Hierarchical Kernel Stick-Breaking Process for Multi-Task Image Analysis

Qi An QA@EE.DUKE.EDU
Chunping Wang CW36@EE.DUKE.EDU
Ivo Shterev IS33@EE.DUKE.EDU
Eric Wang EW28@EE.DUKE.EDU
Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

David B. Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science, Duke University, Durham, NC 27708

Abstract

The kernel stick-breaking process (KSBP) is employed to segment general imagery, imposing the condition that patches (small blocks of pixels) that are spatially proximate are more likely to be associated with the same cluster (segment). The number of clusters is not set *a priori* and is inferred from the hierarchical Bayesian model. Further, KSBP is integrated with a shared Dirichlet process prior to simultaneously model multiple images, inferring their inter-relationships. This latter application may be useful for sorting and learning relationships between multiple images. The Bayesian inference algorithm is based on a hybrid of variational Bayesian analysis and local sampling. In addition to providing details on the model and associated inference framework, example results are presented for several image-analysis problems.

1. Introduction

The segmentation of general imagery is a problem of longstanding interest. There have been numerous techniques developed for this purpose, including K-means and associated vector quantization methods (Ding & He, 2004), statistical mixture models (McLachlan & Basford, 1988), as well as spectral clustering (Ng et al., 2001). This list of existing methods is not exhaustive, although these methods share attributes associated with most existing algorithms. First, the clustering is based on the features of the image, and when clustering these features one typically does not account for their physical location within the image (although the location may be appended as a feature component). Secondly, the segmentation or clustering of images is typically performed one image at a time, and therefore there is no attempt to relate the segments of one image to segments in other images (*i.e.*, to learn inter-relationships between multiple images). Finally, in many of the techniques cited above one must *a priori* set the number of anticipated segments or clusters. The techniques developed in this paper seek to perform clustering or segmentation in a manner that explicitly accounts for the physical locations of the features within the image, and multiple images are segmented simultaneously (termed "multi-task learning") to infer their inter-relationships. Moreover, the analysis is performed in a semi-parametric manner, in the sense that the number of segments or clusters is not set *a priori*, and is inferred from the data. There has been recent research wherein spatial information has been exploited when clustering (Figueiredo et al., 2007), but that segmentation has been performed one image at a time, and therefore not in a multi-task setting.

To address the goals elucidated above within a statistical setting, we employ a class of hierarchical models related to the Dirichlet process (DP) (Ferguson, 1973). The Dirichlet process is a statistical prior that may be summarized succinctly as follows. Assume that the n-th patch is represented by feature vector $\boldsymbol{x}_n$, and the total image is composed of N such feature vectors $\{\boldsymbol{x}_n\}_{n=1,N}$. The feature vector associated with each patch is assumed drawn from a parametric distribution $f(\phi_n)$, where ϕ_n represents the parameters associated with the n-th feature vector. A DP prior can be placed on ϕ_n, which is characterized by the non-negative parameter α and the "base" distribution G_o. We adopt the stick-breaking construction developed by Sethuraman (Sethuraman, 1994), and the hierarchical model may be expressed as

Appearing in *Proceedings of the 25^{th} International Conference on Machine Learning*, Helsinki, Finland, 2008.

$$\begin{aligned} \boldsymbol{x}_n|\phi_n &\overset{ind}{\sim} f(\phi_n) \\ \phi_n|G &\overset{iid}{\sim} G \\ G &= \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h} \\ \pi_h &= V_h \prod_{l=1}^{h-1}(1-V_l) \\ V_h &\overset{iid}{\sim} Beta(1,\alpha) \\ \theta_h &\overset{iid}{\sim} G_o. \end{aligned} \tag{1}$$

This is termed a "stick-breaking" representation of DP because one sequentially breaks off "sticks" of length π_h from an original stick of unit length ($\sum_{h=1}^{\infty}\pi_h = 1$). As a consequence of the properties of the distribution $Beta(1,\alpha)$, for relatively small α it is likely that only a relatively small set of sticks π_h will have appreciable weight/size, and therefore when drawing parameters ϕ_n from the associated G it is probable multiple ϕ_n will share the same "atoms" θ_h (those associated with the large-amplitude sticks). The parameter α therefore plays an important role in defining the number of clusters that are constituted, and therefore in practice one typically places a non-informative Gamma prior on α (Xue et al., 2007).

The form of the model in (1) imposes the prior belief that the feature vectors $\{\boldsymbol{x}_n\}_{n=1,N}$ associated with an image should cluster, and the data are used to infer the most probable clustering distribution, via the posterior distribution on the parameters $\{\phi_n\}_{n=1,N}$. Such semi-parametric clustering has been studied successfully in many settings (Xue et al., 2007; Rasmussen, 2000). However, there are two limitations of such a model, with these defining the focus of this paper. First, while the model in (1) captures our belief that the feature vectors should cluster, it does not impose our additional belief that the probability that two feature vectors are in the same cluster should increase as their physical locations within the image become more proximate; this is an important factor when one is interested in segmenting an image into contiguous regions. Secondly, typical semi-parametric clustering has been performed one image or dataset at a time, and here we wish to cluster multiple images simultaneously, to infer the inter-relationships between clusters in different images, thereby inferring the inter-relationships between the associated multiple images themselves.

As an extension of the DP-based mixture model, we here consider the recently developed kernel stick-breaking process (KSBP) (Dunson & Park, 2008), introduced by Dunson and Park. As detailed below, this model is similar to that in (1), but now the stick-breaking process is augmented to employ a kernel function to quantify the prior belief associated with spatially proximate patches. In (Dunson & Park, 2008) a Markov chain Monte Carlo (MCMC) sampler was used to estimate the posterior on the model parameters. In the work considered here we are interested in relatively large data sets, and therefore we develop an inference engine that exploits ideas from variational Bayesian analysis (Beal, 2003).

There are problems for which one may wish to perform segmentation on multiple images simultaneously, with the goal of inferring the inter-relationships between the different images. This is referred to as multi-task learning (MTL) (Thrun & O'Sullivan, 1996; Xue et al., 2007), where here each "task" corresponds to clustering feature vectors from a particular image. As presented below, it is convenient to simultaneously cluster/segment multiple images by linking the multiple associated KSBP models with an overarching DP. There are at least three applications of MTL in the context of image analysis: (i) one may have a set of images, some of which are labeled, and others of which are unlabeled, and by performing an MTL analysis on all of the images one may infer labels for the unlabeled image segmentation, by drawing upon the relationships to the labeled imagery; (ii) by inferring the inter-relationships between the different images, one may sort the images as well as sort components within the images; (iii) one may identify abnormal images and locations within an image in an unsupervised manner, by flagging those locations that are allocated to a segmentation component that is locally rare. A similar scenario has been studied in (Sudderth et al., 2006), where the spatial translations are handled with transformed Dirichlet processes.

2. Kernel Stick-Breaking Process

2.1. KSBP prior for image processing

The stick-breaking representation of the Dirichlet process (DP) was summarized in (1), and this has served as the basis of a number of generalizations of the DP. The dependent DP (DDP) proposed by MacEachern (MacEachern, 1999) assumes a fixed set of weights, $\boldsymbol{\pi}$, while allowing the atoms $\theta = \{\theta_1, \cdots, \theta_N\}$ to vary with the predictor $\boldsymbol{x}$ according to a stochastic process. Dunson and Park (Dunson & Park, 2008) have proposed the kernel stick-breaking process (KSBP), which is particularly attractive for image-processing applications. Rather than simply considering the feature vector $\{\boldsymbol{x}_n\}_{n=1,N}$, we now consider $\{\boldsymbol{x}_n, \boldsymbol{r}_n\}_{n=1,N}$, where $\boldsymbol{r}_n$ is tied to the location of the pixel or block of pixels used to constitute feature vector $\boldsymbol{x}_n$. We let $K(\boldsymbol{r}, \boldsymbol{r}', \psi) \rightarrow [0,1]$ define a bounded kernel function with parameter ψ, where $\boldsymbol{r}$ and $\boldsymbol{r}'$ represent general locations in the image of interest. One may choose to place a prior on the kernel parameter ψ; this issue is revisited below. A draw $G_{\boldsymbol{r}}$ from a KSBP prior is a function of position $\boldsymbol{r}$, and is represented as

$$\begin{aligned} G_{\boldsymbol{r}} &= \sum_{h=1}^{\infty} \pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)\delta_{\theta_h} \\ \pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi) &= V_h K(\boldsymbol{r}, \Gamma_h, \psi) \textstyle\prod_{l=1}^{h-1}[1 - V_l K(\boldsymbol{r}, \Gamma_l, \psi)] \\ V_h &\overset{iid}{\sim} Beta(a, b) \qquad (2) \\ \Gamma_l &\overset{iid}{\sim} H \\ \theta_h &\overset{iid}{\sim} G_o. \end{aligned}$$

Dunson and Park (Dunson & Park, 2008) prove the validity of $G_{\boldsymbol{r}}$ as a probability measure. Comparing (1) and (2), both priors take the general form of a stick-breaking representation, while the KSBP prior possesses several interesting properties. For example, the stick weights $\pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)$ are a function of $\boldsymbol{r}$. Therefore, although the atoms $\{\theta_h\}_{h=1,\infty}$ are the same for all $\boldsymbol{r}$, the weights effectively shift the probabilities of different θ_h based on $\boldsymbol{r}$. The basis functions Γ_h serve to localize in the space of $\boldsymbol{r}$ regions (clusters) in which the weights $\pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)$ are relatively constant, with the size of these regions tied to the kernel parameter ψ.

If $f(\phi_n)$ is the parametric model (with parameter ϕ_n) responsible for the generation of $\boldsymbol{x}_n$, we now assume that the augmented data $\{\boldsymbol{x}_n, \boldsymbol{r}_n\}_{n=1,N}$ are generated as

$$\begin{aligned} \boldsymbol{x}_n &\overset{ind}{\sim} f(\phi_n) \\ \phi_n &\overset{ind}{\sim} G_{\boldsymbol{r}_n} \qquad (3) \\ G_{\boldsymbol{r}} &\sim KSBP(a, b, \psi, G_o, H). \end{aligned}$$

The notation $G_{\boldsymbol{r}} \sim KSBP(a, b, \psi, G_o, H)$ is meant to convey that $G_{\boldsymbol{r}}$ is drawn *one* time from the KSBP, and is a parametric function of location $\boldsymbol{r}$, and it is evaluated at specific locations $\{\boldsymbol{r}_n\}_{n=1,N}$.

The generative model in (3) states that two feature vectors that come from the same region in the image (defined via $\boldsymbol{r}$) will have similar $\pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)$, and therefore they are likely to share the same atoms θ_h. The settings of a and b control how much similarity there will be in drawn atoms for a given spatial cluster centered about a particular Γ_h. If we set $a = 1$ and $b = \alpha$, analogous to the DP, small concentration parameter α and/or small kernel parameter ψ will impose that π_h is likely to be near one, and therefore only a relatively small number of atoms θ_h are likely to be dominant for a given cluster spatial center Γ_h. On the other hand, if two features are generated from distant parts of a given image, the associated atoms θ_h that may be prominent for each feature vector are likely to be different, and therefore it is of relatively low probability that these feature vectors would have been generated via the same parameters ϕ. It is possible that the model may infer two distinct and widely separated clusters/segments with similar parameters (atoms); if the G_o within the KSBP is itself drawn from a DP, as it will be below when analyzing multiple images, widely separated clusters may share the exact same atoms.

For the case $a = 1$ and $b = \alpha$, which we consider below, we employ the notation $G_{\boldsymbol{r}} \sim KSBP(\alpha, \psi, G_o, H)$. Below we will also assume that $f(\phi)$ corresponds to a multivariate Gaussian distribution.

2.2. Spatial correlation properties

As indicated above, the functional form of the kernel function is important and needs to be chosen carefully. A commonly used kernel is given as $K(\boldsymbol{r}, \Gamma, \psi) = \exp(-\psi\|\boldsymbol{r} - \Gamma\|^2)$ for $\psi > 0$, which allows the associated stick weight to change continuously from $V_h \prod_{l=1}^{h-1}(1 - V_l)$ to 0 conditional on the distance between $\boldsymbol{r}$ and Γ. By choosing a kernel we are also implicitly imposing the dependency between the priors of two samples, $G_{\boldsymbol{r}}$ and $G_{\boldsymbol{r}'}$. Specifically, both priors are encouraged to share the same atoms θ_h if $\boldsymbol{r}$ and $\boldsymbol{r}'$ are close, with this discouraged otherwise. Dunson and Park (Dunson & Park, 2008) derive the correlation coefficient between two probability measures $G_{\boldsymbol{r}}$ and $G_{\boldsymbol{r}'}$ to be

$$\begin{aligned} & corr\{G_{\boldsymbol{r}}, G_{\boldsymbol{r}'}\} \\ &= \frac{\sum_{h=1}^{\infty} \pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)\pi_h(\boldsymbol{r}'; V_h, \Gamma_h, \psi)}{\sqrt{\sum_{h=1}^{\infty} \pi_h(\boldsymbol{r}; V_h, \Gamma_h, \psi)^2}\sqrt{\sum_{h=1}^{\infty} \pi_h(\boldsymbol{r}'; V_h, \Gamma_h, \psi)^2}}. \end{aligned}$$

The coefficient approaches unity in the limit as $\boldsymbol{r} \to \boldsymbol{r}'$. Since the correlation is a strong function of the kernel parameter ψ, below we will consider a distinct ψ_h for each stick. This implies that the spatial extent within the image over which a given stick is important will vary as a function of the stick (to accommodate textural regions of different sizes).

3. Multi-Task Image Segmentation with a Hierarchical KSBP

We now consider the problem for which we wish to jointly segment M images, where each image has an associated set of feature vectors with location information, in the sense discussed above. Aggregating the data across the M images, we have the set of feature vectors $\{\boldsymbol{x}_{nm}, \boldsymbol{r}_{nm}\}_{n=1,N_m;\, m=1,M}$. The image sizes may be different, and therefore the number of feature vectors N_m may vary between images. The premise of the model discussed below is that the cluster or segment characteristics may be similar between multiple images, and the inference of these inter-relationships may be of value. Note that the assumption is that sharing of clusters may be of relevance for the feature vectors, but not for the associated locations.

3.1. Model

A relatively simple means of sharing feature-vector clusters between the different images is to let each image be processed with a separate $KSBP(\alpha_m, \psi_m, G_m, H_m)$. To achieve the desired sharing of feature-vector clusters between the different images, we impose that $G_m \equiv G$ and G is drawn $G \sim DP(\gamma, G_o)$. Recalling the stick-breaking form of a draw from $DP(\gamma, G_o)$, we have $G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}$, in the sense summarized in (1). The discrete form of G is very important, for it implies that the different $G_{\boldsymbol{r}}$ will share the same set of discrete atoms $\{\theta_h\}_{h=1,\infty}$. It is interesting to note that for the case in which the kernel parameter ψ is set such that $K(\boldsymbol{r}, \Gamma_h, \psi) \rightarrow 1$, the hierarchical KSBP (H-KSBP) model reduces to the hierarchical Dirichlet process (HDP) (Teh et al., 2005).

Therefore, the H-KSBP model is represented as

$$\begin{aligned} \boldsymbol{x}_{nm} &\overset{ind}{\sim} \mathcal{N}(\phi_{nm}) \\ \phi_{nm} &\overset{ind}{\sim} G_{\boldsymbol{r}_{nm}} \\ G_{\boldsymbol{r}} &\sim KSBP(\alpha_m, \psi_m, G, H_m) \\ G &\sim DP(\gamma, G_o), \end{aligned} \tag{4}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution. Assume that G is composed of the atoms $\{\theta_h\}_{h=1,\infty}$, from the perspective of the stick-breaking representation in (1). These same atoms are shared across all $\{G_{\boldsymbol{r}_{nm}}\}_{n=1,N_m;m=1,M}$ drawn from the associated KSBPs, but with respective stick weights unique to the different images, and a function of position within a given image. The posterior inference allows one to infer which clusters of features are unique to a particular image, and which clusters are shared between multiple images. The density functions H_m are tied to the support of the m-th image, and in practice this is set as uniform across the image extent. The distinct α_m, for each of which a Gamma hyper-prior may be imposed, encourages that the number of clusters (segments) may vary between the different images, although one may simply wish to set $\alpha_m = \alpha$ for all M tasks.

For notational convenience, in (4) it was assumed that the kernel parameter ψ_m varied between tasks, but was fixed for all sticks within a given task; this is overly restrictive. In the implementation that follows the parameter ψ_{hm} may vary across tasks *and* across the task-specific KSBP sticks.

3.2. Posterior inference

For inference purposes, we truncate the number of sticks in the KSBP to T, and the number of sticks in the truncated DP to K (the truncation properties of the stick-breaking representation of DP are discussed in (Ishwaran & James, 2001), although we emphasize that when truncating KSBP one must take into account the draws from the Beta distribution *and* the properties of the kernel, to assure that the truncated set of sticks sum to one). Due to the discreteness of $G = \sum_{k=1}^{K} \beta_k \delta_{\theta_k}$, each draw of the KSBP, $G_{\boldsymbol{r}_{nm}} = \sum_{h=1}^{T} \pi_{hm} \delta_{\phi_{hm}}$, can only take atoms $\{\phi_{hm}\}_{h=1,T;\, m=1,M}$ from K unique possible values $\{\theta_k\}_{k=1,K}$; when drawing atoms ϕ_{hm} from G, the respective probabilities for $\{\theta_k\}_{k=1,K}$ are given by $\{\beta_k\}_{k=1,K}$, and for a given $\boldsymbol{r}_{nm}$ the respective probabilities for different $\{\phi_{hm}\}_{h=1,T;\, m=1,M}$ are defined by $\{\pi_{hm}\}_{h=1,T;\, m=1,M}$. In order to reflect the correspondences between the data and atoms explicitly, we further introduce two auxiliary indicator variables. One is z_{nm}, this indicating which stick of the KSBP the feature vector $\boldsymbol{x}_{nm}$ is associated, and the other is t_{hm}, this indicating which mixing component θ_k the atom ϕ_{hm} is associated with.

With this specification we can represent our H-KSBP mixture model via a stick-breaking characterization. A graphical representation of the proposed H-KSBP model is provided in Figure 1.

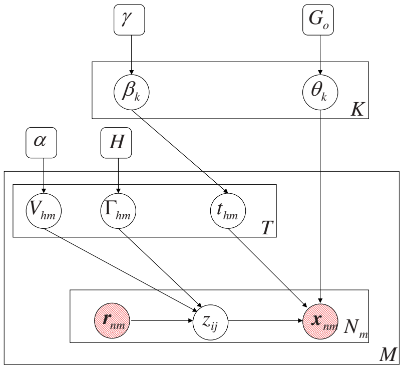

Figure 1. A graphical representation of the H-KSBP mixture model.

For the large-scale problems of interest here we employ variational Bayesian (VB) inference, which has proven to be a relatively fast (compared to MCMC) and accurate inference tool for many models and applications (Beal, 2003; Blei & Jordan, 2004). To employ VB, a conjugate prior is required for all variables in the model. In the proposed model, we however cannot obtain a closed form for the variational posterior distribution of the node V_{hm}, because of the the kernel function. Alternatively, motivated by the Monte Carlo Expectation Maximization (MCEM) algorithm (Wei & Tanner, 1990), we develop a Monte Carlo Variational Bayesian (MCVB) inference algorithm, where the intractable nodes are approximated with Monte Carlo samples from their conditional posterior distributions. The resulting algorithm combines the benefits of both MCMC and VB, and has proven to be effective for the examples we have considered (some of which are presented here).

Given the H-KSBP mixture model detailed in Section 3.1, we can follow standard variational Bayesian inference (Beal, 2003) to infer the variables of interests. All the updates are analytical except for V_{hm}, which is estimated with the samples from its conditional posterior distributions. Due to the limited space, we only consider the update for V_{hm}. To obtain the conditional posterior distribution of V_{hm}, we rewrite $z_{nm} = min\{h : A_{nm,h} = B_{nm,h} = 1\}$, with two auxiliary variables defined as: $A_{nm,h} \sim Bernoulli(V_{hm})$ and $B_{nm,h} \sim Bernoulli(K(\boldsymbol{r}_{nm}, \Gamma_{hm}, \psi_m))$.

The conditional posterior distributions of V_{hm} are

$$Beta(1 + \sum_{n:z_{nm} \geq h} A_{nm,h}, \alpha + \sum_{n:z_{nm} \geq h} (1 - A_{nm,h})),$$

where

$$\begin{aligned}
p(A_{nm,h} = B_{nm,h} = 0) &= \tfrac{(1-V_{hm})(1-K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m))}{1-V_{hm}K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m)} \\
p(A_{nm,h} = 0, B_{nm,h} = 1) &= \tfrac{(1-V_{hm})K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m)}{1-V_{hm}K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m)} \\
p(A_{nm,h} = 1, B_{nm,h} = 0) &= \tfrac{V_{hm}(1-K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m))}{1-V_{hm}K(\boldsymbol{r}_{nm},\Gamma_{hm},\psi_m)},
\end{aligned}$$

for $h = 1, 2, \cdots, z_{nm} - 1$, and $A_{nm,h} = B_{nm,h} = 1$ for $h = z_{nm}$.

The hyper-parameters α, γ, and $\boldsymbol{\psi}$ are assumed to be constant for inference of the other parameters. However, since the model performance may be sensitive to the settings of those hyper-parameters, we can relax this assumption by placing non-informative priors. The updates are straightforward (Beal, 2003) and therefore omitted here.

3.3. Convergence

To monitor the convergence of our MCVB algorithm, we compute the lower bound of the log model evidence at each iteration. Because of the sampling of some variables, the lower bound does not in general increase monotonically, but we observed in all experiments that the lower bound increases sequentially for the first several iterations, with generally small fluctuations after it has converged to the local optimal solution.

4. Experimental Results

We have applied the H-KSBP multi-task image-segmentation algorithm to both synthetic and real images. We first present results on synthesized imagery, wherein we compare KSBP-based clustering of a single image with associated DP-based clustering. We then consider H-KSBP as applied to actual imagery, taken from a widely utilized database. The hyper-priors in the model for the examples are set as follows: Gamma priors, $G(\tau_{10}, \tau_{20})$ and $G(\tau_{30}, \tau_{40})$, for α and γ with parameter $\tau_{10} = 1e^{-2}$, $\tau_{20} = 1e^{-2}$, $\tau_{30} = 3e^{-2}$, $\tau_{40} = 3e^{-2}$, respectively; a normal-Wishart prior, $N(\boldsymbol{\mu}_k|\boldsymbol{\mu}_0, \eta_0\boldsymbol{\Sigma}_k)W(\boldsymbol{\Sigma}_k|w_*, \boldsymbol{\Sigma}_*)$, conjugate to the Gaussian distribution with $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\eta_0 = 1$, $w_* = d + 2$, $\boldsymbol{\Sigma}_* = 5 \times \boldsymbol{I}$; the discrete priors for $\boldsymbol{\Gamma}$ and $\boldsymbol{\psi}$ with uniform weights over all candidates. The stick-breaking truncations are $K = 40, T = 40$.

4.1. Single image segmentation

In this simple illustrative example, each feature vector is associated with a particular pixel, and the feature is simply a real number, corresponding to its intensity; the pixel location is the auxiliary information within the KSBP, while this information is not employed by the DP-based segmentation algorithm. Figure 2 shows the original image and the segmentation results of both algorithms. In Figure 2(a) we note that there are five contiguous regions for which the intensities are similar. There is a background region with a relatively fixed intensity, and within this are four distinct contiguous sub-regions, and of these there are pairs for which the intensities are comparable. The data in Figure 2(a) were generated as follows. Each pixel in each region is generated independently as a draw from a Gaussian distribution; the standard deviation of each of the Gaussians is 10, and the background has mean intensity 5, and the two pairs are generated with mean intensities of 40 and 60. The color bar in Figure 2(a) denotes the pixel amplitudes. The DP and KSBP segmentation results are shown in Figures 2(b) and 2(c), respectively. A distinct color is associated with distinct cluster parameters. In the DP results we note that the four subregions are generally properly segmented, but there is significant speckle in the background region. The KSBP segmentation algorithm is beset by far less speckle. Further, in the KSBP results there are five distinct clusters (dominant KSBP sticks), where in the DP results there are principally three distinct sticks (in the DP, the spatially separated segments with the same features are treated as one cluster, while in the KSBP each contiguous region is represented by its own stick).

In the next set of results, on real imagery, we employ the H-KSBP algorithm, and therefore at the task level segmentation is performed as in Figure 2(c). Alternatively, using the HDP model (Teh et al., 2005), at the task level one employs clustering of the form in Figure 2(b). The relative performance of H-KSBP and HDP is analyzed.

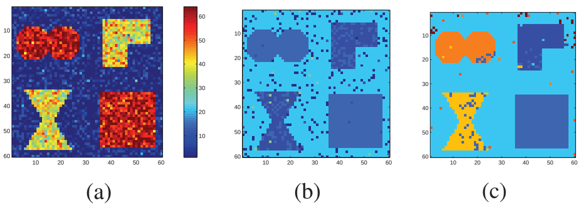

Figure 2. A synthetic image example. (a) Original synthetic image, (b) image-segmentation results of DP-based model, and (c) image-segmentation results of KSBP-based model.

4.2. H-KSBP applied to a set of real images

Within the subsequent image analysis we employ features constituted by the independent feature subspace analysis (ISA) technique, developed by Hyvärinen and Hoyer (Hyvärinen & Hoyer, 2000). These features have proven to be relatively shift or translation invariant, which enables them to be widely applicable to many type of images.

We test the H-KSBP model on a subset of images from Microsoft Research Cambridge, available at http://research.microsoft.com/vision/cambridge/recognition/. There are seven types of images used in this database: buildings, clouds, countryside, faces, fireworks, offices and urban. Twenty images are randomly selected from the database for each type, yielding a total of 140 images. To capture textural information within the features, we first divided each image into a contiguous 24 × 24-pixel non-overlapping patches (more than 70,000 patches in total) and then extract ISA features from each patch; color images are considered, and the RGB colors are handled within ISA feature extraction as in (Hoyer & Hyvärinen, 2000). Concerning learning the ISA independent feature subspaces, we randomly select 150 patches out of each of the 140 images from the seven classes, and these 150 image patches are used for basis training. The posterior on the H-KSBP (and HDP) model parameters is inferred based on the proposed MCVB algorithm, processing all 140 images simultaneously; as discussed in Section 2, the HDP analysis is performed by a special setting of the H-KSBP parameters. To mitigate the influence of random samples and VB initialization, we perform the experiment ten times and report the average results.

Borrowing the successful "bag of words" assumption in text analysis (Blei & Lafferty, 2005), we assume each image is a bag of atoms, which results in a measurable quantity of inter-relationship between images, specifically similar images should share similar distribution over those mixture components. An important aspect of the H-KSBP algorithm is that while in text analysis the "bag of words" may be set *a priori*, here the "bag of atoms" is inferred from the data itself, within the clustering process. Related concepts have been employed previously in image analysis (Quelhas et al., 2007), but in that work one had to set the canonical set of image atoms (shapes) *a priori*, which is somewhat *ad hoc*.

As an example, for the data considered, we show one realization of H-KSBP in Figure 3. In the figure, we display canonical atom usage across all 140 images. Figure 3 is a count matrix, where each square represents the relative number of counts in a given image for a particular atom (atoms indexed along the vertical axis in Figure 3).

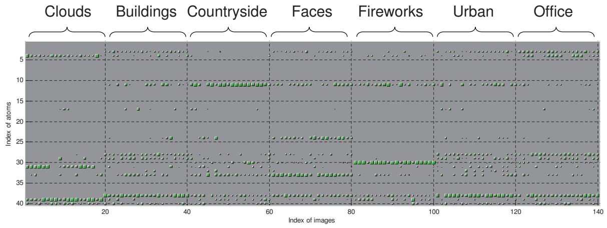

Figure 3. Matrix on the usage of atoms across the different images. The size of each box represents the relative frequency with which a particular atom is manifested in a given image. These results are computed via H-KSBP.

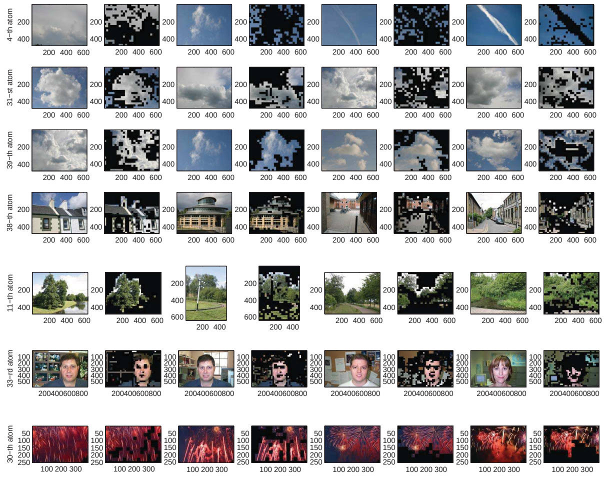

Figure 4. Demonstration of different atoms as inferred by an example run of the H-KSBP algorithm. Each row of the figure corresponds to one atom. Every two images form a set, with the original images at left and areas assigns to a particular atom shown at right.

Figure 4 gives a representation of most of the atoms. For example the 4-th, 31-st and 39-th atoms are associated with clouds and sky; the 38-th atom is principally modeling buildings; and the 11-th atom is associated with trees and grasses. While performing the experiment, we also noticed it was relatively easy to segment clouds, fireworks, countryside, and urban images while harder to obtain contiguous segments within office images (these typically have far more details, and less large regions of smooth texture; this latter issue may be less an issue of the H-KSBP, but rather of the features employed). An example of this difficulty is observable in Figure 5, as office images are composed of many different atoms. Fortunately, the office images still tend to share similar usage of atoms so that they can be grouped together (sorted) when quantifying similarities between images based on the histogram over atoms (discussed next).

The results in Figure 5, in which both H-KSBP and HDP segmentation results are presented, demonstrate general properties observed when analyzing the images considered here: (i) the segmentation characteristics of HDP were generally good, but on some occasions they were markedly worse (less detailed) than those of H-KSBP; and (ii) the H-KSBP was generally more sensitive to detailed textural differences in the images, thereby generally inferring a larger number of principal atoms (increased number of large sticks).

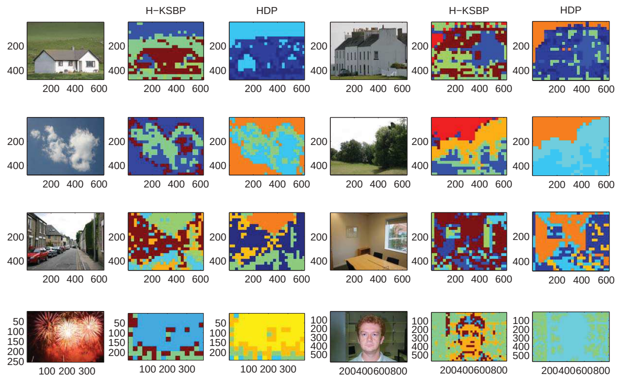

Figure 5. Representative set of segmentation results, comparing H-KSBP and HDP. While these two algorithms tend to generally yield comparable segmentations for the images considered, the H-KSBP is generally more sensitive to details, with this sometimes yielding better segmentations ($e.g.$, the top-level and bottom-right results).

To demonstrate the image-sorting potential of the H-KSBP, we compute the Kullback-Leibler (KL) divergence on the histogram over atoms between any two images, by averaging histograms of the form in Figure 3 over ten random MCVB initializations. For each image, we rank its similarity to all other images based on the associated KL divergence. Performance is addressed quantitatively as follows. For each of the 140 images, we quantify via KL divergence its similarity to all other 139 images, wherein we achieve in ordered list. In Figure 6 we present a confusion matrix, which represents the fraction of the top-ten members of this ordered list that are within the same class (among seven classes) as the image under test.

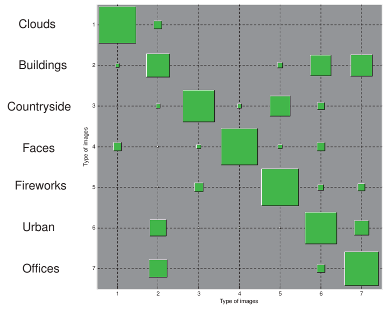

Figure 6. The confusion matrix over image types, generated using H-KSBP.

As demonstrated in Figure 6, the H-KSBP performs well in distinguishing clouds, faces and fireworks images. The buildings and urban images often share some similar atoms, mainly representing buildings, and therefore these are somewhat confused (reasonably, it is felt). The offices images are often related to other relatively complex scenes. Some typical image ranking results are given in Figure 7. It was found that the HDP produced similar sorting results as produced by H-KSBP ($e.g.$, the associated confusion matrix for HDP is similar to that in Figure 6), and therefore the HDP sorting results are omitted here for brevity. This indicates that while in some cases the HDP segmentation results are inferior to those of H-KSBP, in general the ability of HDP and H-KSBP to sort images is comparable (at least for the set of images considered).

The H-KSBP results on the 140-image database were performed in non-optimized $Matlab^{TM}$ software, on a PC with 3 GHz CPU and 2 GB memory. It required about 3 hours to compute one run of the MCVB code for 80 iterations, with typically 40-50 iterations required to achieve convergence. The H-KSBP and HDP algorithms were run with comparable computation times.

5. Conclusions

The kernel stick-breaking process has been extended for use in image segmentation. The algorithm explicitly imposes the belief that feature vectors that are generated from proximate locations in an image are more likely to be associated with the same image segment. We have also extended the KSBP algorithm to the MTL setting, exploring the inter-relationship of images by sharing the same mixing components. Generally superior segmentation performance of H-KSBP was observed relative to HDP, when

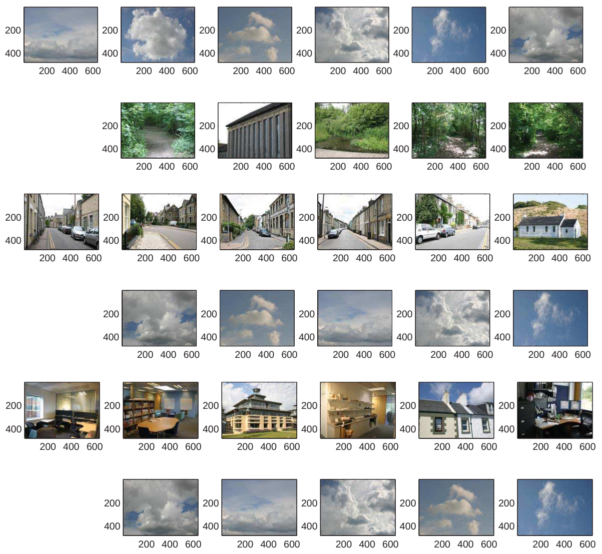

Figure 7. Sample image sorting results, as generated by H-KSBP. The top left image is the original image followed by the five most similar images and then the five most dissimilar images.

segmenting multiple images simultaneously. In addition to segmenting multiple images, the H-KSBP and HDP algorithms also yield information about the inter-relationships between the images, based on the underlying sharing mechanisms inferred among the associated clusters. For the images considered, it was found that the H-KSBP and HDP yielded very similar sorting results.

References

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, University College London.

Blei, D., & Jordan, M. (2004). Variational methods for the Dirichlet process. *Proc. the 21st International Conference on Machine Learning*.

Blei, D., & Lafferty, J. (2005). Correlated topic models. *Advances in Neural Information Processing System*.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proc. the International Conference on Machine Learning* (pp. 225–232).

Dunson, D., & Park, J.-H. (2008). Kernel stick-breaking process. *Biometrika*.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*.

Figueiredo, M., Cheng, D., & Murino, V. (2007). Clustering under prior knowledge with application to image segmentation. *Advances in Neural Information Processing System*.

Hoyer, P., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, *11*, 191–210.

Hyvärinen, A., & Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, *12*, 1705–1720.

Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 161–173.

MacEachern, S. (1999). Dependent nonparametric process. *ASA Proceeding of the Section on Bayesian Statistical Science*. Alexandria, VA.

McLachlan, G., & Basford, K. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 13*.

Quelhas, P., F. Monay, J.-M. O., Gatica-Perez, D., & Tuytelaars, T. (2007). A thosand words in a scenes. *IEEE Trans. Pattern Analysis Machine Intell.*, *9*, 1575–1589.

Rasmussen, C. (2000). The infinite gaussian mixture model. *Advances in Neural Information Processing System* (pp. 554–560).

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2006). Describing visual scenes using transformed Dirichlet processes. *NIPS 18* (pp. 1297–1304).

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2005). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1582.

Thrun, S., & O'Sullivan, J. (1996). Discovering structure in multiple learning tasks: The TC algorithm. *Proc. the 13th International Conference on Machine Learning*.

Wei, G., & Tanner, M. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699–704.

Xue, Y., Liao, X., Carin, L., & Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research*, *8*, 35–63.