Smooth Image Segmentation by Nonparametric Bayesian Inference

Peter Orbanz and Joachim M. Buhmann

Institute of Computational Science, ETH Zurich
{porbanz, jbuhmann}@inf.ethz.ch

Abstract. A nonparametric Bayesian model for histogram clustering is proposed to automatically determine the number of segments when Markov Random Field constraints enforce smooth class assignments. The nonparametric nature of this model is implemented by a Dirichlet process prior to control the number of clusters. The resulting posterior can be sampled by a modification of a conjugate-case sampling algorithm for Dirichlet process mixture models. This sampling procedure estimates segmentations as efficiently as clustering procedures in the strictly conjugate case. The sampling algorithm can process both single-channel and multi-channel image data. Experimental results are presented for real-world synthetic aperture radar and magnetic resonance imaging data.

1 Introduction

Unsupervised data clustering and image segmentation models usually assume that an appropriate number of classes is either known a priori or specified by the data analyst. More sophisticated methods automatically select the number of clusters, e. g. by resampling strategies [1]. Recently, nonparametric Bayesian models based on Dirichlet processes have successfully been applied to machine learning problems such as natural language processing [2] and object categorization [3]. These models perform automatic model selection by supporting a range of prior choices for the number of classes; the different resulting models are then scored by the likelihood according to the observed data.

The question how automatic model selection can be performed in image segmentation for models such as Markov random fields plays an important role in computer vision; see e. g. [4] for recent work employing a Bayesian information criterion. Our approach, which is based on Dirichlet processes, combines spatial constraints on class labels with an estimate of a preferred number of clusters. The smoothness constraints are modeled as a Markov random field (MRF) on a neighborhood graph. To combine MRF image models for segmentation with a nonparametric selection of the segment number, the Dirichlet process prior is enhanced by a smoothness constraint on the label field. Local feature histograms are extracted from the image and grouped by histogram clustering. Adjacent image patches are assigned to the same cluster with high probability if they are neighbors with respect to the neighborhood graph of the MRF.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part I, LNCS 3951, pp. 444–457, 2006.

The paper is organized as follows: Sec. 2 briefly reviews Dirichlet process mixture (MDP) models and their application to data clustering. We discuss their combination with MRFs in Sec. 3, and the histogram clustering model used for application to image segmentation in Sec. 4. Sec. 5 proposes a MCMC sampling algorithm to sample the combined model. Experimental results are given in Sec. 6.

2 Data Clustering with MDP Models

The statistical model considered in this work is a *Dirichlet process mixture* (MDP) model [5]. MDP approaches belong to a class of models referred to as *nonparametric Bayesian models*. A MDP clustering model consists of three principal ingredients: A parametric likelihood function F, a probability distribution G_0, which is referred to as the *base measure*, and a Dirichlet process $\mathrm{DP}(\alpha G_0)$ parameterized by the base measure and a positive constant $\alpha \in \mathbb{R}_+$. In this article, the base measure G_0 will generally be assumed to be infinite. Under the MDP model, a set of distinct classes is assumed to generate the observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Each class has a generative distribution, described by the likelihood F. Each cluster (indexed by k) is characterized by a parameter value θ_k^*, so the data within the cluster is generated according to $\mathbf{x} \sim F(\,.\,|\theta_k^*)$. This makes MDP models conceptually similar to finite parametric mixture models. MDP models generate the parameter values θ_k^*, which characterize the classes, according to a Dirichlet process $\mathrm{DP}(\alpha G_0)$. In contrast to parametric mixture models, the number of classes is not a constant, and will change during the sampling process.

Formally, models based on Dirichlet processes draw a distribution G at random from a stochastic process [5]. The sample values drawn by means of the DP, the mixture parameters $\theta_1, \ldots, \theta_n$, are assumed to be generated by the distribution G:

$$\theta_1, \ldots, \theta_n \sim G \quad \text{with} \quad G \sim \mathrm{DP}(\alpha G_0) \ . \tag{1}$$

The practical applicability of the process, however, is based on the observation that the distribution G can be integrated out. Given a set of samples $\theta_1, \ldots, \theta_n$, a new sample θ_{n+1} has a closed-form conditional distribution:

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta_{\theta_i}(\theta_{n+1}) + \frac{\alpha}{n+\alpha} G_0(\theta_{n+1}) \ , \tag{2}$$

where δ_θ denotes the Dirac measure concentrated at θ. Therefore, sampling the Dirichlet process generates random values in the domain of the base measure G_0, but with a different distribution than the one specified by G_0.

A draw from the distribution (2) will, with probability $\frac{n}{n+\alpha}$, yield a sample value which has already occurred. (Provided that G_0 is infinite, a draw from the second term in (2) will generate a previously unobserved value with probability one.) If any two samples θ_i, θ_j are identical, the corresponding Dirac measures coincide. One may therefore group the samples $\theta_1, \ldots, \theta_n$ into $N_\mathrm{C} \leq n$ classes

containing identical values. Each class $k \in \{1, \ldots, N_C\}$ is characterized by its associated sample value, denoted θ_k^*. Denoting the number of samples in group k by n_k, the distribution (2) may be rewritten as a sum over clusters rather than individual samples:

$$p_{n+1}(\theta_{n+1}|\theta_1, \ldots, \theta_n) := \sum_{k=1}^{N_C} \frac{n_k}{n+\alpha} \delta_{\theta_k^*}(\theta_{n+1}) + \frac{\alpha}{n+\alpha} G_0(\theta_{n+1}). \tag{3}$$

The distribution may be regarded as a mixture model. It contains N_C finite (degenerate) components, which correspond to the clusters already created, and the base measure component, which is responsible for the creation of new classes. The probability of occurrence for each cluster is proportional to its size. The probability for a new class to be created is adjusted by means of the DP parameter α. Definition (3) also implies that $\mathrm{DP}(\alpha G_0)$ can be sampled efficiently, if we provide an algorithm to sample the base measure G_0.

Data generation (of n data values $\mathbf{x}_1, \ldots, \mathbf{x}_n$) according to a MDP model can be summarized by

$$\begin{aligned} \mathbf{x}_i &\sim F(\,.\,|\theta_i) \\ \theta_i &\sim p_i(\theta_i|\theta_1, \ldots, \theta_{i-1})\ . \end{aligned} \tag{4}$$

Inference of this model is not as straightforward as sampling (3), since the observed data is $\mathbf{x}_1, \ldots, \mathbf{x}_n$, whereas the DP distribution is conditional on $\theta_1, \ldots, \theta_n$. The generative model in (4) has to be sampled conditional on the observed data $\mathbf{x}_i$. A sampling algorithm as described in Sec. 5 obtains estimates of the parameters θ_i. MDP models perform automatic model selection, since the number of clusters is determined by the dynamics of the process, i.e., it is not an input parameter. New classes are generated during the sampling process. When sampling the parameter θ_i for a given data value $\mathbf{x}_i$, the data value may be assigned to an existing class k (by setting $\theta_i := \theta_k^*$). The probability for this to happen depends on the likelihood $F(\mathbf{x}_i|\theta_k^*)$ and on the number of points already assigned to the class in question (since large classes, with a large value of n_k, are more probable than small ones). If the cluster distribution provides a good description of the data, the probability of assignment is high, since the likelihood F assumes a large value. If this is not the case for any existing cluster, a new cluster is created for the data value with high probability. Generation of a new cluster corresponds to sampling from the base measure term in (3).

The applicability of MDP models to clustering problems in machine learning and computer vision may be best illustrated by the following observation: Any parametric mixture model of the form

$$m(\mathbf{x}|\mathbf{t}_1, \ldots, \mathbf{t}_K, c_1, \ldots, c_K) = \sum_{k=1}^{K} c_k r(\mathbf{x}|\mathbf{t}_k) \tag{5}$$

can be used within the MDP clustering framework by setting $F = r$ and placing a suitable prior G_0 on the parameter t_k. The prior serves as the base measure.

The class parameters t_k are substituted by samples θ_k^* generated by a process $\mathrm{DP}(\alpha G_0)$ as described above, and the cluster sizes n_k are analogous to the mixture weights c_k in the parametric case. Given a set of observed data values $\mathbf{x}_1, \ldots, \mathbf{x}_n$, sampling the MDP model will result in a set of estimates $\theta_1, \ldots, \theta_n$ for the corresponding class parameters. By grouping identical values, the parameter estimates implicitly determine the number of clusters, class assignments of the data and the mixture proportions of the model.

3 Markov Random Field Constraints and Dirichlet Process Models

This section describes how Markov random field models can be integrated with a MDP clustering approach. Our objective is to obtain a model capable of combining the clustering and model selection performed by the MDP with smoothness constraints on the class labels. The model is applicable to any clustering problem for which it is reasonable to assume a spatially coherent class structure, such as segmentation of noisy images: To obtain smooth segments, a MRF constraint encourages adjacent points in the image to be assigned to the same class.

Consider a clustering problem with vectorial input data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Each point $\mathbf{x}_i$ is assumed to be generated according to a parameter vector θ_i. Two points are considered to originate from the same cluster if their respective parameter vectors are identical. The cluster assignment of feature $\mathbf{x}_i$ is denoted by $S_i \in \{1, \ldots, N_\mathrm{C}\}$. We will use the notation θ_{-i} (or S_{-i}) to denote the set of all parameters (or cluster assignments) with the value corresponding to feature i removed. The MDP clustering model for this problem is once again defined by a likelihood F and a base measure G_0 to parameterize the Dirichlet process.

To combine the MDP model with a MRF, we restrict the choice of MRF constraints to *pairwise difference priors* [6], which are commonly used to model spatial smoothness of the label field. The MRF definition is based on an undirected neighborhood graph $\mathcal{N}$ and we write $l \in \partial(i)$ to denote that the feature of index l is a neighbor of feature i. The MRF prior $\mathit{\Pi}$ consists of two components,

$$\mathit{\Pi}(\theta) \propto P(\theta) M(\theta) \ . \tag{6}$$

P is a parametric prior on the parameter θ, which will be referred to as the *initial prior*. It is used to model initial beliefs about which parameter values are likely to occur. M is a MRF contribution term of the form $M(\theta_i) \propto \exp(-H(\theta_1, \ldots, \theta_n))$, H being a cost function defined on the neighborhood graph $\mathcal{N}$. The term M is used to model constraints such as smoothness, which are conditional on the neighborhood of a feature. M defines a pairwise difference prior if the cost function assumes the form $H(\theta_i|\theta_{-i}) = \sum_{l \in \partial(i)} w_{il} \Phi(\theta_i - \theta_l)$, where Φ is a non-negative, even function and w_{il} are weights associated with the edges of the graph. Conditional on θ_{-i}, the prior for θ_i is given by

$$\mathit{\Pi}(\theta_i|\theta_{-i}) \propto P(\theta_i|\theta_{-i}) \exp(-H(\theta_i|\theta_{-i})) \ . \tag{7}$$

In the above relations, normalization constants have been neglected, because in many practical cases, M will be improper.

The MDP approach and MRF constraints are combined by drawing the initial prior P in (6) from a DP. The resulting generative model is summarized by

$$\begin{aligned} \mathbf{x}_i &\sim F(\mathbf{x}_i|\theta_i) \\ \theta_i &\sim M(\theta_i|\theta_{-i})P(\theta_i) \\ P &\sim \mathrm{DP}\left(\alpha G_0\right) \ . \end{aligned} \tag{8}$$

To obtain a conditional form of this model, i. e. a form in which the random measure P does not occur explicitly, the conditional MDP prior (3) is substituted into (7). For a fixed size data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the sequential form (3) of the conditional prior is rewritten as a prior for θ_i given the remaining parameter values:

$$p_n(\theta_i|\theta_{-i}) \propto \sum_{k=1}^{N_\mathrm{C}} n_{-i}^k \delta_{\theta_k^*}(\theta_i) + \alpha G_0(\theta_i) \ , \tag{9}$$

where n_{-i}^k denotes the number of observations assigned to cluster k when $\mathbf{x}_i$ is removed from the set. The conditional form of the combined MDP/MRF prior is then given by

$$\mathit{\Pi}\left(\theta_i|\theta_{-i}\right) \propto p_n(\theta_i|\theta_{-i}) \exp\left(-H\left(\theta_i|\theta_{-i}\right)\right) \ . \tag{10}$$

Smoothness constraints for clustering problems are formulated on the cluster assignments, so the MRF cost function is a function defined on labels. A cost function modeling spatial smoothness measures whether or not neighboring features are assigned to the same cluster. This binary notion of similarity between neighbors is expressed by cost functions of the general form

$$H\left(S_i|S_{-i}\right) = \sum_{l \in \partial(i)} \delta_{S_i,S_l} \phi(S_1, \ldots, S_n) \ , \tag{11}$$

as proposed by Geman e. a. [7]. A special property of the MDP setting is the one-to-one correspondence between cluster labels and cluster parameters (since two sites belong to the same cluster if and only if their class parameters θ are identical). The correspondence admits an equivalent formulation of the cost function (11) in terms of class parameters:

$$H\left(\theta_i|\theta_{-i}\right) = \sum_{l \in \partial(i)} \delta_{\theta_i,\theta_l} \phi(\theta_1, \ldots, \theta_n) \ . \tag{12}$$

Combination of the resulting MRF with the conditional MDP prior (9) affects only the first, finite term, because the support of H is a subset $\{\theta_1, \ldots, \theta_n\}$. A random value $\theta \sim G_0$ drawn from an infinite base measure will be different from any value in $\mathrm{supp}\left(H\right)$ with probability one, and therefore

$$M(\theta_i|\theta_{-i})G_0(\theta_i) = G_0(\theta_i) \tag{13}$$

almost surely. The relation holds irrespectively of any particular choice of G_0 and H. Intuitively, (13) expresses the modeling assumption that the MRF constraint should encourage uniform assignments of neighbors. The MRF contribution is nontrivial for a given label S_i only if one or more neighbors of $\mathbf{x}_i$ are assigned to the same class as $\mathbf{x}_i$. Since a draw from the base measure will always result in the creation of a new class, the MRF term does not affect the base measure term.

4 The Histogram Clustering Model

The primary focus of this article is on histogram clustering, with application to image segmentation. The input features are composed of a set of n histograms $\mathbf{h}_i = (h_{i1}, \ldots, h_{iN_{\text{bins}}})$, $h_{ij} \in \mathbb{N}_0$, representing local intensity distributions of a digital image. They replace the data values $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the previous sections. Each histogram is associated with a pixel location in the image, referred to as a *site*. All histograms contain an identical number N_{counts} of values.

Given a vector θ_i of bin probabilities, a random histogram $\mathbf{h}_i$ is multinomially distributed with density $F(\mathbf{h}_i|\theta_i) = 1/Z_M(\mathbf{h}_i) \exp\left(\sum_{j=1}^{N_{\text{bins}}} h_{ij} \log(\theta_{ij})\right)$. Each vector θ_i for is assumed to be drawn from the respective conjugate prior, a Dirichlet distribution $G_0(\theta_i|\beta, \boldsymbol{\pi}) = \frac{1}{Z_D(\beta,\boldsymbol{\pi})} \exp\left(\sum_{j=1}^{N_{\text{bins}}} (\beta\pi_j - 1) \log(\theta_{ij})\right)$, where $\beta \in \mathbb{R}_+$ and $\boldsymbol{\pi}$ is a vector representing a finite probability distribution on N_{bins} elements.

To apply MRF constraints to the image segmentation problem, two features are defined as neighbors in the MRF neighborhood graph $\mathcal{N}$ if their associated sites are neighbors in the image. These neighborhoods are either of size $D = 4$ (two horizontal and two vertical neighbors) or $D = 8$ (all direct neighbors), cropped at the image boundaries. The cost function is of the form (12). For the sake of simplicity, ϕ in (12) is chosen to depend only on a scale parameter λ (defined once for the whole image) and the size of the neighborhood:

$$H\left(\theta_i|\theta_{-i}\right) = \lambda \sum_{l \in \partial(i)} \left(D - \delta_{\theta_i,\theta_l}\right) \ , \tag{14}$$

where $D = 4$ or $D = 8$, respectively. Thus, $\exp(-H) = \exp(-\lambda D)$ for feature $\mathbf{h}_i$ if no neighbor is assigned to the same cluster. If one or more neighbors are assigned to the same class, $\exp(-H)$ will increase and thus favor the assignment.

The model may be extended to the case of multiple histograms available at each site. This extension makes the method applicable to color images, where a single one-dimensional histogram is drawn from each color channel at each site, and to radar images with multiple channels representing different frequency bands. Another possible application is the inclusion of additional filter information, by applying a filter transform to the image and drawing histograms from the filter response. For example, texture information may be included in the form of Gabor filter response histograms. Suppose that C histograms $\mathbf{h}_i^l = (h_{i1}^l, \ldots, h_{iN_{\text{bins}}}^l)$, $l = 1, \ldots, C$, are available at each site i. First consider the basic parametric Bayesian model without the DP, consisting of the

multinomial likelihood and Dirichlet prior in the single-channel case. To model multiple channels, the different channels are assumed to be independent. Each marginal histogram $\mathbf{h}_i^l$ is parameterized by its own vector θ_i^l of bin probabilities, and we write $\theta_i := (\theta_i^1, \ldots, \theta_i^C)$. Due to independence, the joint likelihood $F(\mathbf{h}_i^1, \ldots, \mathbf{h}_i^C | \theta_i^1, \ldots, \theta_i^C)$ factors into a product over the channel likelihoods $F(\mathbf{h}_i^l | \theta_i^l)$. Each parameter vector θ_i^l is drawn from a Dirichlet distribution $G_0^l(\theta | \beta^l, \boldsymbol{\pi}^l)$, resulting in the model

$$F(\mathbf{h}|\theta)G_0(\theta) = \prod_{l=1}^{C} F(\mathbf{h}_i^l|\theta_i^l)G_0^l(\theta|\beta^l, \boldsymbol{\pi}^l) \; . \tag{15}$$

The MDP/MRF generative model for multichannel data is then obtained by substituting $F(\mathbf{h}|\theta)$ and $G_0(\theta)$ into the generative model (8).

5 Sampling

The algorithm proposed here to sample the combined MDP/MRF model is a Markov chain Monte Carlo procedure similar to the algorithm proposed MacEachern [8] for sampling MDP models with a conjugate likelihood/base measure pair. Each iteration samples a set of cluster assignments $S_1, \ldots, S_n$ for all sites. New estimates of the cluster parameters θ_k^* are then sampled conditional on the assignments S_i and the observed data. Due to the way in which the finitely supported cost function of the MRF acts on the MDP model, some key formulas reduce to the conjugate case. As a consequence, the sampling approach remains applicable despite the fact that the constrained model is not conjugate. It is easily extended to the case of multiple channels.

To sample a cluster assignment S_i given a current set of parameters $\theta_1, .., \theta_n$ and the datum $\mathbf{x}_i$, the posterior probability of occurrence for each class is computed by integrating the complete model over θ_i:

$$\begin{aligned} &\int_{\Omega_\theta} \exp\left(-H(\theta_i|\theta_{-i})\right) F(\mathbf{x}_i|\theta_i) \left(\sum_{k=1}^{N_\mathrm{C}} n_{-i}^k \delta_{\theta_k^*}(\theta_i) + \alpha G_0(\theta_i) \right) d\theta_i \\ &= \sum_{k=1}^{N_\mathrm{C}} n_{-i}^k \exp\left(-H(\theta_k^*|\theta_{-i})\right) F(\mathbf{x}_i|\theta_k^*) + \alpha \int_{\Omega_\theta} F(\mathbf{x}_i|\theta_i) G_0(\theta_i) d\theta_i \; . \end{aligned} \tag{16}$$

Since $H(\theta|\theta_{-i}) \neq 0$ only if $\theta \in \{\theta_1^*, \ldots, \theta_{N_\mathrm{C}}^*\}$, $\exp\left(-H(\theta|\theta_{-i})\right) \neq 1$ holds only on a set of Lebesgue measure zero. Such a set does not affect the value of the integral, and the MRF contribution term may therefore be neglected in the base measure integral, as we have done above. Each term in (16) corresponds to a single cluster (with the integral involving the base measure G_0 corresponding to the creation of a new group), and we define cluster proportions by setting

$$\begin{aligned} \tilde{q}_{i0} &:= \alpha \int_{\Omega_\theta} F(\mathbf{x}_i|\theta_i) G_0(\theta_i) d\theta_i \\ \tilde{q}_{ik} &:= n_{-i}^k \exp\left(-H(\theta_k^*|\theta_{-i})\right) F(\mathbf{x}_i|\theta_k^*) \; . \end{aligned} \tag{17}$$

These proportions are transformed into cluster probabilities by normalization,

$$q_{ik} := \frac{\tilde{q}_{ik}}{\sum_{j=0}^{N_C} \tilde{q}_{ij}} \ . \tag{18}$$

A cluster assignment S_i is sampled by sampling from the finite probability distribution defined by the vector $(q_{i0}, \ldots, q_{iN_C})$. In the second step, new values for the cluster parameters θ_k^* are chosen by sampling from the class posterior, i. e. the posterior based on all data values currently assigned to the given class:

$$\theta_k^* \sim G_0\left(\theta_k^*\right) \prod_{i|S_i=k} F\left(\mathbf{x}_i|\theta_k^*\right) \ . \tag{19}$$

The combined MDP/MRF model is thus sampled by the following algorithm:

Algorithm 1 (MDP/MRF Sampling)
Initialize: Generate $\theta \sim G_0$ and set $\theta_i = \theta$ for $i = 1, \ldots, n$.
Repeat:

1. For $i = 1, \ldots, n$:
   (a) If $\mathbf{x}_i$ is the only feature assigned to its cluster $k = S_i$, remove this cluster.
   (b) For $k = 0, \ldots, N_C$, compute the component probabilities $q_{i,k}$ according to eqs. (17) and (18).
   (c) Draw a random index k according to the finite distribution $(q_{i,0}, \ldots, q_{i,N_C})$.
   (d) Assignment:
   - If $k \in \{1, \ldots, N_C\}$, assign $\mathbf{x}_i$ to cluster k.
   - If $k = 0$, create a new cluster for $\mathbf{x}_i$.
2. For each cluster $k = 1, \ldots, N_C$: Update the cluster parameters θ_k^* given the class assignments $S_1, \ldots, S_n$ by sampling

$$\theta_k^* \sim G_0\left(\theta_k^*\right) \prod_{i|S_i=k} F\left(\mathbf{x}_i|\theta_k^*\right) \ . \tag{20}$$

In the histogram clustering model introduced above for the single-channel case, F is a multinomial distribution, G_0 a Dirichlet distribution and the observed data $\mathbf{x}_i$ are the histograms $\mathbf{h}_i$. Due to the conjugacy of F and G_0, the integral required for the computation of q_{i0} may be solved analytically:

$$\tilde{q}_{i0} = \alpha \int_{\Omega_\theta} F(\mathbf{x}_i|\theta_i) G_0(\theta_i) d\theta_i = \alpha \frac{Z_D(\mathbf{h}_i + \beta\pi)}{Z_D(\beta\boldsymbol{\pi}) Z_M(\mathbf{h}_i)} \ . \tag{21}$$

Conjugacy also implies that the class posterior (20) is a Dirichlet distribution, with the prior parameters updated by the data assigned to the cluster:

$$G_0\left(\theta_k^*|\beta\boldsymbol{\pi}\right) \prod_{i|S_i=k} F\left(\mathbf{x}_i|\theta_k^*\right) = G_0\left(\theta_k^* \left| \sum_{i|S_i=k} \mathbf{h}_i + \beta\boldsymbol{\pi} \right.\right) \ . \tag{22}$$

Efficient sampling algorithms based on gamma samples are available for this distribution [9], which ensures the feasibility of step 2 of the algorithm.

In the case of multiple channels, products of multinomial and Dirichlet distributions have to be substituted for F and G_0 in the derivation above, assuming that the different channels are statistically independent. Since the MRF term is defined on class labels, it applies to all channels, rather than to each individual channel. The cluster proportions are computed according to

$$\tilde{q}_{i0} := \alpha \int_{\Omega_\theta} \prod_{l=1}^{C} \left(F(\mathbf{h}_i^l|\theta_i^l) G_0^l(\theta_i^l|\beta^l, \boldsymbol{\pi}^l) \right) d\theta_i = \prod_{l=1}^{C} \frac{Z_D(\mathbf{h}_i^l + \beta^l \pi^l)}{Z_D(\beta^l \boldsymbol{\pi}^l) Z_M(\mathbf{h}_i^l)}$$

$$\tilde{q}_{ik} := n_{-i}^k \exp\left(-H(\theta_k^*|\theta_{-i})\right) \prod_{l=1}^{C} F(\mathbf{x}_i^l|\theta_k^{*l}) \; . \tag{23}$$

The class posterior turns into a product of Dirichlet distributions, each of which may be sampled individually:

$$\prod_{l=1}^{C} \left(G_0^l\left(\theta_k^{*l}|\beta^l \boldsymbol{\pi}^l\right) \prod_{i|S_i=k} F\left(\mathbf{h}_i^l|\theta_k^{*l}\right) \right) = \prod_{l=1}^{C} G_0^l \left(\theta_k^{*l} \middle| \sum_{i|S_i=k} \mathbf{h}_i^l + \beta^l \boldsymbol{\pi}^l \right) \; . \tag{24}$$

Sampling of the Dirichlet process for the multichannel model is thus conducted by parallel Dirichlet process sampling procedures applied to the individual channels. The channels couple through the class assignments S_i, and through the MRF contribution defined on these labels.

6 Experimental Results

The experiments presented in this section were conducted on two classes of noisy images, synthetic aperture radar (SAR) and magnetic resonance imaging (MRI) data. Aside from the visual quality of the segmentations, we especially study two model selection questions: (i) How does the hyperparameter of the Dirichlet process influence the model selection (i. e. the number of segments selected)? (ii) How do results compare to other model selection methods?

The histograms used in the experiments shown here where extracted from a digital image by centering a square window around each pixel on an equidistant grid and sorting the intensity values of all pixels within the window into a histogram. Choosing the size of the histogram window generally results in a trade-off between regularity and detail: Using a large window will smooth segmentation results, but coarsen the resolution. Small windows preserve detail, but usually give less robust segmentation results. Using a model with a smoothness constraint permits the choice of small windows. For the experiments shown below, histograms were obtained from a five-by-five pixel sliding window, centered at each node of a rectangular grid of width two.

The nonparametric Bayesian model selection strategy introduced in the previous sections is compared with the stability method [10, 1], a competitive model

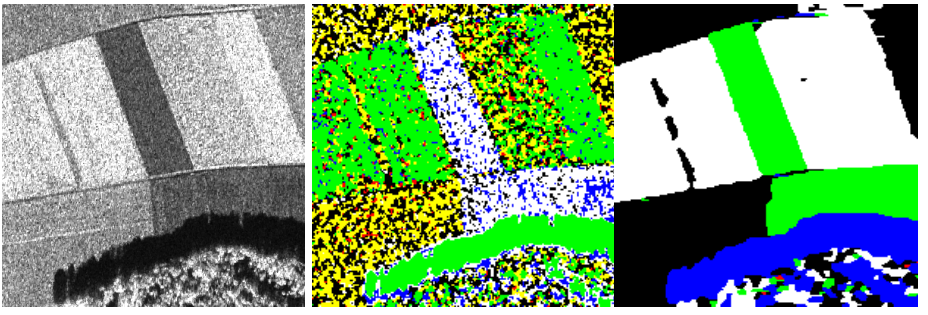

Fig. 1. Segmentation results on real-world radar data. Original image (left), unconstrained MDP segmentation (middle), MDP segmentation with smoothness constraint (right).

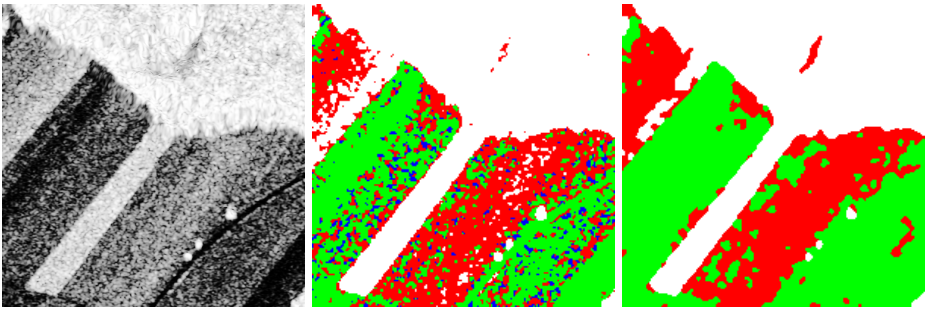

Fig. 2. A SAR image with a high noise level and ambiguous segments (left). Solutions without (middle) and with smoothing (right).

selection technique for clustering. Stability is a cross-validation based wrapper method for an arbitrary clustering algorithm chosen by the user. The method repeatedly computes clustering solutions on randomly chosen subsets of the input data, and evaluates the predictive power of the obtained cluster model on the remaining data. An instability index is computed for different number of clusters, which measures how unstable cluster solutions are under the random split procedure. The chosen model is the one for which the instability index is minimal. Usually, a local rather than the global minimum is chosen, since stability algorithms are known to preferentially estimate a global minimum for very simple solutions (often only two classes). Consider, for example, intensity-based image segmentation: A two-class segmentation, which simply splits the image into light and dark regions, tends to be highly stable with respect to the random split procedure, but is usually not the desired solution.

The MDP/MRF method applied for image segmentation employs a multinomial likelihood. To obtain a valid comparison, the algorithm chosen for use with stability is an EM algorithm which estimates a mixture of multinomial

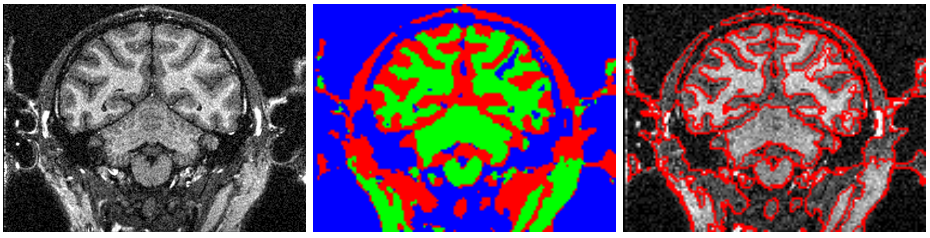

Fig. 3. MR frontal view image of a monkey's head. Original image (left), smoothed MDP segmentation (middle), original image overlaid with segment boundaries (right).

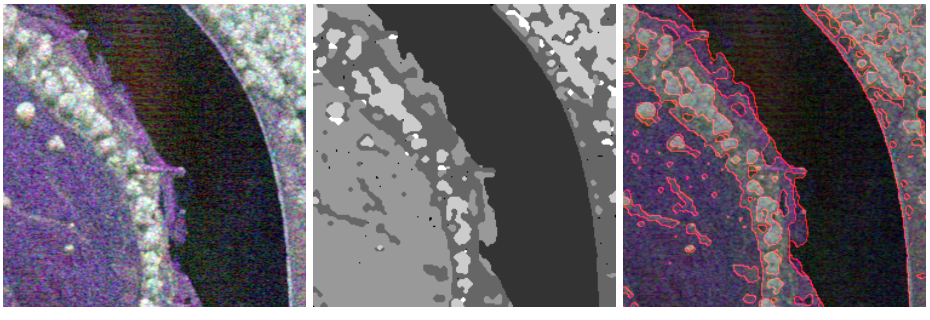

Fig. 4. Segmentation result for multichannel data: A SAR image with three channels (left), segmentation result obtained with the MDP/MRF model, and the original image overlaid with segment boundaries (right)

distributions (also known as the ACM algorithm [11]). Figs. 1 and 2 show results for two SAR images. Segments of the image in Fig. 1 are well separated. As the results show, segmentation quality for noisy data can be improved significantly by a smoothness constraint. Fig. 2 provides an example of ambiguous, poorly separated segments. In this case, both the unconstrained and constrained segmentation results are of limited quality. Another type of noisy data, a MR image, is shown in Fig. 3 together with its (constrained) segmentation result. Fig. 4 shows segmentation results obtained with the multichannel version of the algorithm on a SAR image consisting of three separate frequency bands.

The burn-in phase of the Gibbs sampling algorithm is assumed to have terminated once the number of assignments changed per iteration remains stable below 1% of the total number of sites. This condition is usually met after at most 500-1000 iterations. The behavior of the class assignments during the sampling process visualized by the plot in Fig. 5. In both cases, the algorithm takes about 600 iterations to stabilize (the curves become constant apart from fluctuations). The splitting behavior of the algorithm differs significantly between the two cases: In the unconstrained case, large batches of sites are reassigned at

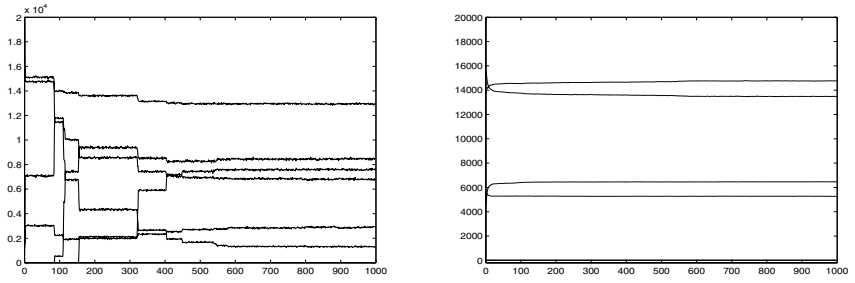

Fig. 5. Cluster sizes during the sampling process for the unconstrained and smoothed version of the MDP method. The number of sites assigned to each cluster (vertical) are drawn against the number of iterations (horizontal), with each graph representing a cluster. Left: Radar image (Fig. 1), no smoothing. Right: Same image, with smoothing.

Table 1. Number of clusters chosen by the algorithm on two radar images for different values of the hyperparameter

α		1e-10	1e-9	1e-8	1e-7	1e-6	1e-5	1e-4	1e-3
Image Fig. 1	MDP	2	4	4	6	5	4	5	6
	smoothed	2	2	3	4	4	4	4	4
Images Fig. 2	MDP	4	3	4	7	6	5	5	9
	smoothed	2	2	3	4	5	3	3	5

once to new clusters (visible as jumps in the diagram). In the constrained case, assignments change gradually.

The influence of the DP hyperparameter α is shown in Tab. 1. In general, the number of clusters increases for larger values of α (i. e. when the probability is high that a new cluster is created by the DP). When the smoothing constraint is activated, the number of clusters becomes more stable with respect to changes of α than without smoothing. We note that the number of clusters selected is more volatile for the poorly separated image in Fig. 2.

For comparison of the model selection results, the stability method has been applied to the two SAR images in Figs. 1 and 2. The resulting instability indices for two to nine clusters are given in Tab. 2. For the image in Fig. 1, the local minimum of the instability index is assumed for five clusters, with the solutions $N_C = 3, 4, 5$ within range of the error bars. This outcome is comparable to the result of the smoothed MDP model, which (except for very small values of α) selects three or four clusters. The unconstrained MDP model tends to select a larger number of clusters. Since the instability index is obtained by averaging over results on random subsets, one should expect its results to be conservative. This is indeed the case, since the smoothed MDP approach produces a comparable number of segments as the stability method does without smoothing. Now consider the image in Fig. 2, for which MDP results, even in the smoothed case, are rather unstable (cf. Tab. 1). The local minimum

Table 2. Stability indices computed with ACM clustering on two radar images for different numbers of clusters

	Stability index			Stability index	
N_C	Image Fig. 1	Image Fig. 2	N_C	Image Fig. 1	Image Fig. 2
2	0.0012 ± 0.0009	0.0003 ± 0.3341	6	0.4740 ± 0.0867	0.2933 ± 0.3437
3	0.3359 ± 0.2324	0.1765 ± 0.2856	7	0.5164 ± 0.0434	0.2907 ± 0.3007
4	0.3204 ± 0.2113	0.1233 ± 0.3481	8	0.5598 ± 0.0728	0.3532 ± 0.2889
5	0.2947 ± 0.0884	0.1436 ± 0.1929	9	0.6637 ± 0.0512	0.3378 ± 0.2801

of the instability index is assumed at $N_C = 4$, but the whole range of computed solutions ($N_C = 2, \ldots, 9$) is within one standard deviation of the local minimum. Thus both the MDP/MRF approach and the stability method give unreliable results on an image with a high noise level and poorly discernible segments. Both methods are constructed around the same probabilistic model of the data (a multinomial histogram clustering model). We therefore conclude that the reliability of model selection results depends, for both approaches, on the ability of the clustering model to resolve differences between segment distributions.

7 Discussion

There exists a considerable number of DP-based models [12] with a wide range of applications in statistics and, more recently, natural language processing and document retrieval [13, 2]. To our knowledge, this paper summarizes the first attempt both to apply the Dirichlet nonparametric approach to image segmentation, and to combine it with Markov random fields, the standard Bayesian approach to image processing and spatial statistics.

We believe that a wide range of applications for MDP models may emerge in computer vision. Despite their mathematical intricacies, the fact that these models may be regarded as mixture distributions with a variable number of mixture components (cf. Sec. 2) makes them an intuitive and powerful tool for probabilistic modeling. Instead of the multinomial distribution employed in our histogram clustering approach, any type of parametric likelihood may be used with the MDP model. If the base measure is set to the respective conjugate prior, standard sampling algorithms are applicable. For example, a nonparametric analogue of the widely used k-means algorithm may be obtained by choosing a Gaussian of fixed, uniform covariance as the likelihood and a Gaussian prior on the mean parameter as the base measure. For applications requiring fast inference, sampling algorithms may be substituted by more efficient approximate methods [14].

We have shown how to combine the MDP clustering model with a spatial smoothness constraint. We like to emphasize that this nonparametric framework is applicable to any type of mixture component distribution and our sampling algorithm remains applicable for any conjugate likelihood/base measure pair.

Our experiments confirm what the structure of the model suggests: The ability of the parametric model used within the nonparametric framework to resolve differences between segments determines the quality of segmentation results. It also determines how stable model selection results are with respect to changes of the hyperparameter.

In summary, the MDP approach can be regarded as a model selection framework built in the style of a wrapper method around an application dependent parametric model. Additionally, it may be equipped with a smoothness constraint for image segmentation. The comparison with the stability framework based on cross-validation yields consistent results for the number of clusters.

References

1. Lange, T., Roth, V., Braun, M., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Computation **16** (2004) 1299–1323
2. Zaragoza, H., Hiemstra, D., Tipping, D., Robertson, S.: Bayesian extension to the language model for ad hoc information retrieval. In: Proc. SIGIR 2003. (2003)
3. Sudderth, E., Torralba, A., Freeman, W.T., Willsky, A.S.: Describing visual scenes using transformed dirichlet processes. In Weiss, Y., Schölkopf, B., Platt, J., eds.: Advances in Neural Information Processing Systems 18, MIT Press (2006)
4. Murtagh, F., Raftery, A.E., Starck, J.L.: Bayesian inference for multiband image segmentation via model-based cluster trees. Image and Vision Computing **23** (2005) 587–596
5. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to bayesian nonparametric estimation. Annals of Statistics **2** (1974) 1152–1174
6. Besag, J., Green, P., Higdon, D., Mengersen, K.: Bayesian computation and stochastic systems. Statistical Science **10** (1995) 3–66
7. Geman, D., Geman, S., Graffigne, C., Dong, P.: Boundary detection by constrained optimization. IEEE Trans. on Pat. Anal. Mach. Intel. **12** (1990) 609–628
8. MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Communications in Statistics: Simulation and Computation **23** (1994) 727–741
9. Devroye, L.: Non-uniform random variate generation. Springer (1986)
10. Breckenridge, J.: Replicating cluster analysis: Method, consistency and validity. Multivariate Behavioral Research **24** (1989) 147–161
11. Puzicha, J., Hofmann, T., Buhmann, J.M.: Histogram clustering for unsupervised segmentation and image retrieval. Pattern Recognition Letters **20** (1999) 899–909
12. MacEachern, S., Müller, P.: Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In Ruggeri, F., Rios Insua, D., eds.: Robust Bayesian Analysis. Springer (2000)
13. Blei, D.M., Griffith, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16, MIT Press (2004)
14. Blei, D.M., Jordan, M.I.: Variational methods for the Dirichlet process. In: Proceedings of the 21st International Conference on Machine Learning. (2004)