

The Infinite Hidden Markov Random Field Model

Sotirios P. Chatzis and Gabriel Tsechpenakis

Abstract

Dirichlet process (DP) mixture models have recently emerged in the cornerstone of nonparametric Bayesian statistics as promising candidates for clustering applications where the number of clusters is unknown a priori. Hidden Markov random field (HMRF) models are parametric statistical models widely used for image segmentation, as they appear naturally in problems where a spatially-constrained clustering scheme is asked for. A major limitation of HMRF models concerns the automatic selection of the proper number of their states, i.e. the number of segments derived by the image segmentation procedure. Typically, for this purpose, various likelihood based criteria are employed. Nevertheless, such methods often fail to yield satisfactory results, exhibiting significant overfitting proneness. Recently, higher order conditional random field models using potentials defined on superpixels have been considered as alternatives tackling these issues. Still, these models are in general computationally inefficient, a fact that limits their widespread adoption in practical applications. To resolve these issues, in this paper we introduce a novel, nonparametric Bayesian formulation for the HMRF model, the infinite HMRF model. We describe an efficient variational Bayesian inference algorithm for the proposed model, and we apply it to a series of image segmentation problems, demonstrating its advantages over existing methodologies.

1. Introduction

Hidden Markov random field (HMRF) models have turned out to be a significant tool for image segmentation, as they provide a powerful and formal way to introduce the information about the mutual influences among image pixels into the image segmentation procedure [13]. A significant limitation of HMRF models as image segmentation tools concerns the automatic determination of the optimal number of their states, and, hence, of the derived image segments. Typically, likelihood-based criteria, inspired by the popular Bayesian information criterion (BIC) for finite mixture models, are employed [24]. For example, in [27], the pseudolikelihood information criterion (PLIC) is proposed,

based on the notion of pseudolikelihood introduced in [22]. In [12], a mean-field principle is utilized to derive a suitable BIC approximation for HMRF models. Nevertheless, likelihood-based model selection methods tend in general to yield noisy model size estimates, being prone to overfitting, and, hence, often leading to over-segmentation [8]. Furthermore, their use entails training of multiple models (to select from), a procedure which can be applied only up to a limited extent, due to its computational demands.

More recently, many researchers have considered higher order conditional random field (CRF) models, using potentials defined on sets of pixels (superpixels), obtained by prior application of unsupervised segmentation algorithms. Initial work on this field has shown quite promising results (e.g., [23]). However, such models suffer from the excessively high computational load imposed by their inference algorithms. Traditional inference algorithms for CRFs, such as belief propagation and graph-cuts, become computationally prohibitive for higher order CRFs. Efforts for mitigating these shortcomings have been only partly successful. For example, *Lan et al.* [18] propose a class of approximation methods for belief propagation to make efficient inference possible in higher order Markov random fields (MRFs). In [20], a methodology is proposed for conducting belief propagation in a computationally tractable manner in graphical models containing moderately large cliques. Nevertheless, these methods continue to be computationally cumbersome, a fact which limits their potential for widespread adoption in practical applications.

To resolve these open issues, in this paper we exploit a relatively new tool in machine learning literature, Dirichlet process mixture (DPM) models [17]. DPM models have emerged as a nonparametric alternative to finite mixture models, with, theoretically, a countably infinite number of mixture components [1]. This thus obtained nonparametric Bayesian formulation eliminates the need of doing inference on the number of mixture components necessary to represent the modeled data. Indeed, as a result of the fitting procedure, rather than selecting a fixed number of mixture components, the nonparametric Bayesian inference scheme induced by a DPM model yields a posterior distribution on the proper number of model component densities [4].

Under this motivation, in this work we propose a novel extension of DPM models to incorporate a spatial distri-

bution over the modeled data similar to that introduced by HMRF models. Dually, we introduce a nonparametric formulation for HMRF models, where the prior probabilities of the modeled data being generated from a model state are considered to follow jointly a DP distribution and a Markov random field (Gibbsian) distribution with a countably infinite number of states; we shall be referring to this new model as the infinite HMRF (iHMRF) model.

The remainder of this paper is organized as follows. In Section 2, a brief overview of DPM and HMRF models is provided. In Section 3, the proposed infinite HMRF model is formulated, and an elegant truncated variational Bayesian inference algorithm for the model is derived. In Section 4, we thoroughly evaluate the efficacy of the proposed model through a number of image segmentation experiments. Finally, in the concluding section, we summarize and discuss our results.

2. Theoretical Background

2.1. Hidden Markov random field models

We consider an alphabet $Q = \{1, \dots, K\}$. Let S be a finite index set, $S = \{1, \dots, N\}$; we shall refer to this set, S , as the set of sites or locations. Let us consider for every site $j \in S$ a finite space \mathcal{X}_j of states x_j , such as $\mathcal{X}_j = \{x_j : x_j \in Q\}$. The product space $\mathcal{X} = \prod_{j=1}^N \mathcal{X}_j$ will be denoted as the space of the configurations of the state values of the considered sites set, $\mathbf{x} = (x_j)_{j \in S}$. A strictly positive probability distribution, $p(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, on the product space \mathcal{X} is called a random field [19]. Let ∂ denote a neighborhood system on S , i.e. a collection $\partial = \{\partial_j : j \in S\}$ of sets, such as $j \notin \partial_j$ and $l \in \partial_j$ if and only if $j \in \partial_l \forall l, j \in S$. Then, the previously considered random field, $p(\mathbf{x})$, is a Markov random field with respect to the introduced neighborhood system ∂ if [13]

$$p(x_j | \mathbf{x}_{S-\{j\}}) = p(x_j | \mathbf{x}_{\partial_j}) \quad \forall j \in S \quad (1)$$

The distribution $p(\mathbf{x})$ of a Markov random field can be shown to be of a Gibbsian form; that is, we have [10]

$$p(\mathbf{x} | \beta) \triangleq \frac{1}{W(\beta)} \exp \left(- \sum_{c \in \mathcal{C}} V_c(\mathbf{x} | \beta) \right) \quad (2)$$

where, β is the inverse temperature of the model, $W(\beta)$ is a normalizing constant, $V_c(\mathbf{x} | \beta)$ are the clique potentials of the model, and \mathcal{C} is the set of the cliques included in the model neighborhood system. Let us, now, consider a second random field, $p(\mathbf{y})$, the d -dimensional state space \mathcal{Y} of which is also indexed by the supposed set of sites S , and is given by

$$\mathcal{Y} = \prod_{j=1}^N \mathcal{Y}_j, \quad \mathcal{Y}_j = \{\mathbf{y}_j : \mathbf{y}_j \in \mathbb{R}^d\}$$

We denote as \mathbf{y} a realization of this field, and it holds $\mathbf{y} = \{\mathbf{y}_j\}_{j=1}^N$. The pair of the above defined random field, $p(\mathbf{y})$, and the supposed Markov random field, $p(\mathbf{x})$, with

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y} | \mathbf{x}) p(\mathbf{x})$$

is widely-known as a hidden Markov random field model [7]. Here, we adopt the typical assumption

$$p(\mathbf{y} | \mathbf{x}) = \prod_{j=1}^N p(\mathbf{y}_j | x_j) \quad (3)$$

which provides a convenient approximation of the posterior field $p(\mathbf{x} | \mathbf{y})$, still guaranteeing its Markovianity; however, more complicated assumptions might be also employed. The field $p(\mathbf{y})$ is called the observed (or emitted) random field of the observed data related to the sites $j \in S$, \mathcal{Y} is called the space of observations, and the field $p(\mathbf{x})$ is called the Markov random field of the unobservable state variables related to the sites $j \in S$ [9]. Usually, the likelihoods $p(\mathbf{y}_j | x_j)$ are taken as multivariate Gaussians

$$p(\mathbf{y}_j | x_j; \Theta_{x_j}) = \mathcal{N}(\mathbf{y}_j | \boldsymbol{\mu}_{x_j}, \mathbf{R}_{x_j}) \quad (4)$$

where, $\boldsymbol{\mu}_i$ and \mathbf{R}_i are the mean and precision matrix of the Gaussian likelihood function, and $\Theta_i = \{\boldsymbol{\mu}_i, \mathbf{R}_i\}$.

A significant problem of HMRF models concerns computation of the posterior probabilities $p(x_j | \mathbf{y})$ and $p(\mathbf{x} | \mathbf{y})$. Usually, these quantities are obtained by means of Bayesian sampling, e.g. using Markov chain Monte Carlo methods [6]. Nevertheless, such methods require a large amount of computation. An alternative to these approaches, yielding good estimates of the Markov posteriors with considerably better scalability in terms of computational cost, is the mean-field approximation [9, 29]. It is based on the idea of neglecting the fluctuations of the sites interacting with a considered site, so that the resulting system behaves as one composed of independent variables for which computation becomes tractable. That is, given an estimate $\hat{\mathbf{x}}$ of the unknown site labels vector \mathbf{x} , obtained by means of a stochastic restoration criterion, such as the iterative conditional modes (ICM) or the marginal posterior modes (MPM) algorithm (see, e.g., [9, 13]), we make the hypothesis [22]

$$p(\mathbf{x} | \beta) = \prod_{j=1}^N p(x_j | \hat{\mathbf{x}}_{\partial_j}; \beta) \quad (5)$$

where

$$p(x_j = i | \hat{\mathbf{x}}_{\partial_j}; \beta) = \frac{\exp(-\sum_{c \ni j} V_c(\tilde{\mathbf{x}}_{ij} | \beta))}{\sum_{h=1}^K \exp(-\sum_{c \ni j} V_c(\tilde{\mathbf{x}}_{hj} | \beta))} \quad (6)$$

$\tilde{\mathbf{x}}_{ij} \triangleq (x_j = i, \hat{\mathbf{x}}_{\partial_j})$, and $\hat{\mathbf{x}}_{\partial_j}$ is the estimate of the j th site's neighborhood.

2.2. Dirichlet process mixture models

Dirichlet process (DP) models were first introduced by Ferguson [11]. A DP is characterized by a base distribution G_0 and a positive scalar α , usually referred to as the innovation parameter, and is denoted as $\text{DP}(G_0, \alpha)$. Essentially, a DP is a distribution placed over a distribution. Let us suppose we randomly draw a sample distribution G from a GP, and, subsequently, we independently draw N random variables $\{\Theta_n^*\}_{n=1}^N$ from G :

$$G|\{G_0, \alpha\} \sim \text{DP}(G_0, \alpha) \quad (7)$$

$$\Theta_n^*|G \sim G, \quad n = 1, \dots, N \quad (8)$$

Integrating out G , the joint distribution of the variables $\{\Theta_n^*\}_{n=1}^N$ can be shown to exhibit a clustering effect. Specifically, given the first $N-1$ samples of G , $\{\Theta_n^*\}_{n=1}^{N-1}$, it can be shown that a new sample Θ_N^* is either (a) drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+N-1}$, or (b) is selected from the existing draws, according to a multinomial allocation, with probabilities proportional to the number of the previous draws with the same allocation [3]. Let $\{\Theta_c\}_{c=1}^K$ be the set of distinct values taken by the set $\{\Theta_n^*\}_{n=1}^N$. Denoting as f_c^{N-1} the number of values in $\{\Theta_n^*\}_{n=1}^{N-1}$ that equal to Θ_c , the distribution of Θ_N^* given $\{\Theta_n^*\}_{n=1}^{N-1}$ can be shown to be of a Pólya urn form [3]

$$p(\Theta_N^*|\{\Theta_n^*\}_{n=1}^{N-1}, G_0, \alpha) = \frac{\alpha G_0 + \sum_{c=1}^K f_c^{N-1} \delta_{\Theta_c}}{\alpha + N - 1} \quad (9)$$

where δ_{Θ_c} denotes the distribution concentrated at a single point Θ_c . These results illustrate two key properties of the DP scheme. First, the innovation parameter α plays a key-role in determining the number of distinct parameter values. A larger α induces a higher tendency of drawing new parameters from the base distribution G_0 ; indeed, as $\alpha \rightarrow \infty$ we get $G \rightarrow G_0$. On the contrary, as $\alpha \rightarrow 0$ all $\{\Theta_n^*\}_{n=1}^N$ tend to cluster to a single random variable. Second, the more often a parameter is shared, the more likely it will be shared in the future.

A characterization of the (unconditional) distribution of the random variable G drawn from a Dirichlet process $\text{DP}(G_0, \alpha)$ is provided by the stick-breaking construction of Sethuraman [25]. Consider two infinite collections of independent random variables $\mathbf{v} = (v_c)_{c=1}^\infty$, $\{\Theta_c\}_{c=1}^\infty$, where the v_c are drawn from the Beta distribution $\text{Beta}(1, \alpha)$, and the Θ_c are independently drawn from the base distribution G_0 . The stick-breaking representation of G is then given by [25]

$$G = \sum_{c=1}^{\infty} \pi_c(\mathbf{v}) \delta_{\Theta_c} \quad (10)$$

where

$$\pi_c(\mathbf{v}) = v_c \prod_{j=1}^{c-1} (1 - v_j) \in [0, 1] \quad (11)$$

and $\sum_{c=1}^{\infty} \pi_c(\mathbf{v}) = 1$. The stick-breaking representation of the DP makes clear that the random variable G drawn from a DP is discrete. It shows explicitly that the support of G consists of a countably infinite sum of atoms located at Θ_c , drawn independently from G_0 . It is also apparent that the innovation parameter α controls the mean value of the stick variables, v_c , as a hyperparameter of their prior distribution; hence, it regulates the effective number of the distinct values of the drawn atoms [25].

Under the stick-breaking representation of the Dirichlet process, the atoms Θ_c , drawn independently from the base distribution G_0 , can be seen as the parameters of the component distributions of a mixture model comprising an unbounded number of component densities, with mixing proportions $\pi_c(\mathbf{v})$. This way, DP mixture (DPM) models are formulated [1].

Let $\mathbf{y} = \{\mathbf{y}_n\}_{n=1}^N$ be a set of observations modeled by a DPM model. Then, each one of the observations \mathbf{y}_n is assumed to be drawn from its own probability density function $p(\mathbf{y}_n|\Theta_n^*)$ parametrized by the parameter set Θ_n^* . All Θ_n^* follow a common DP prior, and given the discreteness of G , may share the same value Θ_c with probability $\pi_c(\mathbf{v})$. Introducing the indicator variables $\mathbf{x} = (x_n)_{n=1}^N$, with $x_n = c$ denoting that Θ_n^* takes on the value of Θ_c , the modeled data set \mathbf{y} can be described as arising from the process

$$\mathbf{y}_n|x_n = c \sim p(\mathbf{y}_n|\Theta_c) \quad (12)$$

$$x_n|\boldsymbol{\pi}(\mathbf{v}) \sim \text{Mult}(\boldsymbol{\pi}(\mathbf{v})) \quad (13)$$

$$v_c|\alpha \sim \text{Beta}(1, \alpha) \quad (14)$$

$$\Theta_c|G_0 \sim G_0 \quad (c = 1, \dots, \infty) \quad (15)$$

where $\boldsymbol{\pi}(\mathbf{v}) = (\pi_c(\mathbf{v}))_{c=1}^\infty$ is given by (11), and $\text{Mult}(\boldsymbol{\pi}(\mathbf{v}))$ is a Multinomial distribution over $\boldsymbol{\pi}(\mathbf{v})$.

3. The Infinite HMRF Model

Let $\mathbf{y} = \{\mathbf{y}_n\}_{n=1}^N$ be a set of d -dimensional multivariate observations associated with N sites arranged on a 2D lattice. We postulate a Gaussian DPM model for the representation of this data set. A drawback of this model is its lack of an explicit consideration of the spatial dynamics (interdependencies) between the neighboring sites on the input lattice. To introduce such spatial constraints into our model, we further impose an additional MRF (Gibbsian) distribution over the component (Gaussian) densities of the postulated model. For convenience, we consider a simplified, pointwise MRF distribution, obtained by application of the mean-field like approximation, as described in Section 2.1. Conversely, the formulated model might be considered as an HMRF model with countably infinite states, where the (approximate) Gibbsian prior over the states generating the data is conjunct with a Dirichlet process. Then, denoting as $\mathbf{x} = (x_n)_{n=1}^N$ the labels of the sites, indicating the states of

the postulated model emitting the associated with the sites observable data, we have

$$\mathbf{y}_n | x_n = c; \Theta_c \sim \mathcal{N}(\boldsymbol{\mu}_c, \mathbf{R}_c) \quad (16)$$

$$p(\mathbf{x}) = \prod_{n=1}^N p(x_n | \boldsymbol{\pi}(\mathbf{v}), \hat{\mathbf{x}}_{\partial_n}) \quad (17)$$

and

$$p(x_n = c | \boldsymbol{\pi}(\mathbf{v}), \hat{\mathbf{x}}_{\partial_n}) \propto p(x_n = c | \hat{\mathbf{x}}_{\partial_n}; \beta) p(x_n = c | \boldsymbol{\pi}(\mathbf{v})) \quad (18)$$

where $p(x_n = c | \hat{\mathbf{x}}_{\partial_n}; \beta)$ are the approximate pointwise prior probabilities of the model states due to the imposed MRF, given by (6), while $p(x_n = c | \boldsymbol{\pi}(\mathbf{v}))$ are the prior probabilities of the model states stemming from the imposed Dirichlet process, given by (13) and (11).

We denote as the infinite hidden Markov random field (iHMRF) model, the generative model of eqs. (16) - (18), defined on data deriving from a 2D lattice of observations.

Inference for DPM-based models can be conducted under a Bayesian setting, typically by means of variational Bayes (e.g., [5]), or Monte Carlo techniques (e.g., [21]). Here, we prefer a variational Bayesian approach, due to its considerably better scalability in terms of computational costs, which becomes of a major importance when having to deal with large data corpora (as in the case of image pixels). Bayesian inference involves introduction of a set of appropriate priors over the model parameters, and derivation of the corresponding (approximate) posterior densities. We choose conjugate-exponential priors, as this selection greatly simplifies inference and interpretability [2]. Hence, we impose a joint Normal-Wishart distribution over the means and the precisions of the state distributions

$$p(\Theta_c) = \mathcal{NW}(\boldsymbol{\mu}_c, \mathbf{R}_c | \lambda_c, \mathbf{m}_c, \omega_c, \boldsymbol{\Psi}_c) \quad (19)$$

Additionally, taking under consideration the effect of the innovation hyperparameter α on the number of effective mixture components of a DPM model, we choose to also impose a (hyper-)prior over the innovation hyperparameter α of the iHMRF model. We use a Gamma prior with

$$p(\alpha) = \mathcal{G}(\alpha | \eta_1, \eta_2) \quad (20)$$

Our variational Bayesian inference formalism for the iHMRF model consists in derivation of a family of variational posterior distributions $q(\cdot)$ which approximate the true posterior distribution over the infinite sets $\mathbf{v} = (v_c)_{c=1}^{\infty}$, and $\{\boldsymbol{\mu}_c, \mathbf{R}_c\}_{c=1}^{\infty}$. Apparently, under this infinite dimensional setting, Bayesian inference is not tractable. For this reason, we employ a common strategy in DPM literature, formulated on the basis of a truncated stick-breaking representation of the DP [5]. That is, we fix a value K and we let the variational posterior over the v_i have the property

$q(v_K = 1) = 1$. In other words, we set $\pi_c(\mathbf{v})$ equal to zero for $c > K$. Note that, under this setting, the treated iHMRF model involves a full DP prior; truncation is not imposed on the model itself, but only on the variational distribution to allow for a tractable inference procedure. Hence, the truncation level K is a variational parameter which can be freely set, and not part of the prior model specification.

Let $W = \{\mathbf{v}, \alpha, \mathbf{x}, \boldsymbol{\mu}_c, \mathbf{R}_c\}_{c=1}^K$ be the set of all hidden variables and unknown parameters of the iHMRF model over which a prior distribution has been imposed, and Ξ be the set of the hyperparameters of the imposed priors, $\Xi = \{\lambda_c, \mathbf{m}_c, \omega_c, \boldsymbol{\Psi}_c, \eta_1, \eta_2\}_{c=1}^K$. Variational Bayesian inference consists in the introduction of an arbitrary distribution $q(W)$ to approximate the actual posterior $p(W | \Xi, \mathbf{y})$, which is computationally intractable [2]. Under this assumption, the log marginal likelihood (log evidence), $\log p(\mathbf{y})$, of the model yields [14]

$$\log p(\mathbf{y}) = \mathcal{L}(q) + \text{KL}(q || p) \quad (21)$$

where

$$\mathcal{L}(q) = \int q(W) \log \frac{p(\mathbf{y}, W | \Xi)}{q(W)} dW \quad (22)$$

and $\text{KL}(q || p)$ stands for the Kullback-Leibler (KL) divergence between the (approximate) variational posterior, $q(W)$, and the actual posterior, $p(W | \Xi, \mathbf{y})$. Since KL divergence is nonnegative, $\mathcal{L}(q)$ forms a strict lower bound of the log evidence, and would become exact if $q(W) = p(W | \Xi, \mathbf{y})$. Hence, by maximizing this lower bound $\mathcal{L}(q)$ (variational free energy) so that it becomes as tight as possible, not only do we minimize the KL-divergence between the true and the variational posterior, but we also implicitly integrate out the unknowns W .

Due to the conjugate exponential prior configuration of the iHMRF model, the variational posterior $q(W)$ is expected to take the same functional form as the prior, $p(W)$, [8], thus factorizing as

$$q(W) = q(\mathbf{x}) q(\alpha) \left(\prod_{c=1}^{K-1} q(v_c) \right) \prod_{c=1}^K q(\boldsymbol{\mu}_c, \mathbf{R}_c) \quad (23)$$

with

$$q(\mathbf{x}) = \prod_{n=1}^N q(x_n) \quad (24)$$

Then, the variational free energy of the model reads

$$\begin{aligned}
 \mathcal{L}(q) = & \sum_{c=1}^K \int d\mathbf{R}_c \int d\boldsymbol{\mu}_c \left[q(\boldsymbol{\mu}_c, \mathbf{R}_c) \right. \\
 & \times \log \frac{p(\boldsymbol{\mu}_c, \mathbf{R}_c | \lambda_c, \mathbf{m}_c, \omega_c, \boldsymbol{\Psi}_c)}{q(\boldsymbol{\mu}_c, \mathbf{R}_c)} \left. \right] \\
 & + \int d\alpha q(\alpha) \left\{ \log \frac{p(\alpha | \eta_1, \eta_2)}{q(\alpha)} \right. \\
 & + \sum_{c=1}^{K-1} \int dv_c q(v_c) \log \frac{p(v_c | a)}{q(v_c)} \left. \right\} \\
 & + \sum_{c=1}^K \sum_{n=1}^N q(x_n = c) \left\{ \log \frac{p(x_n = c | \hat{\mathbf{x}}_{\partial_n}; \beta)}{q(x_n = c)} \right. \\
 & + \int dv q(v) \log p(x_n = c | \boldsymbol{\pi}(v)) \\
 & \left. + \int \int d\mathbf{R}_c d\boldsymbol{\mu}_c q(\boldsymbol{\mu}_c, \mathbf{R}_c) \log p(\mathbf{y}_n | \Theta_c) \right\}
 \end{aligned} \tag{25}$$

Derivation of the variational posterior distribution $q(W)$ involves the maximization of the variational free energy $\mathcal{L}(q)$ over each one of the factors of $q(W)$ in turn, holding the others fixed, in an iterative manner [7]. By construction, this iterative, consecutive updating of the variational posterior is guaranteed to monotonically and maximally increase the free energy $\mathcal{L}(q)$, which functions as the convergence criterion for the inference algorithm [8]. An outline of the updates comprising the proposed variational Bayesian inference algorithm for the iHMRF model is provided in the Appendix.

4. Image Segmentation Using iHMRF

In this Section we provide the experimental evaluation of our method as an image segmentation tool. We use a subset of the MSRC-v2.0 database [26], a moderate difficulty data set, commonly used as benchmark for image segmentation and object recognition algorithms.

In evaluation of our method, an initial rough segmentation of the examined images is performed first, using the k -means algorithm. Typically, each image of the used data set does not contain more than 4 different object classes. For this reason, and to assess how capable the iHMRF model is of obtaining the correct number of object classes in an image, the number of clusters (object classes) computed by the k -means algorithm is set to 9. The k -means output is in the sequel used to initialize the iHMRF model, which, hence, is evaluated on the basis of both the number of the eventually retained object classes as well as the quality of the final image segments. We conduct image segmentation using color and texture image information, with the color information modeled as the R/G/B color components, and the

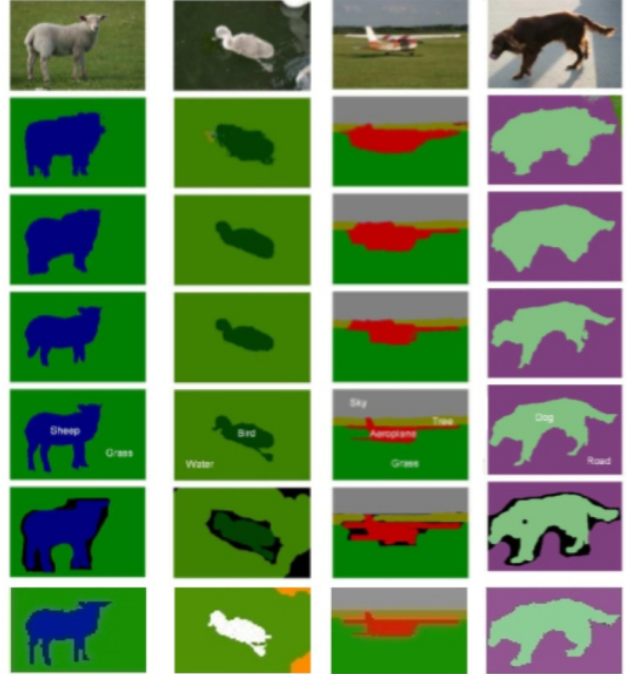


Figure 1: Qualitative evaluation. *First Row*: Original Images. *Second Row*: Unary likelihood labelling from Textonboost [26]. *Third Row*: Result obtained using a pairwise contrast preserving smoothness potential as described in [26]. *Fourth Row*: Results of the method in [15]. *Fifth Row*: Results of the method in [16]. *Sixth Row*: Hand labelled segmentations. *Seventh Row*: Results obtained by the iHMRF model.

texture information obtained by a Gabor wavelet decomposition. Finally, regarding selection of the clique potentials of the iHMRF model, we choose a simple Potts model with a second order (8-neighbor) neighborhood system.

In Figs. 1 and 2, we provide a qualitative comparison of the performance of the iHMRF model with a couple of state-of-the-art CRF based methods [15, 16, 26]. We observe that the proposed iHMRF model yields comparable and slightly better results than the state-of-the-art, high-order CRF method of [16]. More specifically, in Fig. 1 we notice that the iHMRF model retains better the right ear of the sheep (leftmost example), while, contrary to the competition, it manages to capture the small whitish details at the right (top and bottom) corners of the image with the bird (second column), better complying with the hand labeled segmentations. Finally, in the last two cases of the airplane and the dog images, our results are comparable to the best of our competitors.

Regarding Fig. 2, we underline that our method yields a better capture of the contour of the chair at its bottom (top example), whereas it retains the empty space between the

Table 1: Quantitative evaluation: Correct pixel classification (%) for the various object categories

	boat	chair	sheep	grass	airplane	cow	bird	road	sky
[26]	7	15	50	98	60	58	19	86	83
[28]	31	34	84	87	88	73	19	89	94
iHMRF	59	83	97	92	81	85	19	87	93

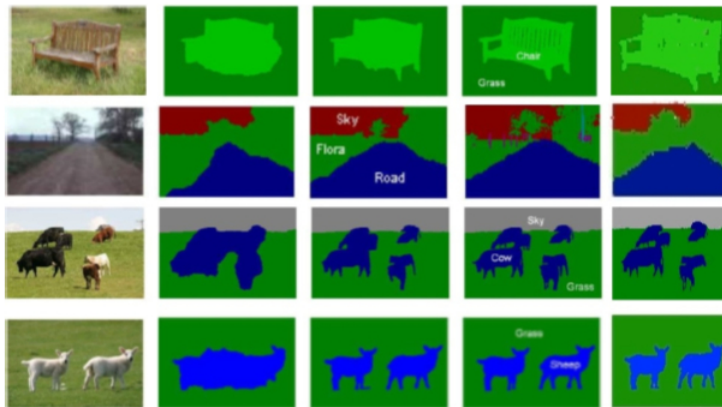


Figure 2: Qualitative evaluation. *First Column:* Original Images. *Second Column:* Segmentation result obtained using the pairwise CRF [26]. *Third Column:* Results obtained by higher order potentials [16]. *Fourth Column:* Hand labelled result used as ground truth. *Fifth Column:* Results obtained by the iHMRF model.

arm and the seat of the chair, contrary to the competition. Regarding the image with the two sheep (bottom example), we notice that our method manages to ignore the soft white spot near the leg of the one of the sheep, contrary to the method of [16] which includes it as a part of the sheep. For the rest of the illustrated experimental cases, our algorithm performance is comparable to that of [16].

In Table 1, we provide a quantitative evaluation of our method, conducted on the basis of the average proportion of correctly classified pixels in the obtained image segments, for each one of the considered object categories from the MSRC-v2.0 database. For comparison, we also cite the performance of the methods in [26] and [28]. As we observe, in several cases the iHMRF model manages to completely outperform the competition, yielding an outstanding result, while in others, it performs comparably to the competition.

Finally, we would also like to underscore that the iHMRF model manages to yield the aforementioned results for only a fraction of the computational time required by high-order CRF methods. For example, as the authors of [16] assert, to obtain an accurate segmentation of a 320×213 image, their method requires approximately 30 minutes. On the contrary, our method, unsophisticatedly implemented in MATLAB, and run on a Macintosh notebook, converges within an average time of 2.5 minutes. To allow for a better insight, in Fig. 3 we illustrate how the convergence time of our algorithm increases with the number of initial object cate-

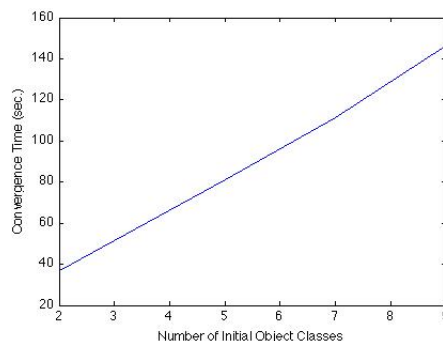


Figure 3: Average convergence time of the iHMRF model for the images of the MSRC-v2.0 database, for various numbers of initial classes.

gories. We observe that the computational complexity of the iHMRF model is linear to the number of initial classes, an advantage which might be of high significance in problems involving detection of objects of multiple categories from complex scenes.

5. Conclusions

In this work, we have described a nonparametric formulation of hidden Markov random field models for robust unsupervised image segmentation. Conversely, the intro-

duced method can be regarded as a spatially-constrained form of Dirichlet process mixture models, where the spatial constraints are encoded in the form of a simplified Gibbsian distribution. Our novel approach is effected by jointly imposing a Dirichlet process and an infinite state Gibbsian distribution over the 2D lattice of the sites associated with the observable data. Approximate inference in our model, needed given its complexity, has been conducted by application of an efficient truncated variational Bayesian algorithm. Experimental evaluation of the iHMRf model using benchmark data sets shows that it manages to yield competitive results with regard to existing state-of-the-art methods, yet with a much better scalability in terms of computational burden. A source code for the replication of the here presented results shall be provided through the website of the authors: <http://web.mac.com/soteri0s>.

6. Acknowledgments

We thank Dr. Robert K. Cowen, from Rosenstiel School of Marine and Atmospheric Sciences (RSMAS), University of Miami, for his support. This work was supported by the National Science Foundation (Geosciences Directorate), and the National Oceanic and Atmospheric Administration (NOAA) (Living Marine Resources Cooperative Science Center, and Advanced Sampling Technology Working Group).

Appendix

We begin with the posteriors over the DP parameters; we have

$$q(v_c) = \text{Beta}(\beta_{c,1}, \beta_{c,2}) \quad (26)$$

where

$$\beta_{c,1} = 1 + \sum_{n=1}^N q(x_n = c) \quad (27)$$

$$\beta_{c,2} = \langle \alpha \rangle + \sum_{c'=c+1}^K \sum_{n=1}^N q(x_n = c') \quad (28)$$

and

$$q(\alpha) = \mathcal{G}(\alpha | \hat{\eta}_1, \hat{\eta}_2) \quad (29)$$

where

$$\hat{\eta}_1 = \eta_1 + K - 1 \quad (30)$$

$$\hat{\eta}_2 = \eta_2 - \sum_{c=1}^{K-1} [\psi(\beta_{c,2}) - \psi(\beta_{c,1} + \beta_{c,2})] \quad (31)$$

and, hence $\langle \alpha \rangle = \frac{\hat{\eta}_1}{\hat{\eta}_2}$. Regarding the posteriors over the likelihood parameters, we have

$$q(\Theta_c) = q(\boldsymbol{\mu}_c, \mathbf{R}_c) = \mathcal{NW}(\boldsymbol{\mu}_c, \mathbf{R}_c | \tilde{\lambda}_c, \tilde{\mathbf{m}}_c, \tilde{\omega}_c, \tilde{\Psi}_c) \quad (32)$$

where we introduce the notation

$$\tilde{\gamma}_c \triangleq \sum_{n=1}^N q(x_n = c) \quad (33)$$

$$\bar{\mathbf{y}}_c \triangleq \frac{\sum_{n=1}^N q(x_n = c) \mathbf{y}_n}{\tilde{\gamma}_c} \quad (34)$$

$$\boldsymbol{\Delta}_c \triangleq \sum_{n=1}^N q(x_n = c) (\mathbf{y}_n - \bar{\mathbf{y}}_c) (\mathbf{y}_n - \bar{\mathbf{y}}_c)^T \quad (35)$$

and, it holds

$$\tilde{\omega}_c = \omega_c + \tilde{\gamma}_c \quad (36)$$

$$\tilde{\Psi}_c = \Psi_c + \boldsymbol{\Delta}_c + \frac{\lambda_c \tilde{\gamma}_c}{\lambda_c + \tilde{\gamma}_c} (\mathbf{m}_c - \bar{\mathbf{y}}_c) (\mathbf{m}_c - \bar{\mathbf{y}}_c)^T \quad (37)$$

$$\tilde{\lambda}_c = \lambda_c + \tilde{\gamma}_c \quad (38)$$

$$\tilde{\mathbf{m}}_c = \frac{\lambda_c \mathbf{m}_c + \tilde{\gamma}_c \bar{\mathbf{y}}_c}{\tilde{\lambda}_c} \quad (39)$$

Last, the posteriors over the states generating the data yield

$$q(x_n = c) \propto p(x_n = c | \hat{\mathbf{x}}_{\partial_n}; \beta) \tilde{\pi}_c(\mathbf{v}) \tilde{p}(\mathbf{y}_n | \Theta_c) \quad (40)$$

where the spatial priors $p(x_n = c | \hat{\mathbf{x}}_{\partial_n}; \beta)$ are given by (6),

$$\begin{aligned} \tilde{\pi}_c(\mathbf{v}) &\triangleq \exp(\langle \log \pi_c(\mathbf{v}) \rangle) \\ &= \exp \left[\sum_{c'=1}^{c-1} \langle \log(1 - v_{c'}) \rangle + \langle \log v_c \rangle \right] \end{aligned} \quad (41)$$

with

$$\langle \log v_c \rangle = \psi(\beta_{c,1}) - \psi(\beta_{c,1} + \beta_{c,2}) \quad (42)$$

$$\langle \log(1 - v_c) \rangle = \psi(\beta_{c,2}) - \psi(\beta_{c,1} + \beta_{c,2}) \quad (43)$$

and, finally

$$\begin{aligned} \tilde{p}(\mathbf{y}_n | \Theta_c) &\triangleq \exp(\langle \log p(\mathbf{y}_n | \Theta_c) \rangle) \\ &= \exp \left[-\frac{d}{2} \log 2\pi + \frac{1}{2} \langle \log |\mathbf{R}_c| \rangle \right. \\ &\quad \left. - \frac{1}{2} \langle (\mathbf{y}_n - \boldsymbol{\mu}_c)^T \mathbf{R}_c (\mathbf{y}_n - \boldsymbol{\mu}_c) \rangle \right] \end{aligned} \quad (44)$$

where

$$\begin{aligned} \langle (\mathbf{y}_n - \boldsymbol{\mu}_c)^T \mathbf{R}_c (\mathbf{y}_n - \boldsymbol{\mu}_c) \rangle &= \frac{d}{\tilde{\lambda}_c} + \tilde{\omega}_c (\mathbf{y}_n - \tilde{\mathbf{m}}_c)^T \tilde{\Psi}_c^{-1} \\ &\quad \times (\mathbf{y}_n - \tilde{\mathbf{m}}_c) \end{aligned} \quad (45)$$

$$\langle \log |\mathbf{R}_c| \rangle = -\log \left| \frac{\tilde{\Psi}_c}{2} \right| + \sum_{k=1}^d \psi \left(\frac{\tilde{\omega}_c + 1 - k}{2} \right) \quad (46)$$

In the above, $\psi(\cdot)$ denotes the Digamma function, and $\langle \cdot \rangle$ the variational posterior expectation of a quantity. Finally, as a concluding step, at the end of each iteration of the inference algorithm for the iHMRf model, the estimates $\hat{\mathbf{x}}$ of the site labels are updated according to an MPM criterion, i.e., by maximization of $q(x_n = c)$ over c , yielding

$$\hat{x}_n = \operatorname{argmax}_{c=1}^K q(x_n = c) \quad (47)$$

References

- [1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- [2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.
- [3] D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.
- [4] D. Blei and M. Jordan. Variational methods for the Dirichlet process. In *21st Int. Conf. Machine Learning*, 2004.
- [5] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.
- [6] B. Chalmond. An iterative Gibbsian technique for reconstruction of m -ary images. *Pattern Recognition*, 22(6):747–761, 1989.
- [7] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987.
- [8] S. Chatzis, D. Kosmopoulos, and T. Varvarigou. Signal modeling and classification using a robust latent space model based on t distributions. *IEEE Trans. Signal Processing*, 56(3):949–963, March 2008.
- [9] S. P. Chatzis and T. A. Varvarigou. A fuzzy clustering approach toward hidden Markov random field models for enhanced spatially constrained image segmentation. *IEEE Transactions on Fuzzy Systems*, 16(5):1351–1361, October 2008.
- [10] P. Clifford. Markov random fields in statistics. In G. Grimmett and D. Welsh, editors, *Disorder in physical systems. A volume in honour of John M. Hammersley on the occasion of his 70th birthday*. Oxford Science Publication, Clarendon Press, Oxford, 1990.
- [11] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.
- [12] F. Forbes and N. Peyrard. Hidden Markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101, 2003.
- [13] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- [14] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, Dordrecht, 1998.
- [15] P. Kohli, M. Kumar, and P. Torr. p^3 and beyond: Solving energies with higher order cliques. In *Proceedings CVPR*, 2007.
- [16] P. Kohli, L. Ladický, and P. Torr. Robust higher order potentials for enforcing label consistency. In *Proceedings CVPR, 2008*. [Online]. Available: http://research.microsoft.com/en-us/um/people/pkohli/papers/klt_CVPR08_slides.pdf.
- [17] K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational Dirichlet process mixtures. In *Advances in Neural Information Processing Systems*, 2006.
- [18] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order Markov random fields. In *Proceedings ECCV*, pages 269–282, 2006.
- [19] J. Maroquin, S. Mitte, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of The American Statistical Association*, 82:76–89, 1987.
- [20] B. Potetz. Efficient belief propagation for vision using linear constraint nodes. In *Proceedings CVPR*, pages 1–8, 2007.
- [21] Y. Qi, J. W. Paisley, and L. Carin. Music analysis using hidden Markov mixture models. *IEEE Transactions on Signal Processing*, 55(11):5209–5224, 2007.
- [22] W. Qian and D. Titterton. Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London A*, 337:407–428, 1991.
- [23] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proceedings CVPR*, pages 860–867, 2005.
- [24] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 4(461-464), 1978.
- [25] J. Sethuraman. A constructive definition of the Dirichlet prior. *Statistica Sinica*, 2:639–650, 1994.
- [26] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In 1-15, editor, *Proceedings ECCV*, 2006.
- [27] D. Stanford and A. E. Raftery. Determining the number of colors or gray levels in an image using approximate bayes factors: The pseudolikelihood information criterion (PLIC). Technical report, Dept. of Statistics, Univ. of Washington, 2001.
- [28] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *Proceedings CVPR*, 2007.
- [29] J. Zhang. The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Image Processing*, 2(1):27–40, 1993.