\mathbf{z}})}{2}, \tag{16}$$

where $\mathbf{g}(\widehat{\mathbf{z}})$ is the gradient of $q(\mathbf{z})$ computed at $\widehat{\mathbf{z}}$ and $\mathbf{B}$ is a positive definite matrix which provides a lower bounds for the (negative definite) Hessian $\mathcal{H}(\mathbf{z})$ of $q(\mathbf{z})$, *i.e.*, $\mathcal{H}(\mathbf{z}) \succeq -\mathbf{B}$ (in the matrix sense, *i.e.*, $\mathcal{H}(\mathbf{z}) + \mathbf{B}$ is positive semi-definite). Since $q(\mathbf{z}) - l(\mathbf{z}|\widehat{\mathbf{z}}) \geq 0$, with equality if and only if $\mathbf{z} = \widehat{\mathbf{z}}$, $l(\mathbf{z}|\widehat{\mathbf{z}})$ is a valid surrogate function; any other function differing from it by an additive constant (irrelevant for (15)) is also a valid surrogate. Matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \frac{1}{2} \left(\mathbf{I}_{K-1} - \frac{\mathbf{1}_{K-1} \, \mathbf{1}_{K-1}^T}{K} \right) \otimes \mathbf{I}_{|\mathcal{L}|}, \tag{17}$$

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

The following simple Lemma (proved in the Appendix) will allow further simplification of the algorithm, by using a less tight, but simpler bound matrix.

Lemma 1. *Let us define* ξ_K *as*

$$\xi_K = \begin{cases} 1/2 & \text{if} \quad K > 2 \\ 1/4 & \text{if} \quad K = 2. \end{cases} \tag{18}$$

Then, $\mathbf{B} \preceq \xi_K \, \mathbf{I}_{(K-1)|\mathcal{L}|}$, *with equality if* $K = 2$.

This lemma allows us to replace $\mathbf{B}$ by $\xi_K \, \mathbf{I}_{(K-1)|\mathcal{L}|}$ in (16) and still have a valid surrogate; the advantage is that in this new surrogate the several $\mathbf{z}^{(k)}$ become decoupled. Performing some simple manipulation, using the fact that one is free to add to the surrogate any terms independent of $\mathbf{z}$ (thus irrelevant for the maximization), leads to

$$l(\mathbf{z}|\widehat{\mathbf{z}}) = -\frac{\xi_K}{2} \sum_{k=1}^{K-1} \|\mathbf{z}^{(k)} - \mathbf{v}^{(k)}\|_2^2, \quad \text{with} \quad \mathbf{v}^{(k)} = \widehat{\mathbf{z}}^{(k)} + \frac{\mathbf{d}^{(k)}}{\xi_K}, \tag{19}$$

where $\| \cdot \|_2^2$ denotes squared Euclidean norm,

$$\mathbf{d}^{(k)} = \begin{bmatrix} \widehat{y}_1^{(k)} - p(y_1^{(1)} = 1 | \widehat{\mathbf{z}}_1) \\ \vdots \\ \widehat{y}_{|\mathcal{L}|}^{(k)} - p(y_{|\mathcal{L}|}^{(k)} = 1 | \widehat{\mathbf{z}}_{|\mathcal{L}|}) \end{bmatrix}, \tag{20}$$

and the $p(y_1^{(k)} = 1|\widehat{\mathbf{z}}_1)$ are given by the logistic model (3).

Since a surrogate for $Q(\mathbf{z}|\widehat{\mathbf{z}})$ is also valid for $Q(\mathbf{z}|\widehat{\mathbf{z}})+\log p(\mathbf{z})$, and (see (5)-(7))

$$\log p(\mathbf{z}) = \sum_{k=1}^{K-1} \log p(\mathbf{z}^{(k)}) = A - \frac{1}{2} \sum_{k=1}^{K-1} (\mathbf{z}^{(k)})^T \boldsymbol{\Delta} \, (\mathbf{z}^{(k)}), \tag{21}$$

(A is an irrelevant constant) the following decoupled update equation results:

$$\widehat{\mathbf{z}}_{\text{new}}^{(k)} = \arg\min_{\mathbf{z}} \left\{ \|\mathbf{z} - \mathbf{v}^{(k)}\|_2^2 + \frac{\mathbf{z}^T \boldsymbol{\Delta} \, \mathbf{z}}{\xi_K} \right\} = \xi_K \left(\xi_K \mathbf{I}_{|\mathcal{L}|} + \boldsymbol{\Delta} \right)^{-1} \mathbf{v}^{(k)}, \tag{22}$$

for $k = 1, ..., K-1$.

3.5 FFT-Based Implementation of the M-Step

For a general matrix $\boldsymbol{\Delta}$ (*i.e.*, an arbitrary choice of $\mathbf{W}$), (22) is computationally very expensive, requiring $O(|\mathcal{L}|^3)$ operations. However, for certain choices of $\mathbf{W}$ (correspondingly of $\boldsymbol{\Delta}$), we can resort to fast frequency-domain methods. Suppose that $w_{i,j}$ only depends on the relative position of i and j (the Gaussian field prior is stationary) and that the neighborhood system has periodic boundary condition; in this case, both $\mathbf{W}$ and $\boldsymbol{\Delta}$ are block-circulant matrices, with circulant[2] blocks [1]. It is well known that block-circulant matrices with circulant blocks can be diagonalized by a two-dimensional discrete Fourier transform (2D-DFT): $\boldsymbol{\Delta} = \mathbf{U}^H \mathbf{D} \mathbf{U}$, where $\mathbf{D}$ is a diagonal matrix, $\mathbf{U}$ is the matrix representation of the 2D-DFT, and the superscript $(\cdot)^H$ denotes conjugate transpose. Since $\mathbf{U}$ is an orthogonal matrix ($\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathbf{I}$), the inversion in (22) can be written as

$$\widehat{\mathbf{z}}_{\text{new}}^{(k)} = \xi_K \mathbf{U}^H \left(\xi_K \mathbf{I}_{|\mathcal{L}|} + \mathbf{D} \right)^{-1} \mathbf{U} \mathbf{v}^{(k)}, \tag{23}$$

where $(\xi_K \mathbf{I}_{|\mathcal{L}|} + \mathbf{D})^{-1}$ is a trivial diagonal inversion, and the matrix-vector products by $\mathbf{U}$ and $\mathbf{U}^H$ (the 2D-DFT and its inverse) are not carried out explicitly but via the efficient ($O(|\mathcal{L}| \log |\mathcal{L}|)$) fast Fourier transform (FFT). Notice that this can be seen as a smoothing operation, applied to each $\mathbf{v}^{(k)}$ in the discrete Fourier domain. Since the computational cost of the E-step is essentially $O(|\mathcal{L}|)$, as is obvious from (13), the leading cost of the proposed algorithm is $O(|\mathcal{L}| \log |\mathcal{L}|)$.

Finally, we should mention that the condition of periodic boundary conditions can be relaxed; in that case, the resulting matrix $\boldsymbol{\Delta}$ is block-Toeplitz with Toeplitz blocks, but not block-circulant. Nevertheless, it is still possible to embed a block-Toeplitz matrix into a larger block-circulant one, and still work in the DFT domain [15].

[2] Recall that a circulant matrix is characterized by the fact that each row is a circularly shifted version of the first (or any other) row.

3.6 Summary of the Algorithm

We now summarize the algorithm, showing that it is in fact very simple.

Inputs: Observed image $\mathbf{x}$, number of regions K, observation models $p(\cdot|\boldsymbol{\phi}^{(k)})$, matrix $\mathbf{W}$ or $\boldsymbol{\Delta}$, stopping threshold ε, number of inner iterations r.
Output: Estimates $\widehat{\mathbf{z}}^{(k)}$, for $k = 1, ..., K-1$.
Initialization: For $k = 1, ..., K-1$, set $\widehat{\mathbf{z}}^{(k)} = 0$.
Step 1: Run the E-step (13), producing K images $\{\widehat{\mathbf{y}}^{(0)}, ..., \widehat{\mathbf{y}}^{(K-1)}\}$.
Step 2: Store the current estimate: $\widehat{\mathbf{z}}_{\text{old}} \leftarrow \widehat{\mathbf{z}}$.
Step 3: Repeat r times (for $k = 1, ..., K-1$):
Step 3.a: Compute the images $\mathbf{d}^{(k)}$ (according to (20)).
Step 3.b: Compute the images $\mathbf{v}^{(k)} = \widehat{\mathbf{z}}^{(k)} + \mathbf{d}^{(k)}/\xi_K$ (see (19)).
Step 3.c: Compute $\widehat{\mathbf{z}}^{(k)}_{\text{new}}$ according to (23). Update $\widehat{\mathbf{z}}^{(k)} \leftarrow \widehat{\mathbf{z}}^{(k)}_{\text{new}}$.
Step 3.d: Go back to **Step 3.a**.
Step 4: If $\max_k \|\widehat{\mathbf{z}}^{(k)}_{\text{old}} - \widehat{\mathbf{z}}^{(k)}\|_\infty < \varepsilon$, then stop; otherwise, return to **Step 1**.

4 Extensions

4.1 Unsupervised and Semi-supervised Segmentation

The model and algorithm above described can be extended to the unsupervised case, where the parameters $\boldsymbol{\phi}^{(k)}$ of the observation models $p(\cdot|\boldsymbol{\phi}^{(k)})$ are considered unknown. In this case, the full posterior in (9) has to be modified to

$$p(\mathbf{z}, \boldsymbol{\phi}, \mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})\, p(\mathbf{y}|\mathbf{z})\, p(\mathbf{z}). \tag{24}$$

where $\boldsymbol{\phi} = \{\boldsymbol{\phi}^{(0)}, ..., \boldsymbol{\phi}^{(K-1)}\}$, assuming the absence of any prior on $\boldsymbol{\phi}$ (although one could easily be considered with little additional cost). Let us adopt again the MMAP criterion, now jointly w.r.t. $\mathbf{z}$ and $\boldsymbol{\phi}$. The following observations can now be added to those made in Section 3.2:

- If $\mathbf{y}$ was observed, estimating $\boldsymbol{\phi}$ would be a simple ML parameter estimation problem, based on the complete log-likelihood $\log p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi})$.
- The complete log-likelihood (see (1)) is linear w.r.t. the missing variables $\mathbf{y}$:

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

The algorithm presented in Section 3.6 can thus be modified by inserting an extra step, say between steps 2 and 3:

Step 2.5: Update the observation model parameters according to the following weighted ML criterion:

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

If, for example, the feature densities are Gaussians, $p(\cdot|\boldsymbol{\phi}^{(k)}) = \mathcal{N}(\cdot|\boldsymbol{\mu}^{(k)}, \mathbf{C}^{(k)})$, these update equations coincide with those of the EM algorithm for Gaussian mixture estimation:

$$\widehat{\boldsymbol{\mu}}^{(k)} = \frac{\sum_{i \in \mathcal{L}} \widehat{y}_i^{(k)} \, x_i}{\sum_{i \in \mathcal{L}} \widehat{y}_i^{(k)}}, \quad \widehat{\mathbf{C}}^{(k)} = \frac{\sum_{i \in \mathcal{L}} \widehat{y}_i^{(k)} \, (x_i - \widehat{\boldsymbol{\mu}}^{(k)})(x_i - \widehat{\boldsymbol{\mu}}^{(k)})^T}{\sum_{i \in \mathcal{L}} \widehat{y}_i^{(k)}}. \tag{25}$$

In the semi-supervised case, instead of previous knowledge of $\{\boldsymbol{\phi}^{(0)},...,\boldsymbol{\phi}^{(K-1)}\}$, one is given a subset of pixels for which the exact true label/region is known. In this case, the EM algorithm derived 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 cases, the log-posterior is no longer concave, and the results will depend critically on the initialization.

4.2 Discriminative Features

The formulation presented above (and most of the work on probabilistic segmentation) uses what can be classified as "generative feature models": each $p(\cdot|\boldsymbol{\phi})$ is a probabilistic model that is assumed to describe how features/pixel values are generated in each region. However, discriminative models, such as logistic regression, Gaussian processes [31], support vector machines, or boosting (see references in [12]) are currently considered the state-of-the-art in classification.

Observe that all the EM segmentation algorithm requires, in the E-step defined in (13), is the posterior class probabilities, given the pixel values and the current estimates $\widehat{\mathbf{z}}^{(k)}$. These estimates provide some prior class probabilities in (13). Consider a probabilistic discriminative classifier, that is, a classifier that, for each pixel x_i, provides estimates of the posterior class probabilities $p(y_i^{(k)} = 1|x_i)$, for $k = 0, ..., K-1$ (this can be obtained, *e.g.*, by logistic regression, or a tree classifier). 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 $p(y_i^{(k)} = 1)$, for $k = 0, ..., K-1$, by computing

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

This procedure allows using a pre-trained probabilistic discriminative classifier, which yields $p(y_i^{(k)} = 1|x_i)$, 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

In the first experiment, we consider a simple synthetic segmentation problem, with known class models. Each of the four regions follows a Gaussian distribution

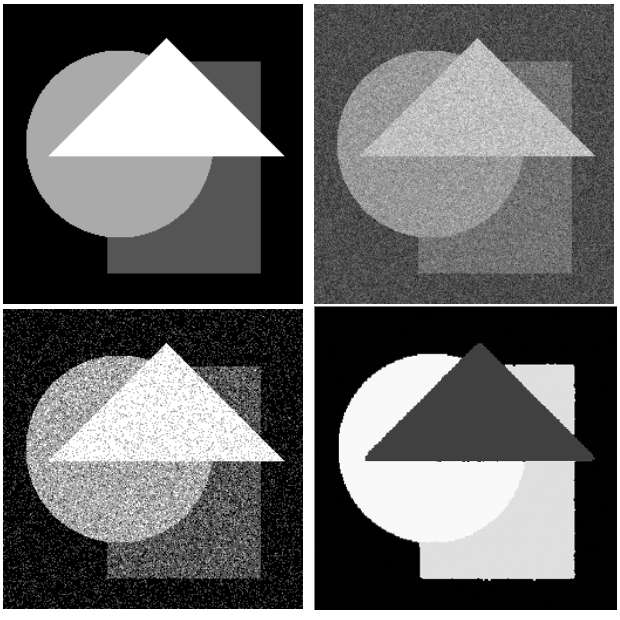

Fig. 1. Top row: true regions and observed image. Bottom row: maximum likelihood segmentation and the one obtained by our algorithm.

with standard deviation 0.6 and means 1, 2, 3, and 4. We have used (in this and all the following examples) $r = 4$, and $\varepsilon = 0.001$. We choose the simplest possible GMRF prior: $w_{i,j} = \gamma$, if j is one of the four nearest neighbors of i, and is zero otherwise. The true regions, observed image, the maximum likelihood segmentation (obtained by maximizing (1) with respect to $\mathbf{y}$), and the (hard, obtained via (11)) segmentation produced by our algorithm are shown in Fig. 1. This is comparable to what would be obtained by an MRF-based method; however, it must be stressed that the algorithm herein proposed is optimal (in the sense that we are minimizing a convex objective function), fully deterministic, and fast (due to the use of the FFT-based M-step). This result illustrates the ability of the proposed method to use Gaussian priors to regularize image segmentation via the logistic modelling approach, producing well defined boundaries.

In Fig. 2 we show the final estimates $\widehat{\mathbf{z}}^{(1)}$, $\widehat{\mathbf{z}}^{(2)}$, and $\widehat{\mathbf{z}}^{(3)}$ as well as the corresponding $\widehat{\mathbf{y}}^{(1)}$, $\widehat{\mathbf{y}}^{(2)}$, $\widehat{\mathbf{y}}^{(3)}$, and $\widehat{\mathbf{y}}^{(4)}$, obtained from the $\widehat{\mathbf{z}}^{(k)}$ via the logistic model (3). Notice the higher uncertainty near the region boundaries. The hard segmentation shown in Fig. 1 was obtained by choosing, for each site, the maximum of the four $\widehat{\mathbf{y}}^{(k)}$ images.

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

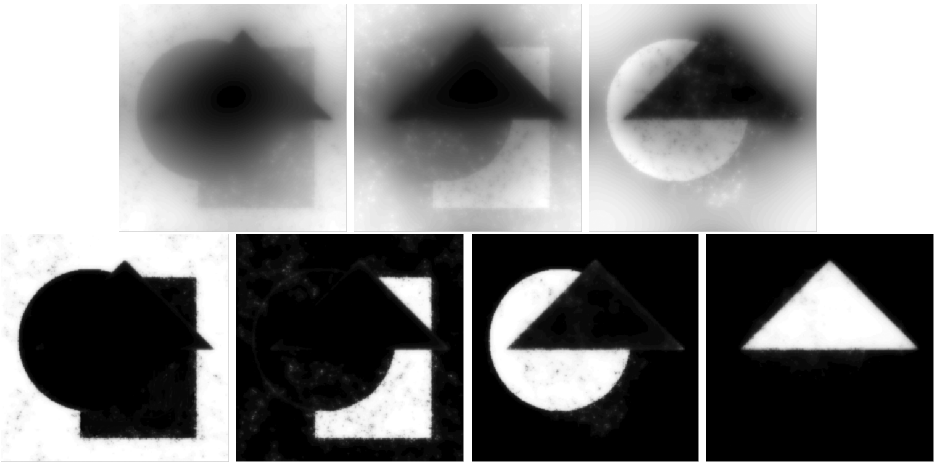

Fig. 2. Top row: final estimates $\widehat{\mathbf{z}}^{(1)}$, $\widehat{\mathbf{z}}^{(2)}$, and $\widehat{\mathbf{z}}^{(3)}$. Bottom row: corresponding $\widehat{\mathbf{y}}^{(1)}$, $\widehat{\mathbf{y}}^{(2)}$, $\widehat{\mathbf{y}}^{(3)}$, and $\widehat{\mathbf{y}}^{(4)}$, obtained by the logistic model (3).

Fig. 3. Observed image, maximum likelihood segmentation, and segmentation obtained by our algorithm

Fig. 4. Observed image, maximum likelihood segmentation, and segmentation obtained by our algorithm

For segmentation of real images, the results depend critically on the features and feature models used, and that is not the focus of this paper. We will only show two examples of color image segmentation ($d = 3$), using Gaussian densities for each region. In Fig. 3, the goal is to segment the image into three regions: clothe, skin, and background. Fig. 4 shows a figure-ground segmentation problem. The results shown were produced by the unsupervised version of our algorithm, initialized with the ML segmentations which result from fitting mixtures of Gaussians to the observed (RGB) pixels.

6 Summary and Conclusions

A new formulation for Bayesian image segmentation was introduced. This approach allows using priors for continuous-valued fields as regularizers for image segmentation; in particular, it was used with Gaussian field priors, which (if stationary) can be easily and efficiently manipulated in the frequency domain using the FFT algorithm. 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 [10].

Appendix: Proof of Lemma 1

Recall (see (17)) that

$$\mathbf{B} = \frac{1}{2}\left(\mathbf{I}_{K-1} - \frac{\mathbf{1}_{K-1}\,\mathbf{1}_{K-1}^T}{K}\right) \otimes \mathbf{I}. \tag{26}$$

For $K = 2$, it is obvious that $\mathbf{B} = \mathbf{I}/4$.

For $K > 2$, the matrix inequality $\mathbf{I}/2 \succeq \mathbf{B}$ is equivalent to $\lambda_{\min}(\mathbf{I}/2 - \mathbf{B}) \geq 0$. Now, since $\lambda_i(\mathbf{I}/2 - \mathbf{B}) = (1/2) - \lambda_i(\mathbf{B})$, we need to show that $\lambda_{\max}(\mathbf{B}) \leq (1/2)$.

To study the eigenvalues of $\mathbf{B}$, the following fact (see, *e.g.*, [21]) is used: let $\mathbf{M}$ and $\mathbf{P}$ be $m \times m$ and $p \times p$ matrices, with eigenvalues $\{\lambda_1, ..., \lambda_m\}$ and $\{\gamma_1, ..., \gamma_p\}$, respectively; then, $\mathbf{M} \otimes \mathbf{P}$ has eigenvalues $\{\lambda_i\,\gamma_j,\ i = 1, ..., m,\ j = 1, ..., p\}$. Since $\mathbf{1}$ is a vector with $K - 1$ ones, $\mathbf{1}\,\mathbf{1}^T$ is a rank-1 matrix with eigenvalues $\{0, ..., 0, K - 1\}$; thus, the eigenvalues of $(\mathbf{I} - (1/K)\,\mathbf{1}\,\mathbf{1}^T)$ are $\{1, ..., 1, 1/K\}$. Because the eigenvalues of $\mathbf{I}$ are of course all ones, the maximum eigenvalue of $\mathbf{B}$ is $\lambda_{\max}(\mathbf{B}) = 1/2$. ■

References

1. N. Balram and J. Moura, "Noncausal Gauss-Markov random fields: parameter structure and estimation", *IEEE Trans. Information Theory*, vol. 39, pp. 1333–1355, 1993.
2. J. Bernardo and A. Smith, *Bayesian Theory*, J. Wiley & Sons, 1994.
3. D. Böhning. "Multinomial logistic regression algorithm." *Annals Inst. Stat. Math.*, vol. 44, pp. 197–200, 1992.
4. D. Böhning and B. Lindsay. "Monotonicity of quadratic-approximation algorithms." *Annals Inst. Stat. Math.*, vol. 40, pp. 641–663, 1988.
5. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts." *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 23, pp. 1222–1239, 2001.
6. F. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.
7. G. Cross and A. Jain. "Markov random field texture models." *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 5, pp. 25-39, 1983.
8. H. Derin and H. Elliot. "Modelling and segmentation of noisy and textured images in Gibbsian random fields." *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 9 , pp. 39-55, 1987.
9. M. Figueiredo, "Bayesian image segmentation using wavelet-based priors", *Proc. of IEEE CVPR'2005*, San Diego, CA, 2005.
10. M. Figueiredo and A.K.Jain. "Unsupervised learning of finite mixture models." *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 24, pp. 381-396, 2002.
11. R. Haralick, K. Shanmugan, and I. Dinstein, "Textural features for image classification." em IEEE Trans. Syst., Man, and Cybernetics, vol. 8, pp. 610-621, 1973.
12. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Verlag, New York, 2001.
13. L. Hermes and J. Buhmann, "A minimum entropy approach to adaptive image polygonization," *IEEE Trans. Image Proc.*, vol. 12, pp. 1243–1258, 2003.
14. T. Hofmann, J. Puzicha, and J. Buhmann. "Unsupervised texture segmentation in a deterministic annealing framework," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 20, pp. 803–818, 1998.
15. A. Jain. *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.
16. A. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters." *Pattern Recognition*, vol. 24, pp. 1167-1186, 1991.
17. J. Kim, J. Fisher, A. Yezzi, M. Çetin, and A. Willsky, "A nonparametric statistical method for image segmentation using information theory and curve evolution," *IEEE Trans. Image Proc.*, to appear, 2005.
18. 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.
19. K. Lange, D. Hunter, and I. Yang. "Optimization transfer using surrogate objective functions." *Jour. Comp. Graph. Stat.*, vol. 9, pp. 1–59, 2000.
20. S. Z. Li, *Markov Random Field Modelling in Computer Vision*. Springer Verlag, 2001.
21. J. Magnus and H. Neudecker. *Matrix Differential Calculus*. John Wiley & Sons, 1988.
22. J. Marroquin, E. Santana, and S. Botello. "Hidden Markov measure field models for image segmentation." *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 25, pp. 1380–1387, 2003.

23. G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997.
24. R. Nowak and M. Figueiredo, "Unsupervised progressive parsing of Poisson fields using minimum description length criteria," *Proc. IEEE ICIP'99,* Kobe, Japan, vol. II, pp. 26-29, 1999.
25. T. Randen and J. Husoy. "Filtering for texture classification: a comparative study." *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 21, pp. 291-310, 1999.
26. D. Martin, C. Fowlkes, and J. Malik. "Learning to detect natural image boundaries using local brightness, color and texture cues." *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 26, pp. 530-549, 2004.
27. E. Sharon, A. Brandt, R. Basri. "Segmentation and boundary detection using multiscale intensity measurements." *Proc. IEEE CVPR*, vol. I, pp. 469-476, Kauai, Hawaii, 2001.
28. J. Shi and J. Malik, "Normalized cuts and image segmentation." *IEEE Trans. Patt. Anal. Mach. Intell.*, vol.22, pp. 888-905, 2000.
29. M. Unser, "Texture classification and segmentation using wavelet frames." *IEEE Trans. Image Proc.*, vol. 4, pp. 1549 1560, 1995.
30. Y. Weiss, "Segmentation using eigenvectors: a unifying view." *Proc. Intern. Conf. on Computer Vision – ICCV'99*, pp. 975-982, 1999.
31. C. Williams and D. Barber. "Bayesian classification with Gaussian priors." *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 20, pp. 1342–1351, 1998.
32. Z. Wu and R. Leahy, "Optimal graph theoretic approach to data clustering: theory and its application to image segmentation." *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 15, pp. 1101-1113, 1993.
33. R. Zabih and V. Kolmogorov, "Spatially coherent clustering with graph cuts." *Proc. IEEE-CVPR*, vol. II, pp. 437–444, 2004.
34. S. C. Zhu and A. Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 18, pp. 884–900, 1996.