Nonparametric Clustering with Distance Dependent Hierarchies

Soumya Ghosh
Dept. of Computer Science,
Brown Univ., Providence, RI
sghosh@cs.brown.edu

Michalis Raptis
Comcast Labs,
Washington, D.C.
mraptis@cable.comcast.com

Leonid Sigal
Disney Research,
Pittsburgh, PA
lsigal@disneyresearch.com

Erik B. Sudderth
Dept. of Computer Science,
Brown Univ., Providence, RI
sudderth@cs.brown.edu

Abstract

The distance dependent Chinese restaurant process (ddCRP) provides a flexible framework for clustering data with temporal, spatial, or other structured dependencies. Here we model multiple groups of structured data, such as pixels within frames of a video sequence, or paragraphs within documents from a text corpus. We propose a hierarchical generalization of the ddCRP which clusters data within groups based on distances between data items, and couples clusters across groups via distances based on aggregate properties of these local clusters. Our hddCRP model subsumes previously proposed hierarchical extensions to the ddCRP, and allows more flexibility in modeling complex data. This flexibility poses a challenging inference problem, and we derive a MCMC method that makes coordinated changes to data assignments both within and between local clusters. We demonstrate the effectiveness of our hddCRP on video segmentation and discourse modeling tasks, achieving results competitive with state-of-the-art methods.

1 INTRODUCTION

The recent explosive growth of image and video repositories, and of structured data collections more broadly, motivates methods for the unsupervised discovery of informative latent structures. Image sequences of course exhibit strong spatio-temporal dependencies: objects typically occupy blocks of spatially contiguous pixels, and their movements induce strong dependencies among video frames. Nevertheless, many previous nonparametric models for visual data have mostly ignored such relationships, relying on careful feature engineering to make local likelihoods informative (Sudderth et al., 2008; Haines & Xiang, 2012). While accounting for spatial dependencies can be technically challenging, it produces image partitions which much more accurately reflect real-world scene structure (Orbanz & Buhmann, 2008; Sudderth & Jordan, 2008). However, these methods treat images as an unordered, or *exchangeable*, collection; they thus fail to capture the strong temporal dependencies found in video sequences.

Blei & Frazier (2011) proposed the *distance dependent Chinese restaurant process* (ddCRP) as a flexible distribution over partitions of data with temporal, spatial, or other non-exchangeable structure. The ddCRP represents partitions via links between data instances: each observation links to one other, and the probability of linking to nearby instances is higher. Closeness is measured according to a distance which may be arbitrarily specified to capture domain knowledge. The connected components of the induced link graph then partition the dataset into clusters. Previous work has used the ddCRP to effectively cluster data with sequential, temporal, or spatial structure (Ghosh et al., 2011; Socher et al., 2011; Ghosh et al., 2012).

In this paper, we propose a *hierarchical ddCRP* (hddCRP) that captures local relationships like these, but also uses distances among latent clusters to extract further global dependencies. After an initial ddCRP partitioning, local clusters are grouped via additional links that depend on a user-specified measure of cluster similarity. This framework allows the hddCRP to model relationships that depend on *aggregate* properties of clusters such as size and shape, which may be difficult to capture with likelihoods alone. Given arbitrary cluster and data affinity functions, which need not arise from true distance metrics, the hddCRP always defines a valid joint probability distribution on partitions.

The hddCRP is a hierarchical generalization of the ddCRP which unifies and generalizes existing models. Simpler hierarchical extensions of the ddCRP employing restricted distance functions (Ghosh et al., 2011; Kim & Oh, 2011), as well as the "Chinese restaurant franchise" representation of the *hierarchical Dirichlet process* (HDP, Teh et al. (2006)), are special cases of the hddCRP. The HDP and related dependent Dirichlet process models (MacEachern, 1999) define dependent random measures from which allocations of data to clusters are sampled, indirectly inducing dependencies in the resulting partitions. For example, Griffin & Steel (2006), Dunson & Park (2008), Rao & Teh (2009), and Lin et al. (2010) define priors which encourage "close" data points to have similar allocation distributions.

In contrast, the hddCRP directly specifies distributions over partitions via a flexible set of user-specified affinity functions. This allows structural constraints on clusters, such as connectivity (Ghosh et al., 2011), to be directly enforced. The hddCRP does not require its "distance" functions to be true metrics or have any special properties, and thus provides an extremely flexible framework for modeling complex data. Alternative models based on latent Gaussian processes (Duan et al., 2007; Sudderth & Jordan, 2008) require appropriate positive-definite kernel functions, whose specification can be challenging in non-Euclidean spaces (e.g., of object shapes). By working directly with discrete partitions, rather than latent continuous measures, the hddCRP also allows more computationally efficient inference.

The hddCRP generative process defined in Section 2 is simple, but the data-level and cluster-level link variables are strongly coupled in the posterior. Section 3 develops a *Markov chain Monte Carlo* (MCMC) method that makes coordinated changes to links at both levels, and thus more effectively explores clustering hypotheses. This sampler is also a novel inference algorithm for the HDP that makes large changes to the partition structure, without needing to explicitly craft split or merge proposals (Jain & Neal, 2004; Wang & Blei, 2012). By reasoning about data and cluster links, our sampler changes cluster allocations at varying resolutions, perturbing both memberships of data instances to local clusters and clusters to global components.

In Section 4, we demonstrate the versatility of the hddCRP by applying it to the problems of video segmentation and discourse analysis. In addition to having diverse data types (video sequences versus text documents), these two problems exhibit very different kinds of relationships among data instances and latent clusters. Nevertheless, our hddCRP model and inference framework easily applies to both domains by selecting appropriate data and cluster-level affinity functions. In both domains, explicit modeling of dependencies between latent clusters boosts performance over models that ignore such relationships.

2 HIERARCHICAL DISTANCE DEPENDENT CLUSTERS

The distance-dependent CRP (Blei & Frazier, 2011) defines a distribution over partitions indirectly via distributions over links between data instances. A data point i has an associated link variable c_i which links to another data instance j, or itself, according to the following distribution:

$$p\left(c_i = j \mid A, \alpha\right) \propto \begin{cases} A_{ij} & i \neq j, \\ \alpha & i = j. \end{cases} \qquad (1)$$

The *affinity* $A_{ij} = f(d(i,j))$ depends on a user-specified *distance* $d(i,j)$ between pairs of data points, and a monotonically decreasing *decay function* $f(d)$ which makes links to nearby data more likely. The resulting link structure induces a partition, where two data instances are assigned to the same cluster if and only if one is reachable from the other by traversing the link edges. Larger self-affinity parameters α favor partitions with more clusters.

2.1 THE HIERARCHICAL ddCRP

We propose a novel generative model that applies the ddCRP formalism twice, first for clustering data within each group into local clusters, and then for coupling the local clusters across groups. Like the ddCRP, our hddCRP defines a valid distribution over partitions of a dataset. It places higher probability mass on partitions that group nearby data points into latent clusters, *and* couple similar local clusters into global components. Examples of these data and cluster links are illustrated in Figure 1.

Consider a collection of G groups, where group g contains N_g observations. We denote the i^{th} data point of group g by x_{gi}, and the full dataset by $\mathbf{x}$. The data link variable c_{gi} for x_{gi} is sampled from a group-specific ddCRP:

$$p(c_{gi} = gj \mid \alpha_g, A^g) \propto \begin{cases} A^g_{ij} & i \neq j, \\ \alpha_g & i = j. \end{cases} \qquad (2)$$

At this first level of link variables, we set the probability of linking observations in different groups to zero. The connected components of the links $c_g = \{c_{gi} \mid i = 1, \ldots, N_g\}$ then determine the local clustering for group g.

Data links $\mathbf{c} = \{c_1, \ldots, c_G\}$ across all groups divide the dataset into group-specific local clusters $T(\mathbf{c})$. The hddCRP then associates each cluster $t \in T(\mathbf{c})$ with a cluster link k_t drawn from a global ddCRP distribution:

$$p(k_t = s \mid \alpha_0, A^0(\mathbf{c})) \propto \begin{cases} A^0_{ts}(\mathbf{c}) & t \neq s, \\ \alpha_0 & t = s. \end{cases} \qquad (3)$$

Here α_0 is a global self-affinity parameter, and $A^0(\mathbf{c})$ is the set of pairwise affinities between the elements of $T(\mathbf{c})$. We let $A^0_{ts}(\mathbf{c}) = f_0(d_0(t,s,\mathbf{c}))$, where $d_0(t,s,\mathbf{c})$ is a "distance" based on arbitrary properties of clusters t and s, and $f_0(d_0)$ a decreasing decay function. The connected components of $\mathbf{k} = \{k_t \mid t \in T(\mathbf{c})\}$ then couple local clusters into global components shared across groups. Let z_{gi} denote the component associated with observation i in group g, and $\mathbf{z} = \{z_{gi} \mid g = 1, \ldots, G; i = 1, \ldots, N_g\}$. Data instances x_{gi} and x_{hj} are clustered ($z_{gi} = z_{hj}$) if and only if they are reachable via some combination of data and cluster links.

Given this partition structure, we endow component m with likelihood parameters $\phi_m \sim G_0(\lambda)$, and generate observations $x_{gi} \sim p(x_{gi} \mid \phi_{z_{gi}})$. Let $M(\mathbf{c}, \mathbf{k})$ equal the number of global components induced by the cluster links $\mathbf{k}$ and data links $\mathbf{c}$. Because data links $\mathbf{c}$ are conditionally independent given $A^{1:G}$, and cluster links $\mathbf{k}$ are conditionally independent given $\mathbf{c}$ and the cluster affinities $A^0(\mathbf{c})$, the hddCRP

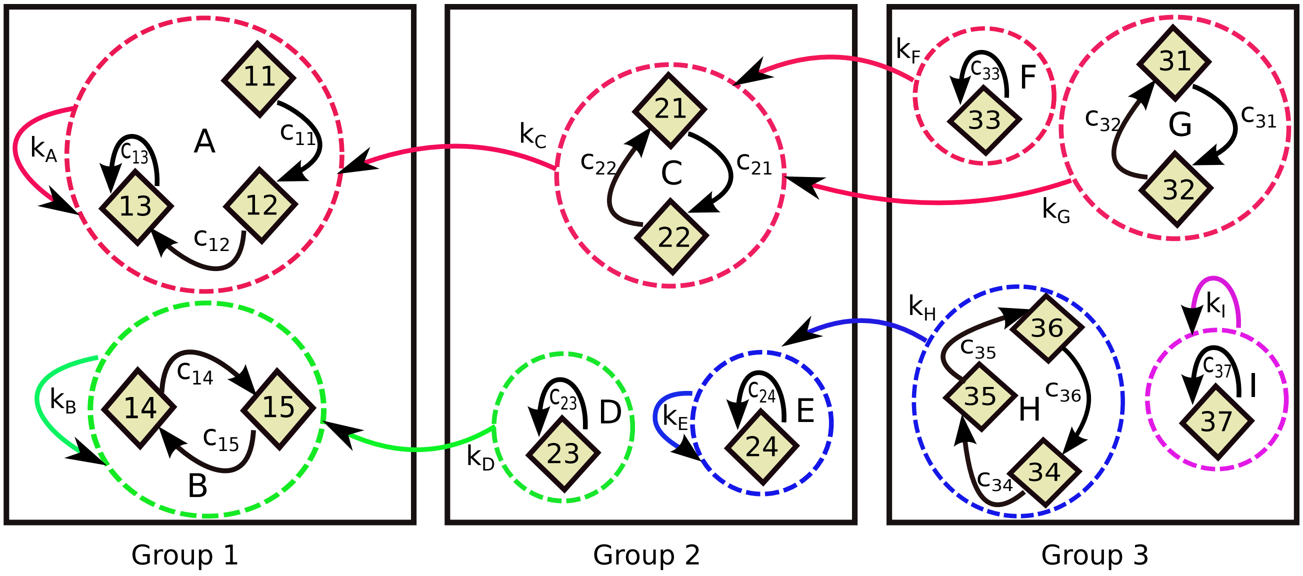

Figure 1: An example link variable configuration for a hierarchical ddCRP model of three groups (rectangles). Observed data points (customers, depicted as diamonds) link to other data points in the same group (black arrows), producing local clusters (dashed circles, labeled A through I). Cluster links (colored arrows) then join clusters to produce (in this case, four) global mixture components.

joint distribution on partitions and observations equals

$$p(\mathbf{x}, \mathbf{k}, \mathbf{c} \mid \alpha_{1:G}, \alpha_0, A^{1:G}, A^0, \lambda) = \prod_{m=1}^{M(\mathbf{c},\mathbf{k})} p(x_{\mathbf{z}=m} \mid \lambda)$$
$$\prod_{g=1}^{G}\prod_{i=1}^{N_g} p(c_{gi} \mid \alpha_g, A^g) \prod_{k_t \in \mathbf{k}} p(k_t \mid \mathbf{c}, \alpha_0, A^0(\mathbf{c})) \quad (4)$$

The set of data in component m is denoted by $x_{\mathbf{z}=m}$, and

$$p(x_{\mathbf{z}=m} \mid \lambda) = \int \prod_{gi \mid z_{gi}=m} p(x_{gi} \mid \phi_m)\, dG_0(\phi_m \mid \lambda), \quad (5)$$

where λ are hyperparameters specifying the prior distribution G_0. Our inference algorithms assume this integral is tractable, as it always is when an exponential family likelihood is coupled with an appropriate conjugate prior. We emphasize that for arbitrary data and cluster affinities, the sequential hddCRP generative process defines a valid joint distribution $p(\mathbf{x}, \mathbf{k}, \mathbf{c}) = p(\mathbf{c})p(\mathbf{k} \mid \mathbf{c})p(\mathbf{x} \mid \mathbf{k}, \mathbf{c})$.

2.2 RELATED HIERARCHICAL MODELS

The hddCRP subsumes several recently proposed hierarchical extensions to the ddCRP, as well as the HDP itself, by defining appropriately restricted data affinities and local cluster affinities. Blei & Frazier (2011) show that the CRP is recovered from the ddCRP by arranging data in an arbitrary sequential order, and defining affinities as

$$A_{ij} = \begin{cases} 1 & \text{if } i < j, \\ 0 & \text{if } i > j. \end{cases} \quad (6)$$

Data points link to all previous observations with equal probability, and thus the probability of joining any existing cluster is proportional to the number of other data points already in that cluster. The probability of creating a new cluster is proportional to the self-connection weight α. The resulting distribution on partitions can be shown to be invariant to the chosen sequential ordering of the data, and thus the standard CRP is *exchangeable* (Pitman, 2002).

Hierarchical Chinese Restaurant Process (hCRP) The hCRP representation of the HDP, which Teh et al. (2006) call the "Chinese restaurant franchise", is recovered from the hddCRP by first defining group-specific affinities as in Eq. (6). We then arrange local clusters (tables, in the CRF metaphor) t sequentially with distances $A^0_{ts}(\mathbf{c}) = 1$ if $t < s$, and $A^0_{ts}(\mathbf{c}) = 0$ if $t > s$. Just as the two-level hCRP arises from a sequence of CRPs, the hddCRP is defined from a sequence of two ddCRP models.

Naive Hierarchical ddCRP (naive-hddCRP) The image segmentation model of Ghosh et al. (2011) clusters data within each group via a ddCRP based on an informative distance (in their experiments, spatial distance between image pixels). A standard CRP, as in the upper level of the HDP, is then used to combine these clusters into larger segments. Inference is substantially simpler for this special case, because cluster distances do not depend on properties of the data assigned to those clusters.

Distance Dependent Chinese Restaurant Franchise An alternate approach to capturing group-specific metadata uses a standard CRP to locally cluster data, but then uses the group labels to define affinities between clusters. Kim & Oh (2011) use this model to learn topic models of time-stamped documents. By constraining cluster affinities to depend on group labels, but not properties of the data assigned to within-group clusters, inference is simplified.

3 MCMC INFERENCE

The posterior distribution over the data and cluster links $p(\mathbf{c}, \mathbf{k} \mid \mathbf{x}, \alpha_{1:G}, \alpha_0, A^{1:G}, A^0, \lambda)$ is intractable, and we thus explore it via a Metropolis-Hastings MCMC method. Our approach generalizes the non-hierarchical ddCRP Gibbs sampler of Blei & Frazier (2011), which iteratively samples single data links conditioned on the observations and other data links. Evolving links lead to splits, merges, and other large changes to the partition structure. In the hddCRP, local clusters belong to global components, and these component memberships must be sampled as well.

3.1 MARKOV CHAIN STATE SPACE

The number of possible non-empty subsets (clusters) of N data points is $2^N - 1$. The state space of our Markov chain consists of the data links $\mathbf{c}$, and the set of *all* possible cluster links $\mathcal{K}$, one for each candidate non-empty cluster. For instance, given three observations $\{h, i, j\}$ the set of non-empty subsets is $\mathcal{T} = \{[h], [i], [j], [hi], [ij], [jh], [hij]\}$, and the corresponding set of possible cluster links is $\mathcal{K} = \{k_h, k_i, k_j, k_{hi}, k_{ij}, k_{jh}, k_{hij}\}$, where $|\mathcal{K}| = 2^3 - 1$.

For any configuration of $\mathbf{c}$, a strict subset of $\mathcal{T}$ will have data associated with it. We call this the *active set*. For instance, if $c_h = h, c_i = i, c_j = j$, then only the clusters $\{[h], [i], [j]\}$ and the corresponding links $\{k_h, k_i, k_j\}$ are active. Given $\mathbf{c}$, we split $\mathcal{K}$ into the active set $\mathbf{k}$, and the remaining inactive cluster links $\tilde{\mathbf{k}} = \mathcal{K} \setminus \mathbf{k}$. We account for the inactive clusters by augmenting $A^0(\mathbf{c})$ as follows:

$$\tilde{A}^0(\mathbf{c}) = \begin{bmatrix} A^0(\mathbf{c}) & \mathbf{0} \\ \mathbf{0} & \alpha_0 \mathbf{I} \end{bmatrix}. \tag{7}$$

Here, we have sorted the links so that affinities among the active clusters are listed in the upper-left quadrant of $\tilde{A}^0(\mathbf{c})$. As indicated by the identity matrix $\mathbf{I}$, inactive clusters have zero affinity with all other clusters, and link to themselves with probability one. Under this augmented model, the joint probability factorizes as follows:

$$p(\mathbf{x}, \mathbf{k}, \tilde{\mathbf{k}}, \mathbf{c}) = p(\mathbf{c}) p(\mathbf{k} \mid \mathbf{c}) p(\tilde{\mathbf{k}} \mid \mathbf{c}) p(\mathbf{x} \mid \mathbf{c}, \mathbf{k}, \tilde{\mathbf{k}}) =$$

$$p(\mathbf{c}) p(\mathbf{k} \mid \mathbf{c}) p(\tilde{\mathbf{k}} \mid \mathbf{c}) p(\mathbf{x} \mid \mathbf{c}, \mathbf{k}) = p(\mathbf{x}, \mathbf{k}, \mathbf{c}) p(\tilde{\mathbf{k}} \mid \mathbf{c}). \tag{8}$$

Here, we have recovered the joint distribution of Eq. (4) because given $\mathbf{c}$, the observations $\mathbf{x}$ are conditionally independent of the inactive links $\tilde{\mathbf{k}}$. Crucially, because inactive cluster links have no uncertainty, we must only explicitly represent the active clusters at each MCMC iteration.

As the Markov chain evolves, clusters are swapped in and out of the active set. Although the number of active clusters varies with the state of the chain, the dimensionality of the augmented state space $(\mathbf{c}, \mathbf{k}, \tilde{\mathbf{k}})$ remains constant, allowing us to ignore complications that arise when dealing with chains whose state spaces have varying dimensionality. In particular, we employ standard *Metropolis-Hastings* (MH) proposals to change data and cluster links, and need not resort to reversible jump MCMC (Green, 1995).

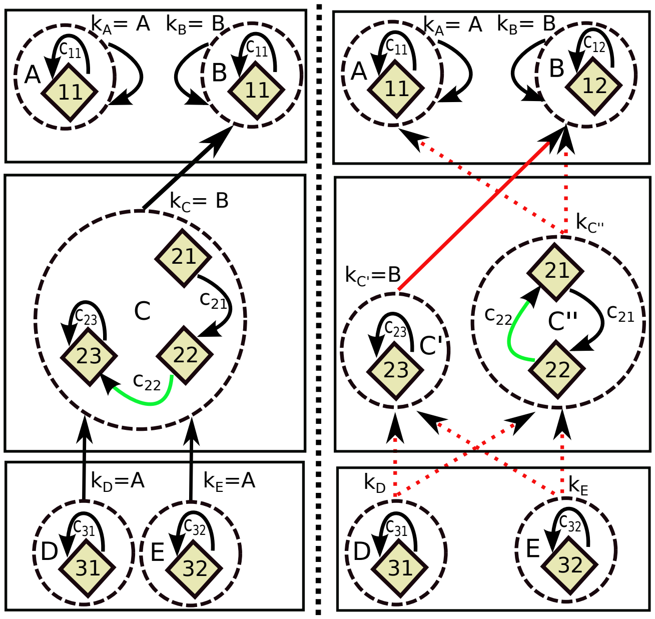

Figure 2: Illustration of changes induced by a data link proposal. Changing c_{22} (in the left configuration) splits cluster C into two clusters C' and C''. The cluster links associated with C (shown in red) must also be resampled. The MH step of the sampler proposes a joint configuration of the links $\{c_{22}, k_{C'}, k_{C''}, k_D, k_E\}$. The dashed red arrows illustrate the possible values the resampled cluster links could take. A single data link can create large changes to the partition structure, with local clusters splitting or merging, and groups of clusters shifting between components.

3.2 LINK PROPOSAL DISTRIBUTIONS

In samplers previously developed for the hCRP (Teh et al., 2006) and the naive-hddCRP (Ghosh et al., 2011), local clusters directly sample their global component memberships. However for the hddCRP, cluster links indirectly determine global component memberships. This complicates inference, as any change to the cluster structure necessitates coordinated changes to the cluster links. As illustrated in Figure 2, consider the case where a data link proposal causes a cluster to break into two components. The new cluster must sample a cluster (outgoing) link, and cluster links pointing to the old cluster (incoming links) must be divided among the newly split clusters. Thus, we use a MH proposal to jointly resample data and affected cluster links.

To simplify the exposition, we focus on a particular group g and denote c_{gi} as c_i. Let the current state of the sampler be $\mathbf{k}(\mathbf{c})$ and $\mathbf{c} = \{\mathbf{c}_{-i}, c_i = j\}$, so that i and j are members of the same cluster t_{ij}. Let $\mathcal{K}_{t_{ij}} = \{k_s \mid k_s = t_{ij}, s \neq t_{ij}\}$ denote the set of other clusters linking to t_{ij}.

Split? To construct our link proposal, we first set $c_i = i$. This may split current cluster t_{ij} into two new clusters, in which case we let t_i denote the cluster containing data i, and t_j the cluster containing formerly linked data j. Or, the partition structure may be unchanged so that $t_i = t_{ij}$.

Incoming links $k_s \in \mathcal{K}_{t_{ij}}$ to a split cluster are independently assigned to the new clusters with equal probability:

$$q_{\text{in}}(\mathcal{K}_{t_{ij}}) = \prod_{k_s \in \mathcal{K}_{t_{ij}}} \left(\frac{1}{2}\right)^{\delta(k_s,t_i)} \left(\frac{1}{2}\right)^{\delta(k_s,t_j)}. \tag{9}$$

The current outgoing link is retained by one of the split clusters, $k_{t_j} = k_{t_{ij}}$. To allow likelihood-based link proposals, we *temporarily* fix the other cluster link as $k_{t_i} = t_i$.

Propose Link We compare two proposals for c_i, the ddCRP prior distribution $q(c_i) = p(c_i \mid \alpha, A)$, and a data-dependent "pseudo-Gibbs" proposal distribution:

$$q(c_i) \propto p(c_i \mid \alpha, A)\Gamma(\mathbf{x}, \mathbf{z}(c_i, \mathbf{c}_{-i}, \mathbf{k})), \tag{10}$$

$$\Gamma(\mathbf{x}, \mathbf{z}(c_i, \mathbf{c}_{-i}, \mathbf{k})) = \begin{cases} \dfrac{p(\mathbf{x}_{\mathbf{z}=m_a} \cup \mathbf{x}_{\mathbf{z}=m_b} \mid \lambda)}{p(\mathbf{x}_{\mathbf{z}=m_a} \mid \lambda) p(\mathbf{x}_{\mathbf{z}=m_b} \mid \lambda)} & \text{if } c_i \text{ merges } m_a, m_b, \\ 1 & \text{otherwise.} \end{cases}$$

The prior proposal, although naïve, can perform reasonably when A is sparse. The pseudo-Gibbs proposal is more sophisticated, as data links are proposed conditioned on both the observations $\mathbf{x}$ and the current state of the sampler. Our experiments in Sec. 4 show it is much more effective.

Merge? Let $c_i = j^*$ denote the new data link sampled according to either the ddCRP prior or Eq. (10). Relative to the reference configuration in which $c_i = i$, this link may either leave the partition structure unchanged, or cause clusters t_i and t_{j^*} to merge into t_{ij^*}. In case of a merge, the new cluster retains the current outgoing link $k_{t_{ij^*}} = k_{t_{j^*}}$, and inherits the incoming links $\mathcal{K}_{t_{ij^*}} = \mathcal{K}_{t_i} \cup \mathcal{K}_{t_{j^*}}$.

If a merge does not occur, but t_{ij} was previously split into t_i and t_j, the outgoing link $k_{t_j} = k_{t_{ij}}$ is kept fixed. For newly created cluster t_i, we then propose a corresponding cluster link k_{t_i} from its full conditional distribution:

$$q_{\text{out}}(k_{t_i}) = p(k_{t_i} \mid \alpha_0, A^0(\mathbf{c}), \mathbf{x}, \mathbf{k}_{-t_i}, \mathbf{c}). \tag{11}$$

Note that the proposal $c_i = j^*$ may leave the original partition unchanged if $c_i = i$ does not cause t_{ij} to split, and $c_i = j^*$ does not result in a merge. In this case, the corresponding cluster links are also left unchanged.

Accept or Reject Combining the two pairs of cases above, our overall proposal distribution equals

$$q(\mathbf{c}^*, \mathbf{k}^* | \mathbf{c}, \mathbf{k}, \mathbf{x}) = \begin{cases} q(c_i^*) q_{\text{in}}(\mathcal{K}_{t_{ij}}^*) & \text{split, merge,} \\ q(c_i^*) & \text{no split, merge,} \\ q(c_i^*) q_{\text{out}}(k_{t_i}^*) q_{\text{in}}(\mathcal{K}_{t_{ij}}^*) & \text{split, no merge,} \\ p(c_i^* \mid \alpha, A) & \text{otherwise.} \end{cases}$$

Here, $\mathbf{c}^*$ and $\mathbf{k}^*$ denote the proposed values, which are then accepted or rejected according to the MH rule. For acceptance ratio derivations and further details, please see the supplemental material. After cycling through all data links $\mathbf{c}$, we use the Gibbs update of Eq. (11) to resample the cluster links $\mathbf{k}$, analogously to a standard ddCRP.

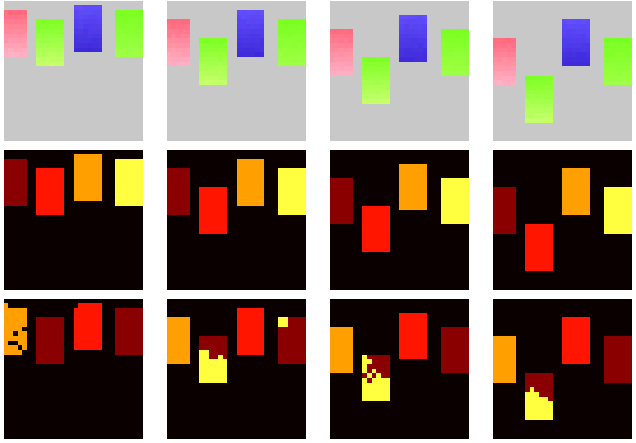

Figure 3: Experiments on synthetic data. *Top:* Ground truth partitions of a toy dataset containing four groups. Each group contains four objects exhibiting motion and color gradients. *Middle:* MAP partitions inferred by an hddCRP using size and optical flow-based cluster affinities. *Bottom:* MAP partitions discovered by a baseline hCRP using only color-based likelihoods.

4 EXPERIMENTS

In this section we present a series of experiments investigating the properties of the hddCRP model and our proposed MCMC inference algorithms. We examine a pair of challenging real-world tasks, video and discourse segmentation. We quantify performance by measuring agreement with held-out human annotations via the Rand index (Rand, 1971) and the WindowDiff metric (Pevzner & Hearst, 2002), demonstrating competitive performance.

To provide intuition, we first compare the hddCRP with the hCRP on a synthetic dataset (Figure 3) with four 30 × 30 frames (groups). Each frame contains four objects moving from top to bottom at different rates, and object appearances exhibit small color gradients. The hddCRP utilizes data link affinities that allow pixels (data instances) to connect to one of their eight spatial neighbors with equal probability. To exploit the differing motions of the objects, we define optical flow-based cluster affinities (Sun et al., 2010). Letting $w(t)$ denote the spatial positions occupied by cluster t after being warped by optical flow, and $w(s)$ the corresponding support of cluster s, the affinity is defined as $A^0_{ts}(\mathbf{c}) = (w(t) \cap w(s))/(w(t) \cup w(s)), t \neq s$, encouraging clusters to link to other clusters with similar spatial support. Given this affinity function, hddCRP was able to robustly disambiguate the four uniquely moving objects, while the hCRP produced noisy segmentations and consistently confused objects with local similarity but distinct motion.

4.1 VIDEO SEGMENTATION

Likelihood As a preprocessing step, we divide each frame into approximately 1200 superpixels using the

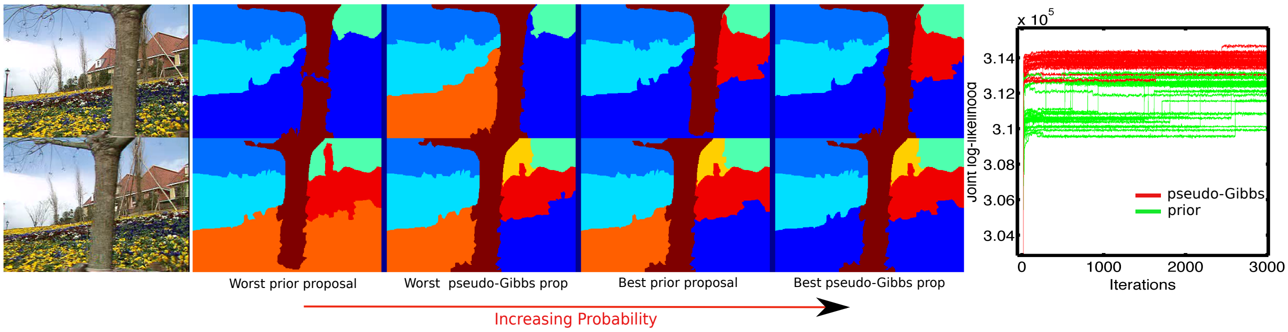

Figure 4: Data link proposal comparisons. *Left:* Two frames from the "garden" sequence, and partitions corresponding to the best and worst MAP samples using prior or pseudo-Gibbs proposals. *Right:* Joint log-likelihood trace plots for 25 trials of each proposal.

method proposed by Chang et al. (2013).[1] Each superpixel is described by L_2 unit-normalized 120-bin HSV color and 128-bin local texton histograms. Unit normalization projects raw histograms to the surface of a hypersphere, where we use *von-Mises Fisher* (vMF) distributions (Mardia & Jupp, 2009) shared across all clusters of a global component. In preliminary experiments, we found that the vMF produced more accurate segmentations than multinomial models of raw histograms; similar L_2 normalizations are useful for image retrieval (Arandjelović & Zisserman, 2012). We also extracted optical flow using the "Classic+NL" algorithm (Sun et al., 2010), and associated a two-dimensional flow vector to each super-pixel, the median flow of its constituent pixels.

The color, texture, and flow features for super-pixel i in video frame g are denoted by $x_{gi} = \{x^c_{gi}, x^t_{gi}, x^f_{gi}\}$, where

$$x^c_{gi} \sim \text{vMF}(\mu^c_{z_{gi}}, \kappa^c), \mu^c_{z_{gi}} \sim \text{vMF}(\mu^c_0, \kappa^c_0), \tag{12}$$

where κ^c, μ^c_0, and κ^c_0 are hyper-parameters controlling the concentration of color features around the direction $\mu^c_{z_{gi}}$, the mean color direction μ^c_0, and the concentration of $\mu^c_{z_{gi}}$ around μ^c_0. Texture features are generated similarly. Flow features are modeled via Gaussian distributions with conjugate, normal-inverse-Wishart priors:

$$\begin{aligned} x^f_{gi} \sim \mathcal{N}(\mu^{fg}_{z_{gi}}, \Sigma^{fg}_{z_{gi}}),\ \Sigma^{fg}_{z_{gi}} \sim \mathcal{IW}(n_0, S_0), \\ \mu^{fg}_{z_{gi}} \mid \Sigma^{fg}_{z_{gi}} \sim \mathcal{N}(\mu_0, \tau_0 \Sigma^{fg}_{z_{gi}}). \end{aligned} \tag{13}$$

Requiring all clusters in a global component, which may span several video frames, to share a single flow model is too restrictive. Instead we model the flow for each frame independently, requiring only that clusters in frame g assigned to the same component share a common flow model. Our model requires motion of a component to be locally (within a frame) coherent, but allows for large deviations between frames.[2] This assumption more closely reflects the motion statistics of objects in real videos.

Prior We used data affinities that encourage spatial neighbors not separated by strong intervening contours to connect to one another. We computed them by independently running the Pb edge detector (Martin et al., 2004) on each video frame and computing $A_{ij} = (1 - b_{ij})^3 \times \mathbf{1}[i, j]$ for each superpixel pair. Here, $0 \leq b_{ij} \leq 1$ is the maximum edge response along a straight line segment connecting the centers of superpixels i, j, and $\mathbf{1}[i, j]$ takes a value of 1 if i and j are spatial neighbors, and 0 otherwise.

Flow-based affinities, as in the earlier toy example, were used to specify the cluster affinity functions. All $\alpha_{1:G}$ and α_0 were set to 10^{-8}. The naive-hddCRP used identical data affinities and hyper-parameters, but used sequential distances between clusters (see Sec. 2.2). The hCRP used sequential affinities to govern both the data and cluster links. For a CRP, the expected number of clusters given N data points is roughly αlog(N). We set $\alpha_{1:G}$ such that the expected number of clusters in a video frame matches the number of observed ground truth clusters, and $\alpha_0 = 1$.

Data link proposals We compare the two data link proposals on 10 frames from the classic "garden" sequence. For each proposal, we ran 3000 iterations of 25 MCMC chains. The results, including MAP samples from the highest and lowest probability chains and log-likelihood trace plots, are summarized in Figure 4. The visualized MAP partitions demonstrate that all chains eventually reach reasonable configurations, but segmentations nevertheless improve qualitatively with increasing model likelihood. This suggests a correspondence between the biases captured by the hddCRP and the statistics of video partitions.

We find that pseudo-Gibbs proposals reach higher probability states more rapidly than prior proposals, and have much lower sensitivity to initialization. Overall, 24 of the 25 pseudo-Gibbs chains reach states that are more probable than the best prior proposal trial. Subsequent experiments thus focus solely on the superior pseudo-Gibbs proposal.

Empirical evaluation We compare our performance against a popular non-probabilistic *hierarchical graph-based video segmentation* (HGVS) algorithm (Grundmann et al., 2010), against the naive-hddCRP variant that was recently used for video co-segmentation (Chiu & Fritz,

[1] Chang et al. (2013) also estimate temporal correspondences between superpixels, but we do not utilize this information.

[2] See the supplement for specific hyper-parameter settings.

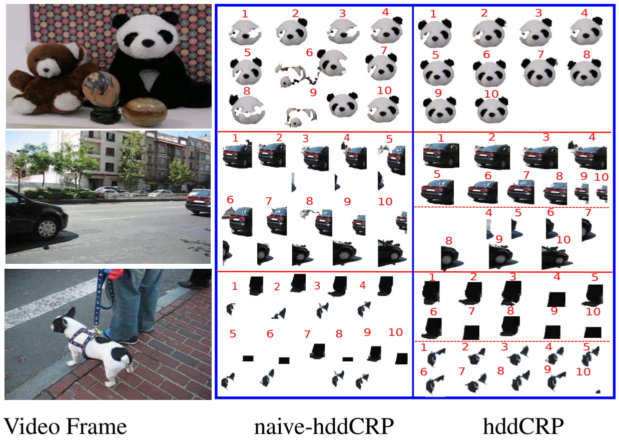

Figure 5: Video segments discovered by the naive-hddCRP and hddCRP. *Left to right:* A representative video frame, segments discovered by the naive-hddCRP and the corresponding segments discovered by hddCRP. Dashed red horizontal lines indicate different segments, and numbers indicate video frame numbers.

2013), and against the hCRP (Teh et al., 2006). For a controlled comparison, all three CRP models use identical likelihoods and hyperparameters. We use the MIT human annotated video dataset (Liu et al., 2008), which contains 9 human annotated videos, to quantitatively measure segmentation performance. We benchmark performance using the first 10 frames of each sequence.

Figure 6 summarizes this experiment. For HGVS the displayed segmentations were produced at 90 percent of the highest hierarchy level, which appears to produce the best visual and quantitative results. For the hddCRP variants, the segmentations correspond to the MAP sample of five MCMC chains, each run for 400 iterations[3]. We decided to run the samplers for 400 iterations based on the results shown in Figure 4, where a large majority of the pseudo-Gibbs chains converged within the first 300 iterations.

The Rand index was computed by treating the entire video sequence as one spatio-temporal block. This penalizes spatially coherent, but temporally inaccurate, segmentations that exhibit frequent "label switching" between frames. HGVS operates on pixels rather than superpixels and consequently produces finer-scale segmentations. However, these segmentations exhibit large segmentation errors (for instance, the neck and face regions get merged with the background in the second sequence). The hddCRP produces more coherent segmentations and in terms of Rand index, outperforms HGVS on all but one video sequence. The hddCRP also performs substantially better than the hCRP which ignores both superpixel and segment-level correlations; "bag of feature" assumptions are insufficient for this task. The gains over the naive-hddCRP appear to be more modest. However, a closer inspection (Figure 5) reveals that the hddCRP segments are visually cleaner and more coherent. Additionally, naive-hddCRP often falsely merges visually similar but distinctly moving objects together, while the hddCRP recognizes them as distinct segments. The videos in our dataset have large background regions with no significant motion. Both the hddCRP and the naive-hddCRP models tend to agree on such regions, while disagreeing on smaller foreground objects with distinct motions. Large regions dominate the Rand index, which explains the similar global performance by that metric.

4.2 DISCOURSE SEGMENTATION

Next we consider the problem of discourse segmentation. Given a collection of documents, the goal is to partition each document into a sequence of topically coherent non-overlapping discourse fragments. Previous work by Riedl & Biemann (2012) found that sharing information across documents tends to produce better segmentations, motivating the development of several text segmentation algorithms that exploit document relationships.

We conducted experiments on the *wikielements* dataset (Chen et al., 2009), which consists of 118 Wikipedia articles (at paragraph resolution) describing chemical elements. Although not explicitly made available in the dataset, each article corresponds to a chemical element that is characterized by its chemical properties and has a unique location in the periodic table. Our distance-dependent models are capable of exploiting this additional information to produce better discourse segmentations. As an illustration, consider the alternative problem of clustering articles. Figure 7 illustrates such a clustering where we leverage element properties by defining distances between documents as the Manhattan distance between corresponding element locations in the periodic table. The discovered clustering corresponds well with known element groupings. Discourse segmentation requires clustering the paragraphs describing documents, instead of the documents themselves. Nonetheless, we find that exploiting the periodic table location of each document's element leads to noticeable performance gains.

We compare two versions of the hddCRP to the naive-hddCRP and hCRP. To encourage topic contiguity, naive-hddCCRP and hddCRP allowed paragraphs to either link to themselves or to other paragraphs immediately preceding or succeeding them. We experimented with two affinity functions to capture the intuitions that similar documents tend to present similar topics in similar orders, and that clusters are more likely to be shared among articles about similar elements. The first function (hddCRP1) biased clusters of paragraphs to connect to those that occur at similar locations within other documents. Further, clusters were constrained to connect to only those that were contained in documents with lower atomic numbers. A second variant (hddCRP2) modeled distances between articles using the

[3] Roughly 6 hours on a 2.3 GHz intel core i7.

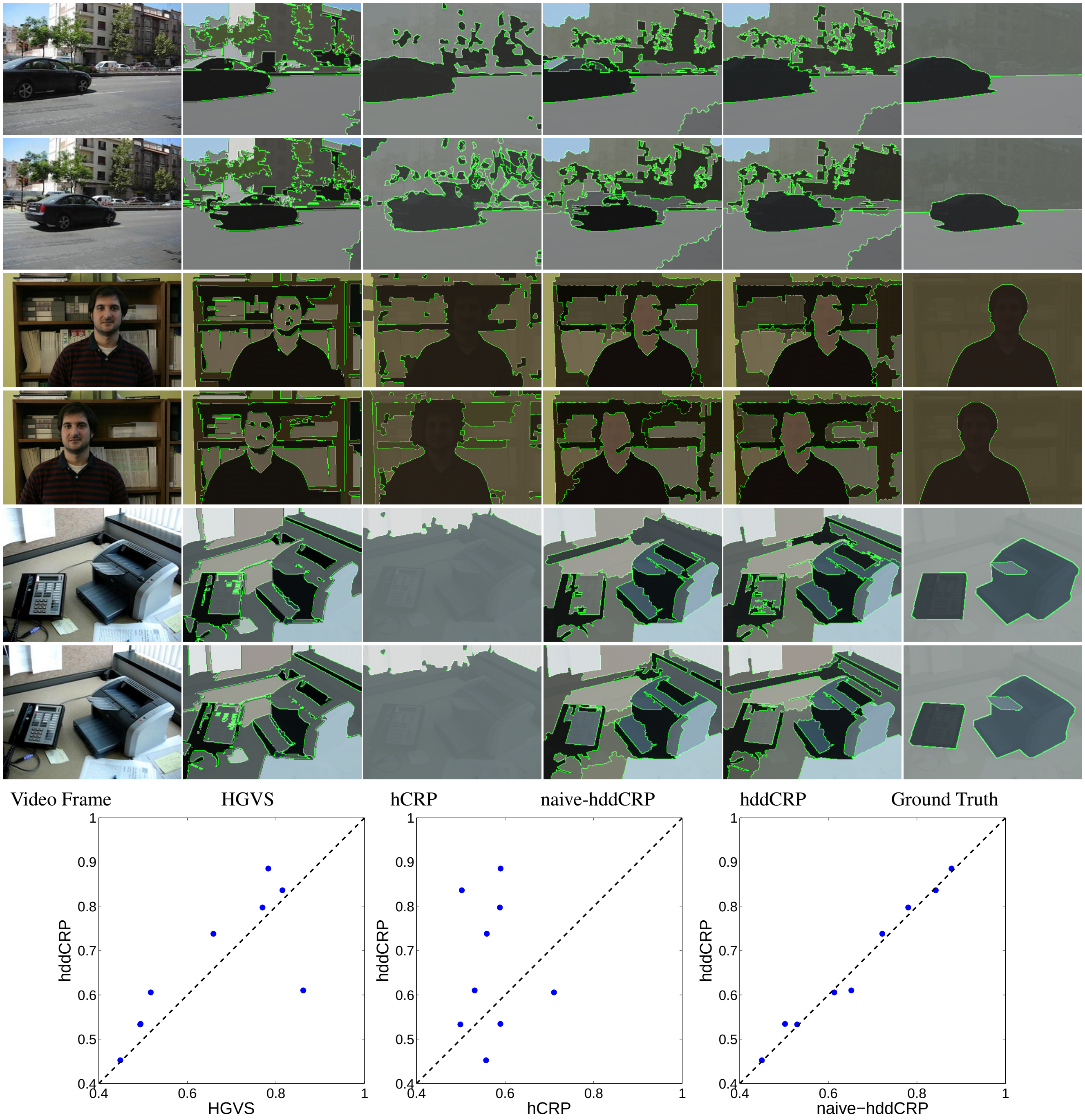

Figure 6: Video segmentation results. The top eight rows show the first and tenth frames of four videos from the MIT human annotated video dataset. *Left to right:* original video frames, segmentations produced by HGVS, hCRP, naive-hddCRP, and hddCRP, and the ground truth segmentations. *Bottom row:* Scatter plots comparing hddCRP, HGVS, naive-hddCRP, and hCRP in terms of Rand index achieved on all nine human annotated videos. Higher scores are better, and more points above the diagonal indicate favorable performance of hddCRP over competitors.

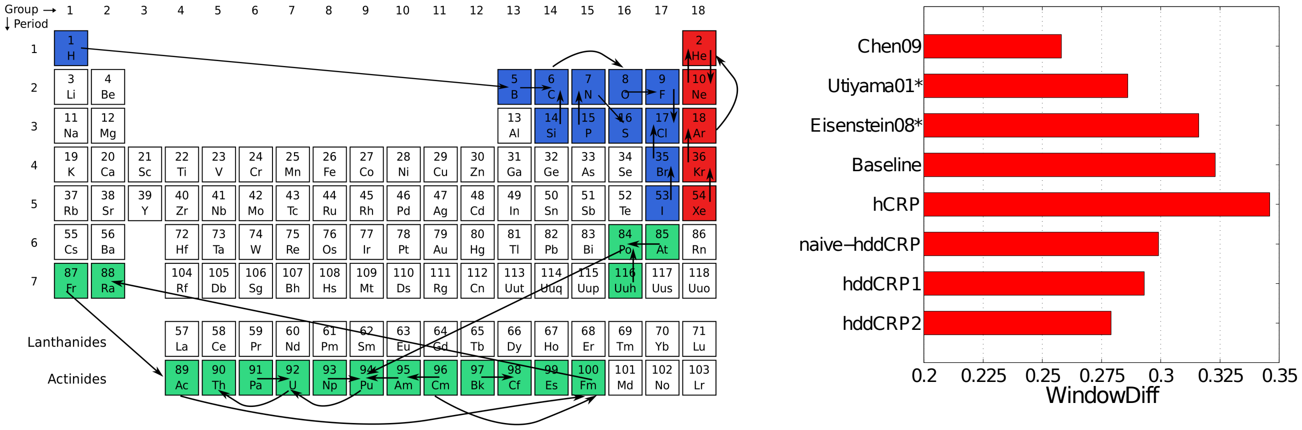

Figure 7: Discourse segmentation results on the *wikielements* dataset. *Left:* A partial visualization of the inferred customer links when clustering Wikipedia articles describing 118 chemical elements. The distance between articles equals the Manhattan distance between their locations in the periodic table. Only three of the nine discovered clusters have been visualized. *Right*: windowDiff scores achieved by competing methods, where lower scores indicate better performance. Asterisks indicate numbers reproduced from Chen et al. (2009).

Manhattan distance between the corresponding element locations in the periodic table, and defined cluster affinities as the logistic decay $f(d) = (1 + \exp(d))^{-1}$ of distances between their corresponding documents. In all cases, we model observed word counts using cluster-specific multinomial distributions with Dirichlet priors. The reported results correspond to the MAP sample of 5 MCMC chains, each run for 400 iterations.

We also compared against established text segmentation methods (Chen et al., 2009; Eisenstein & Barzilay, 2008; Utiyama & Isahara, 2001), and a naïve baseline that groups the entire dataset into one segment. We quantified performance using the windowDiff metric, which slides a window through the text incurring a penalty on discrepancies between the number of segmentation boundaries in the inferred segmentation and a gold standard segmentation. Figure 7 summarizes the performance of the competing models 7. Both hddCRP1 and hddCRP2 outperform naive-hddCRP and hCRP, showing that our cluster-level affinities capture important additional dataset structure. The hddCRP2 model was superior to all other hddCRP variants, as well as to the specialized text segmentation algorithms of Eisenstein & Barzilay (2008) and Utiyama & Isahara (2001). However, the generalized Mallows model (Chen et al., 2009) achieved the best performance; it is able to both globally bias segment orderings to be similar across related documents, while guaranteeing spatially connected topics. In contrast, the hddCRP weakly constrains segment order through local cluster affinities and while it encourages contiguity, the likelihood may prefer disconnected segments, resulting in a poorer match with the reference segmentation. We nevertheless find it encouraging that the general hddCRP framework, with appropriate affinities, is competitive with specialized text segmentation methods.

5 DISCUSSION AND FUTURE WORK

We have developed a versatile probabilistic model for clustering groups of data with complex structure. Applying it to diverse domains is straightforward: one need only specify appropriate distance functions. Our hierarchical ddCRP defines a valid joint probability distribution for any choice of “distances”, which need not be metrics or have any special properties. Using distances based on pixel locations and optical flow estimates, the hddCRP compares favorably to contemporary video segmentation methods. Using distances based on paragraph order and element positions in the periodic table, it outperforms several established textual discourse segmentation techniques.

While our MCMC inference methods are highly effective for moderate-sized datasets, further innovations will be needed for computational scaling to very large datasets. In cases where training examples of appropriate clusterings are available, we would also like to automatically learn effective hddCRP distance functions.

Acknowledgements

This research supported in part by ONR Award No. N00014-13-1-0644. Portions of this work were completed while S. Ghosh and M. Raptis were at Disney Research, Pittsburgh, PA, USA.

References

Arandjelović, R. and Zisserman, A. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

Blei, D. M. and Frazier, P. I. Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.*, 12:2461–2488, November 2011.

Chang, J., Wei, D., and Fisher III, J. W. A Video Representation Using Temporal Superpixels. In *CVPR*, 2013.

Chen, H., Branavan, S. R. K., Barzilay, R., and Karger, D. R. Content modeling using latent permutations. *J. Artif. Intell. Res. (JAIR)*, 36:129–163, 2009.

Chiu, W. and Fritz, M. Multi-class video co-segmentation with a generative multi-video model. In *CVPR*, 2013.

Duan, J. A., Guindani, M., and Gelfand, A. E. Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825, 2007.

Dunson, D. B. and Park, J. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

Eisenstein, J. and Barzilay, R. Bayesian unsupervised topic segmentation. In *EMNLP*, pp. 334–343, 2008.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B., and Blei, D. Spatial distance dependent Chinese restaurant processes for image segmentation. In *NIPS*, pp. 1476–1484, 2011.

Ghosh, S., Sudderth, E. B., Loper, M., and Black, M. J. From deformations to parts: Motion-based segmentation of 3D objects. In *NIPS*, pp. 2006–2014, 2012.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Griffin, J. E. and Steel, M. F. J. Order-based dependent Dirichlet processes. *JASA*, 101(473):179–194, March 2006.

Grundmann, M., Kwatra, V., Han, M., and Essa, I. Efficient hierarchical graph based video segmentation. In *CVPR*, 2010.

Haines, T. S. F. and Xiang, T. Background subtraction with Dirichlet processes. In *ECCV*, pp. 99–113. Springer, 2012.

Jain, S. and Neal, R. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. of Comput. and Graph. Stats*, 13:158–182, 2004.

Kim, D. and Oh, A. Accounting for data dependencies within a hierarchical Dirichlet process mixture model. In *CIKM*, pp. 873–878, 2011.

Lin, D., Grimson, E., and Fisher III, J. W. Construction of dependent Dirichlet processes based on Poisson processes. In *NIPS*, 2010.

Liu, C., Freeman, W. T., Adelson, E. H., and Weiss, Y. Human-assisted motion annotation. In *CVPR*, 2008.

MacEachern, S. N. Dependent nonparametric processes. In *Proc. Section on Bayesian Statistical Science*, pp. 50–55. American Statistical Association, 1999.

Mardia, K. V. and Jupp, P. E. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

Martin, D. R., Fowlkes, C.C., and Malik, J. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. PAMI*, 26(5):530–549, 2004.

Orbanz, P. and Buhmann, J. M. Nonparametric Bayesian image segmentation. *IJCV*, 77:25–45, 2008.

Pevzner, L. and Hearst, M. A. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, March 2002.

Pitman, J. Combinatorial stochastic processes. Technical Report 621, U.C. Berkeley Department of Statistics, August 2002.

Rand, W. M. Objective criteria for the evaluation of clustering methods. *JASA*, 66(336):846–850, 1971.

Rao, V. A. and Teh, Y. W. Spatial normalized gamma processes. In *NIPS*, pp. 1554–1562, 2009.

Riedl, M. and Biemann, C. How text segmentation algorithms gain from topic models. In *HLT-NAACL*, pp. 553–557, 2012.

Socher, R., Maas, A., and Manning, C. D. Spectral Chinese restaurant processes: Nonparametric clustering based on similarities. In *AISTATS*, 2011.

Sudderth, E. B. and Jordan, M. I. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *NIPS*, pp. 1585–1592, 2008.

Sudderth, E. B., Torralba, A., Freeman, W. T., and Willsky, A. S. Describing visual scenes using transformed objects and parts. *IJCV*, 77:291–330, 2008.

Sun, D., Roth, S., and Black, M. J. Secrets of optical flow estimation and their principles. In *CVPR*, pp. 2432–2439. IEEE, June 2010.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of American Statistical Association*, 25(2):1566–1581, 2006.

Utiyama, M. and Isahara, H. A statistical model for domain-independent text segmentation. In *ACL*, pp. 499–506, 2001.

Wang, C. and Blei, D. M. A split-merge MCMC algorithm for the hierarchical Dirichlet process. *arXiv preprint arXiv:1201.1657*, January 2012.