Nonparametric Discovery of Activity Patterns from Video Collections

Michael C. Hughes and Erik B. Sudderth
Department of Computer Science, Brown University, Providence, RI, USA
mhughes@cs.brown.edu, sudderth@cs.brown.edu

Abstract

We propose a nonparametric framework based on the beta process for discovering temporal patterns within a heterogenous video collection. Starting from quantized local motion descriptors, we describe the long-range temporal dynamics of each video via transitions between a set of dynamical behaviors. Bayesian nonparametric statistical methods allow the number of such behaviors and the subset exhibited by each video to be learned without supervision. We extend the earlier beta process HMM in two ways: adding data-driven MCMC moves to improve inference on realistic datasets and allowing global sharing of behavior transition parameters. We illustrate discovery of intuitive and useful dynamical structure, at various temporal scales, from videos of simple exercises, recipe preparation, and Olympic sports. Segmentation and retrieval experiments show the benefits of our nonparametric approach.

1. Introduction

We consider the problem of grouping similar short-term motions, such as jogging or grating cheese, that might occur repeatedly in a video collection. Our goal is to obtain a temporal segmentation of each video into coherent *behaviors*, and identify how these behaviors are reused across videos and through time. These behaviors provide a step towards the perceptual organization of large video collections, by identifying segments of video arising from similar physical causes. We adapt a Bayesian nonparametric model of sequential data, the *beta process hidden Markov model* (BP-HMM) [4], to allow completely unsupervised activity discovery. We need not predefine the relevant behaviors or even their number, as both are learned directly from data.

The BP-HMM defines an unbounded global library of behaviors and describes each sequence with a sparse subset of these behaviors. Sequences are generated by Markovian dynamics which allocate a single behavior state to each timestep. In our video analysis application, these behavior states then generate vector-quantized motion descriptors. By clustering motion codewords, our behaviors describe activity at a coarser temporal scale than typical spatio-temporal features. We also extend the BP-HMM's dynamical model and develop improved learning algorithms.

1.1. Previous Work

Video understanding, and in particular activity recognition, is a widely studied area [1]. Many contemporary approaches begin by extracting descriptors of local spatio-temporal interest points, which are then vector quantized into a "bag of features" (BoF) [13]. While this holistic representation has proven useful for activity recognition due to its robustness and efficiency, it does not capture temporal information crucial for distinguishing complicated actions (e.g. the long jump and triple jump). Simple extensions have built independent BoF models for each segment in some fixed, coarse temporal segmentation [9], but cannot adequately describe more complex and variable chronologies. Other work has adapted probabilistic topic models by associating each activity category with a unique latent topic [23]. However, this rigid structure cannot learn behaviors shared across categories or model details which distinguish subtypes of the same category.

Among approaches that try to model chronology, many presume external, expert knowledge of the activity domain, either by specifying the action semantics [10] or predefining motion templates for every possible action [11]. Linear dynamical systems have been used for unsupervised temporal learning [20], but without notions of discrete behaviors or shared structure among multiple videos. More recently, Niebles *et al*. [12] proposed a discriminative recognition framework that builds a set of bag-of-words classifiers for each action type, each with an associated temporal range. This approach presumes all videos in a category have similar durations and temporal patterns. Hoai *et al*. [7] present a discriminative model for segmentation and classification that is more robust, but this requires a training set of preselected activities with ground-truth temporal labels.

1.2. Contributions

As one of the first applications of Bayesian nonparametrics to video analysis, we make several important contri-

butions. First and foremost, we improve *unsupervised* recovery of activity patterns. We can learn detailed temporal structure at multiple scales, from repetitive short-term dynamics (e.g., handwaving, as in Fig. 3) to the more structured patterns of sporting events (e.g., the gymnast's vault routine in Fig. 7). Via Bayesian nonparametric priors, such learning is possible without requiring detailed manual model design or dataset-specific tuning. Unlike discriminative classifiers, the dynamical behaviors inferred by the BP-HMM can be used for multiple purposes; we demonstrate visualization of the shared dynamical structure of video collections and retrieval of related sequences.

Additionally, we introduce novel *data-driven* moves which improve reversible jump MCMC posterior inference algorithms. Previous work [4] employed simple feature creation proposals, and their experiments with motion capture data consider at most 6 sequences. Our novel data-driven proposals allow efficient inference with hundreds of videos.

We begin in Sec. 2 by describing the BP-HMM model and the underlying video representation. Sec. 3 then derives MCMC methods for learning and inference, with a focus on our data-driven proposals. Sec. 4 demonstrates activity discovery on three datasets: KTH exercises [15], CMU kitchen activities [3], and Olympic sports [12].

2. Beta Processes for Video Analysis

Here, we describe our video representation (Sec. 2.1), and then review existing Bayesian nonparametric binary featural models (Sec. 2.2) and the BP-HMM (Sec. 2.3). We then extend the BP-HMM to allow behavior dynamics parameters to be shared across sequences, and discuss related nonparametric models (Sec. 2.4).

2.1. Sparse Representation of Video Sequences

Following several recent papers, we use *spatio-temporal interest points* (STIPs) to compactly describe video sequences. We use existing STIP code [9] to detect interest points and obtain *histogram of gradients* (HOG) and *histogram of optical flow* (HOF) descriptors. Separately for each dataset, we build a codebook with $V = 1000$ codewords using the K-means algorithm. Each STIP is then mapped to the nearest codeword, providing a standard "bag of words" representation [21].

To represent videos as discrete time series, we choose a temporal bin-width w (in seconds for invariance to frame rate), divide video i into T_i bins of width w, and count the number of occurrences of each codeword across all STIPs within each bin. The parameter w indirectly influences the time-scale of the learned dynamics.

2.2. Bayesian Nonparametric Featural Models

Feature-based representations provide intuitive descriptions of the high-level actions found in any video corpus. We assume there exists a global set of possible atomic actions, which we will call behaviors or *features*.[1] Each feature is characterized by a distribution on the set of STIP codewords, and hence captures a particular pattern of short-term movements. We posit that semantically meaningful, long-term *activities* can be understood as compositions of these features. Each video sequence in the corpus exhibits a sparse subset of the global features: a clip might contain running and jumping, but neither diving nor lifting.

Each video "object" in the corpus is associated with a sparse binary vector $f_i = [f_{i1}, f_{i2}, \ldots]$ indicating the presence or absence of each feature in the unbounded global collection. Corpus-wide behavior assignments are denoted by F, a binary matrix whose i-th row is f_i. Feature k has an associated corpus frequency b_k and STIP distribution parameterized by θ_k. These global variables are generated by an underlying stochastic process, the *beta process*:

$$B \mid B_0, \gamma, \beta \sim \mathrm{BP}(\beta, \gamma B_0), \quad B = \sum_{k=1}^{\infty} b_k \delta_{\theta_k}. \quad (1)$$

Here $\theta_k \sim B_0$, and the unbounded collection of feature weights b_k is determined by an underlying Poisson process [18]. The binary feature vector for object i is then determined by independent Bernoulli draws $f_{ik} \sim \mathrm{Ber}(b_k)$. Marginalizing over B, the total number of active features in object i has distribution $\mathrm{Poisson}(\gamma)$ determined by the *mass parameter* γ. The *concentration parameter* β controls the degree to which features are shared between objects.

Thibaux and Jordan [18] show that marginalizing B from this construction leads to an exchangeable prediction rule for f_i known as the *Indian buffet process* (IBP) [6]. In this analogy, objects (videos) are customers, and features (behaviors) are dishes in a buffet. The first customer (video) samples $\mathrm{Poisson}(\gamma)$ unique dishes (behaviors). Successive customer i chooses previously sampled dish k with probability $\frac{m_k}{i}$ proportional to the number of previous videos m_k exhibiting it, and also samples $\mathrm{Poisson}(\frac{\gamma}{i})$ new behaviors. This IBP representation is useful for MCMC inference.

2.3. Beta Process Hidden Markov Models

To model a collection of video sequences via partially shared dynamical behaviors, we begin with the BP-HMM [4] shown in Fig. 1. As above we define binary features f_i indicating the behaviors observed in video sequence i, which are coupled by a global feature distribution $B \sim \mathrm{BP}(\beta, \gamma B_0)$. To model discrete STIP encodings, we associate each feature k with a multinomial distribution θ_k on the V possible codewords. A natural conjugate prior here is a symmetric V-dimensional Dirichlet with mass λ_θ:

$$\theta_k \mid B_0 \sim \mathrm{Dirichlet}(\lambda_\theta, \lambda_\theta, \ldots \lambda_\theta) \quad (2)$$

[1] This terminology comes from the machine learning literature on latent feature models [6]. Our learned features should not be confused with the so-called "visual features" generated by bottom-up interest point detectors.

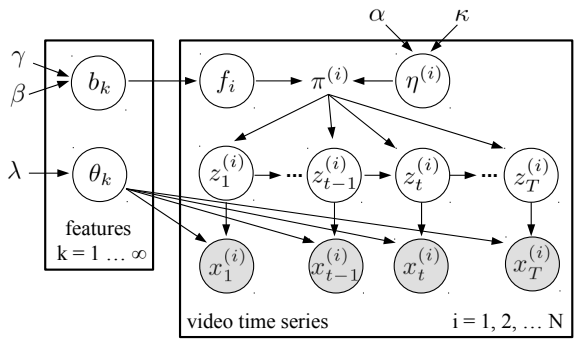

Figure 1. The BP-HMM as a directed graphical model. Binary features f_i determine the set of possible temporal states $z_t^{(i)}$ in sequence i, which in turn generate observed STIPs $x_t^{(i)}$. We illustrate sequence-specific dynamics $\eta^{(i)}$. Sec. 2.3 describes an alternative model which shares dynamics across all videos.

We consider two different approaches for coupling these emission parameters with Markov state dynamics.

Our baseline model, proposed by Fox *et al.* [4], associates *independent* transition dynamics with each video. In particular, the transition distribution $\pi_j^{(i)}$ from each state j for the HMM of video i is obtained by drawing a set of individual transition weights $\eta^{(i)}$, and then normalizing these according to the feature assignments f_i as follows:

$$\eta_{jk}^{(i)} \sim \text{Gam}(\alpha + \kappa\delta_{j,k}, 1), \quad \pi_j^{(i)} = \frac{\eta_j^{(i)} \circ f_i}{\sum_k f_{ik}\eta_{jk}^{(i)}} \tag{3}$$

Here, $\delta_{j,k} = 1$ if $j = k$, and 0 otherwise. The element-wise vector product, denoted by $\circ$, assigns positive transition probability $\pi_{jk}^{(i)}$ only to features k active in f_i. The *sticky* parameter κ places extra expected mass on self-transitions [5], encouraging the model to learn state sequences with the temporal persistence of real activities.

The transition matrix $\pi^{(i)}$ and emission distributions θ fully parameterize the HMM which generates the observed STIPs. For each timestep t, we draw its feature assignment $z_t^{(i)} \in \{k \mid f_{ik} = 1\}$ according to $z_t^{(i)} \sim \pi_{z_{t-1}^{(i)}}^{(i)}$. The L_t spatio-temporal codewords in bin t, whose histogram we denote by $x_t^{(i)}$, are then emitted according to

$$x_t^{(i)} \sim \text{Multinomial}(\theta_{z_t^{(i)}}, L_t) \tag{4}$$

The number of emissions L_t can vary with time, but we assume that L_t is *independent* of the current state $z_t^{(i)}$.

While the preceding prior on transition dynamics is flexible, in situations where behavior transitions across many videos are very similar we expect that sharing transition weights across all videos will be more appropriate. We thus also consider the following, alternative prior:

$$\eta_{jk}^{(0)} \sim \text{Gam}(\alpha + \kappa\delta_{j,k}, 1), \quad \pi_j^{(i)} = \frac{\eta_j^{(0)} \circ f_i}{\sum_k f_{ik}\eta_{jk}^{(0)}} \tag{5}$$

Here, a single *common* set of weights $\eta^{(0)}$ is normalized by sequence-specific feature activations. This sharing of transition information across sequences can provide noticeable gains on real data (Fig. 3). Note that there can still be significant variability across multiple state sequences $z^{(i)}$ sampled from common Markov dynamics.

2.4. Related Work

There have been few applications of Bayesian nonparametric models to video analysis. For general nonparametric modeling of sequential data, alternatives to the BP-HMM include the earlier infinite HMM [2] and the hierarchical Dirichlet process (HDP) HMM [17]. The HDP-HMM has been used in detect unusual events in video sequences [14]. For problems of far-field surveillance from static cameras, the HDP has been used to model interactions among simple activities [22]. Later work proposed a dependent Dirichlet process HMM that uncovers temporal rules for traffic motion patterns in a single scene [8].

Our model is novel in its emphasis on understanding *collections* of videos rather than individual clips, and in its featural representation of behavior-video relationships via the beta process. Hierarchical Dirichlet process models would force all videos to have positive probability of displaying all behaviors, but the beta process elegantly allows a video to contain only a sparse subset of relevant behaviors. The beta process has been used for image denoising [24], but has not yet been used to model temporal video sequences.

3. Learning via MCMC

Due to the complex combinatorial structure of the BP-HMM, we employ Markov chain Monte Carlo (MCMC) methods for learning and inference. We base our algorithms on the exact MCMC procedure proposed by Fox *et al.* [4], whose collapsed sampler marginalizes over feature inclusion parameters $\mathbf{b}$ and state assignments $\mathbf{z}$. Conditional updates to the feature matrix F, emission distributions θ_k, and transition weights η proceed in an iterative fashion.

3.1. Resampling Feature Indicator Variables

We proceed sequentially through the time series i and sample their features f_i in two stages: features shared by some other time series, and features unique to time series i. For shared features f_{ik}, we propose flipping their binary values one at a time, and accept or reject according to the Metropolis-Hastings rule [4].

For sequence-specific features, Fox *et al.* [4] used reversible jump MCMC to define a pair of feature birth and death moves. While this approach elegantly avoids the need to approximate the infinite BP-HMM, their birth proposals use the (typically vague) prior to propose parameters θ_{k^*} for new features k^*. Such proposals rarely explain the high-dimensional observed codewords found in realistic data, re-

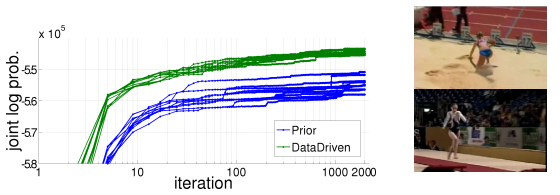

Figure 2. Comparison of different proposal distributions for the birth move of reversible jump MCMC, using vault and triple jump videos. *Left:* Log joint probability versus iteration, across 10 runs for each method. The data-driven proposals quickly reach higher probability posterior configurations. *Right:* Example key frames. All 10 prior runs assigned (a) and (b) to the same behavior due to poor exploration of the feature space, while 5 data-driven runs discover different behaviors for these intuitively distinct activities.

sulting in low acceptance rates and slow exploration. We instead consider a *data-driven* proposal distribution [19], based on the posterior of θ_{k^*} given data in a randomly chosen subwindow W of the current sequence:

$$\theta_{k^*}|W \sim \text{Dir}(C_1 + \lambda_\theta, C_2 + \lambda_\theta, \dots C_V + \lambda_\theta) \quad (6)$$

Here, C_v counts occurrences of codeword v in window W.

To validate this contribution, we compare data-driven (DD) and Prior birth proposals in Fig. 2. Here, we apply the BP-HMM to a collection of 20 vault and 20 triple jump sequences. We start from a poor configuration of just 10 states, and run 10 chains for each proposal. In less than 50 iterations, all DD runs achieve high probabilities that are superior to all Prior runs even after 2000 iterations. Each DD run recovers about 50 total behaviors, while each Prior run finds only 25. To illustrate qualitative benefits, we consider the behaviors assigned to fragments (5 timesteps, 0.8 seconds) from two videos: (a) standing up after a triple jump, and (b) running towards a vault (see Fig. 2). These clips show very different activities, but share some STIP codewords in common. We measure how often the dominant behavior in (b) is re-used within clip (a). All 10 Prior runs assign (b)'s dominant behavior to 2-3 timesteps in clip (a). DD proposals show a clear improvement, with 5 runs sharing *no* behaviors and the remaining 5 using (b)'s majority behavior in just 1 of (a)'s 5 timesteps.

3.2. Resampling HMM Parameters

Given fixed F, θ, and η, we can draw state sequences $\mathbf{z}$ as auxiliary variables using dynamic programming [16]. Each θ_k can then be sampled from a conjugate posterior. Similar closed-form updates exist for sequence-specific transition parameters $\eta^{(i)}$, though we note importantly that the update equations given in [4] are slightly incorrect. The correct posterior for $\eta^{(i)}$, up to a normalization constant, equals

$$p(\eta_{jk}^{(i)} \mid \mathbf{z}_i, f_{ik} = 1) \propto \frac{(\eta_{j,k}^{(i)})^{N_{jk}^{(i)}+\alpha+\delta_{j,k}\kappa-1} e^{-\eta_{jk}^{(i)}}}{\left[\sum_\ell f_{i\ell}\eta_{j\ell}^{(i)}\right]^{N_{j\cdot}^{(i)}}} \quad (7)$$

where $N_{jk}^{(i)}$ counts the number of transitions from state j to k in sequence $z^{(i)}$, and $N_{j\cdot}^{(i)} = \sum_k f_{ik} N_{jk}^{(i)}$. Draws from this posterior can be obtained by sampling a vector $\text{Dir}(\dots, N_{jk}^{(i)} + \alpha + \delta_{j,k}\kappa, \dots)$, and then scaling it by a gamma random variable with mass determined by the prior distributions on the active features (see supplement).

Updates for globally-shared $\eta^{(0)}$ have similar posteriors, but require terms from all sequences that make closed-form posterior draws intractable. We use Metropolis-Hastings updates based on a gamma random walk proposal with mean equal to the current value, and variance 10.

4. Experiments

We investigate the capabilities of the BP-HMM on three video datasets. Our goals here are twofold. First, we illustrate that our approach recovers qualitatively useful hidden structure. Second, we quantitatively demonstrate that the BP-HMM provides useful representations for activity segmentation and retrieval tasks.

For all experiments, we run MCMC inference for at least 2000 iterations and fix model parameters $\hat{\theta}, \hat{z}, \hat{\eta}, \hat{F}$ to the final sample. For all datasets except KTH, we use sequence-specific dynamics, as global sharing is likely not beneficial when temporal variability is significant. We use a fixed set of hyperparameters for all datasets, with transition weights $\alpha = 2$ and $\kappa = 10\alpha$, BP mass $\gamma = 2$, and BP concentration $\beta_0 = 1$ (as in the conventional IBP [6]). We encourage moderately sparse emission distributions via $\lambda_\theta = 0.75$.

4.1. KTH Exercise Dataset

The KTH actions dataset contains simple exercises performed by 25 actors. Rather than conventional supervised activity recognition, we explore two unsupervised analyses of this data. We first compare different dynamics sharing schemes, and later demonstrate recovery of meaningful temporal segmentations. For both tasks, we use only HOF descriptors and set the bin width w for the video time series to 0.08 seconds (2 frames) to capture intricate motion.

Due to the strong similarity in temporal structure between videos, we intuitively expect that a model with globally shared dynamics will perform well. Our first experiment studies qualitatively what benefits this change brings. We train two BP-HMM models on all 378 training videos in the categories `clap`, `wave`, and `jog`, one using global dynamics and the other sequence-specific. Fig. 3 compares each model's MAP estimate for $z^{(i)}$ for several example sequences. For both `clap` and `wave`, global sharing increases the level of detail (number of behaviors) found in a typical sequence, while also producing more consistent segmentations across videos.

Next, we evaluate the BP-HMM's ability to recover meaningful segmentations. We construct a dataset of 12

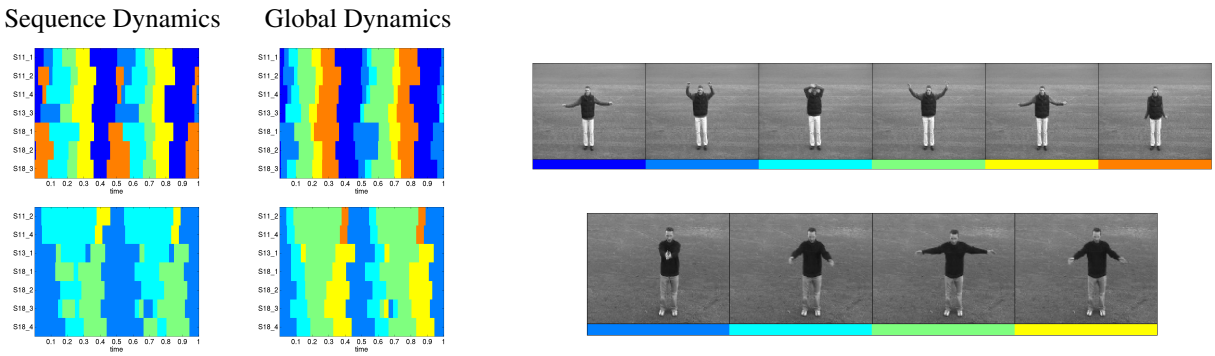

Figure 3. Qualitative results for `wave` (top) and `clap` (bottom) videos from KTH. Colors indicate distinct behaviors. *Left:* Behavior segmentations for example sequences from actors 11, 13, and 18. Each row in the image represents a single video clip, labeled at left by actor and trial. Each sequence was aligned by hand to show two complete cycles of the wave/clap action. Far left uses sequence-specific dynamics, near left uses globally shared dynamics (Sec. 2.3). Sharing dynamics across all sequences yields more detailed segmentations (6 phases for wave, 4 for clap) that are also more consistent across videos. *Right:* Key frames with behavior annotations, using global model.

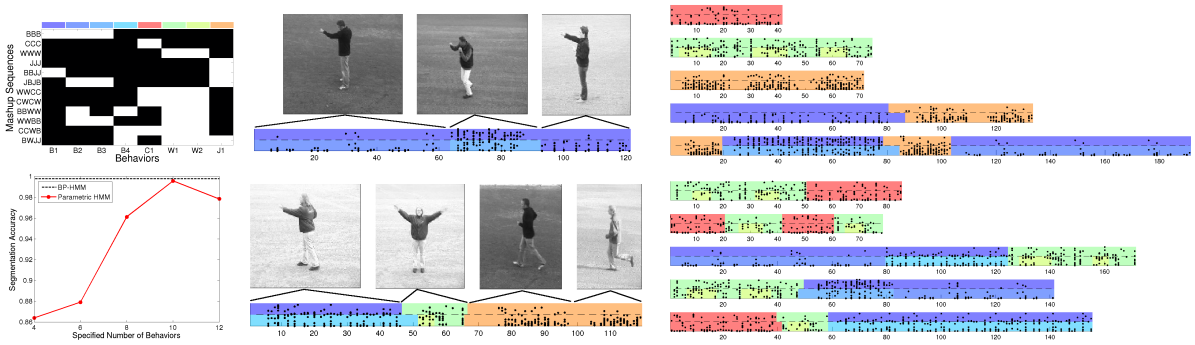

Figure 4. Results on KTH mashup sequences. *Top Left:* Binary feature matrix recovered by BP-HMM. White indicates presence, black absence. Each row shows behaviors assigned to a single mashup sequence. Each sequence is assigned a text label based on the categories of its source clips (B=box, C=clap, W=wave, J=jog). Each behavior is assigned to a true category by state sequence proximity (indicated by text label on horizontal axis). *Bottom Left:* Segmentation accuracy. Scores computed for run with highest joint probability among 5 initializations of MCMC. *Middle:* Estimated state sequences for BBB (top) and BWJJ (bottom) mashups, shown with key frames from each source clip. Distinct boxing behaviors correspond to clips taken from different actors. Colors above dotted line indicate true categories at each timestep, colors below indicate recovered behaviors. Multiple behaviors can map to one true category (indicated by color similarity), so mismatch does not necessarily indicate error. Black dots indicate STIP codewords observed at each timestep: each location along vertical axis marks presence of a unique codeword (vocab size=1000). *Right:* Ground truth and estimated behaviors for other 10 mashups.

sequences which are each "mashups" of 3-4 KTH clips, inspired by [7]. Each mashup concatenates clips at random from `box`, `clap`, `wave`, and `jog` actions, with variability in the actions present across sequences. This `KTHMashup` dataset, illustrated in Fig. 4, allows us to define a ground-truth action category label to each timestep, and thus quantify the accuracy of a BP-HMM segmentation.

The BP-HMM discovers 9 behaviors, whose binary presence in each mashup is depicted in Fig. 4. We observe that 4 of these 9 correspond to boxing activities, with different states allocated to each of the four actors whose clips generated the data. Each actor has a slightly different style of boxing (moving feet, punch out vs. up, etc.), which produces different motions and thus distinct codeword distributions. The model cleanly finds a single state for both clapping and jogging, and splits the waving clips into two phases (arms up, arms down). Note that with only 12 sequences, we expect the level of detail in wave segmentation to be lower than that recovered using 120 sequences in the first experiment. This showcases the power of the nonparametric approach to adapt to the available data.

To compare the BP-HMM to a conventional parametric HMM, we compute segmentation accuracy by first mapping each estimated state to its closest true label, and then computing the number of timesteps where this relabeled estimate matches ground truth across all 12 `KTHMashup` sequences. Fig. 4 shows that our BP-HMM achieves better segmentation accuracy than any HMM model found in a search over a reasonable range of the number of hidden states. Our nonparametric approach, which automatically explores the number of total behaviors as well as the sparse subset available to each sequence, is clearly beneficial.

4.2. CMU Kitchen Dataset

We next apply the BP-HMM model to videos from the CMU Multi-Modal Activity Database [3]. Each video is several minutes long and depicts a single actor in the same kitchen cooking a prescribed dish from start to finish. We chose 3 distinct recipes (`Sandwich`, `Pizza`, and `Brownie`) and downloaded 10 training videos for each recipe. Although simple in the dimensions of scene and object variability, the activities in these videos are complex over time. We evaluate via quantitative retrieval performance, as well as unsupervised exploration of the latent behaviors discovered by our model of these sequences. We set the window size w to be 0.5 seconds (15 frames), to capture the coarser scale behaviors of these longer videos.

4.2.1 Retrieval Evaluation

Our first goal is to determine if the hidden structure found by BP-HMM can generalize to novel videos. We propose a retrieval task: rank videos in a held-out test set by similarity to a query video from training. We take recipe labels to be ground truth, which means that `Brownie` videos should be judged more similar to other `Brownie` videos than any other recipe. We train our model on the 30 training videos, and then estimate a state sequence z for each of 30 test videos (10 per recipe). We summarize each video i by a histogram ϕ_i indicating how many timesteps are assigned to each behavior. We compute similarity between videos (using either the baseline BoF codeword histogram or our BP-HMM behavior histogram ϕ_i) via the χ^2 kernel [21].

After computing rankings for each query video in training, we obtain the class-specific precision-recall curves of Fig. 5. At all values of recall, our BPHMM representation provides the same or better precision compared to bag-of-features. Overall, we find BP-HMM's F-score to be 0.804, which compares favorably to 0.703 for BoF. As a further test, we compare to the rigid temporal discretization of BoF proposed by [9], which given 2 bins builds a separate BoF histogram for the first half and second half of each video. We consider both 2 bins and 3 bins, finding the best choice (2 bins) yields only 0.713. These results suggest that BP-HMM's flexible approach to temporal structure is very useful for measuring similarity in this challenging dataset.

4.2.2 Unsupervised Learning of Behavior Patterns

Next, we explore the BP-HMM's utility in unsupervised activity *discovery*. After training on all 30 `CMUKitchen` videos, we examine key behaviors and their sharing patterns across videos in Fig. 6. For this illustration, we manually selected a handful of features that best matched human interpretable action concepts. We then plot the appearance patterns of these features across all videos and time, as well as example frames sampled randomly from those assigned to each feature. Note that our visualization only shows detections for a single hand-picked feature linked to each concept. This doesn't necessarily mean a `Pizza` video lacking the "Grate Cheese" feature lacked that activity, but just that the visualized feature was not used.

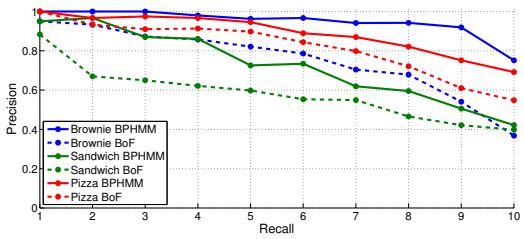

Figure 5. Comparison of BP-HMM with bag-of-features (BoF) representations on retrieval of 10 test videos for each `CMUKitchen` recipe. The BPHMM's temporal representation provides a superior video similarity measure.

Overall, Fig. 6 suggests that the BP-HMM successfully identifies interesting behaviors and intuitive sharing patterns despite its completely unsupervised approach. For example, the "Grate Cheese" and "Slice Pepperoni" behaviors are almost exclusive to videos from the `Pizza` recipe, while both `Pizza` and `Brownie` recipes use the oven near the end. All actors are required by data collection protocol to switch a light on and off at the start and end of their sessions, and we find a corresponding behavior. Furthermore, we discover that only `Sandwich` and `Brownie` recipes require ingredients stored in the overhead cupboard. Some of the depicted feature assignments are false positives. For example, the first "Spread Peanut Butter" frame shown is actually from a `Pizza` video, probably identified based on local motion of the hands. Nevertheless, we observe that behaviors are quite consistent across subjects.

The BP-HMM often discovers *multiple* features that correspond to what a human might consider a single behavior (e.g., stirring ingredients in a bowl). This is driven by subtle differences in observed motion, which produce different codewords and thus distinct states. For example, the "Stir Bowl Unique" feature is unique to subject 13. Inspection reveals that his stirring technique is noticeably different from peers (see supplementary video). This example highlights the ability of our model to identify idiosyncracies and unusual behaviors, which can be useful in many applications.

4.3. Olympic Sports Dataset

The Olympic Sports dataset, introduced by [12], contains sports videos collected from YouTube that have significant temporal structure as well as variability in viewpoint, background clutter, and camera motion. We use release `2010.09.07`, which contains 16 action categories represented by 640 training and 132 test video; this is a

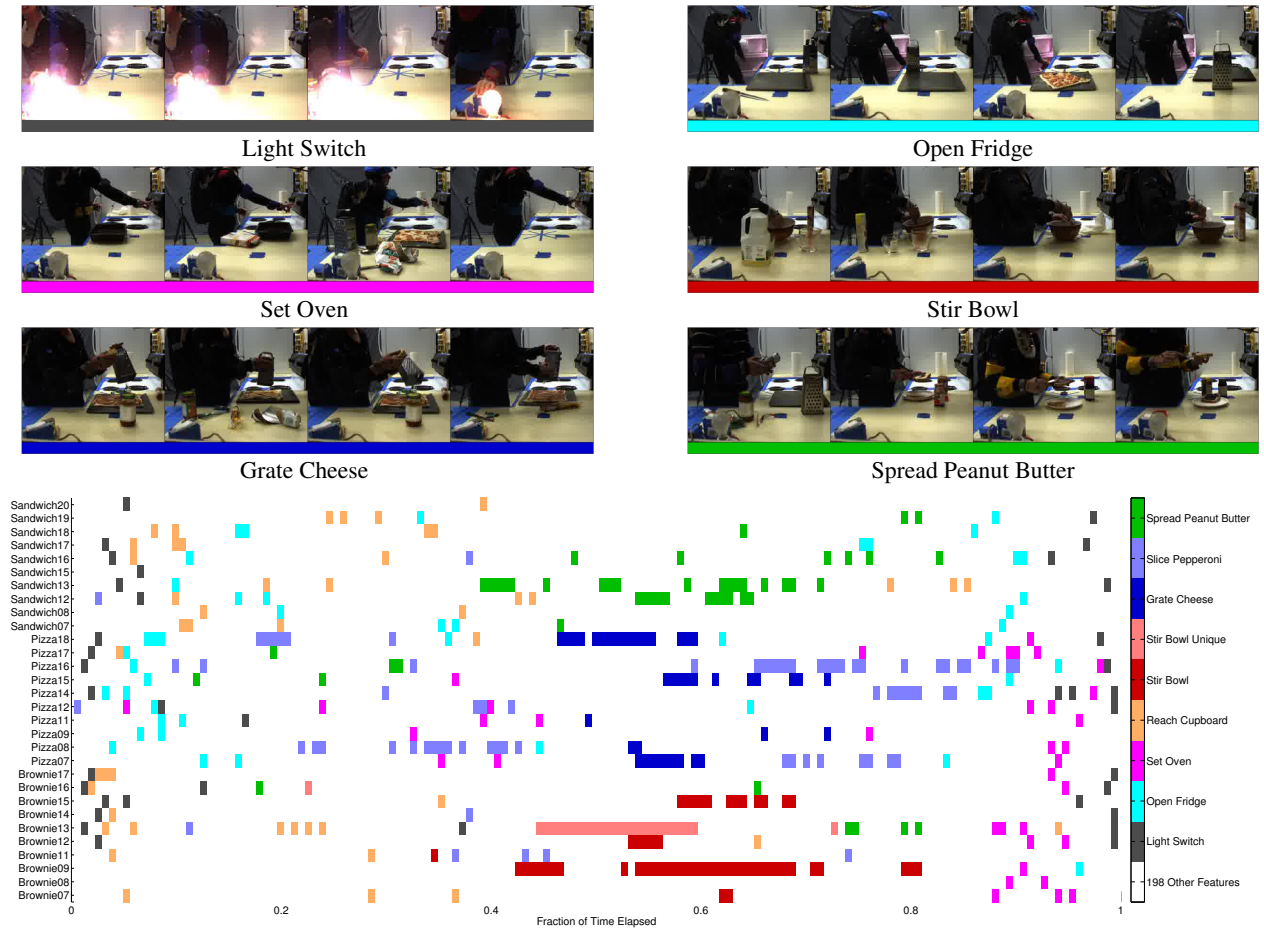

Figure 6. Behaviors found for `CMUKitchen`. *Top*: Example frames assigned to discovered behaviors (text labels are manually assigned). *Bottom*: Assigned locations of behaviors in time across all 30 videos in the corpus. Each row represents a single video, labeled at left by recipe type and actor ID. We show only locations where the feature is assigned to at least two time steps in a local window.

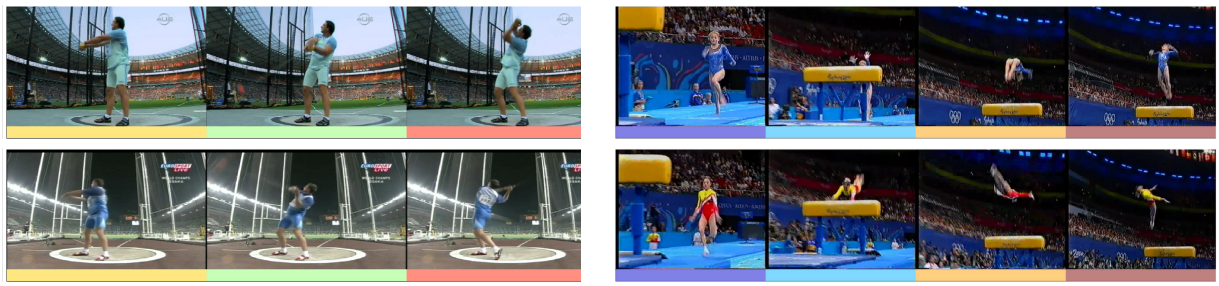

Figure 7. Example frames and behavior annotations recovered by the BP-HMM on `OlympicSports`. Colors indicate distinct behaviors. *Left*: 3 phase `hammer throw` wind-up. *Right*: 4 phase `vault` progression. Annotated videos available in supplementary material.

subset of the data used by [12]. For this dataset, we set our temporal width w to 0.16 s (4 frames per bin). The size and complexity of the corpus requires that we train on data from each category separately.

Qualitative results are shown in Fig. 7. We recover an intuitive breakdown of the periodic wind-up an athlete performs in a `hammer throw`, swinging the hammer around multiple times to build momentum before the throw. We also find sensible `vault` segmentations, with separate phases for the approach, jump, acrobatics, and landing. These results highlight the flexibility and expressiveness our approach can bring to video collections.

In quantative evaluations, however, the BP-HMM does not outperform a simple BoF in Olympic Sports retrieval (F1-score of 0.25 vs. 0.32), though both scores are poor in an absolute sense. The BP-HMM's lower score is likely ex-

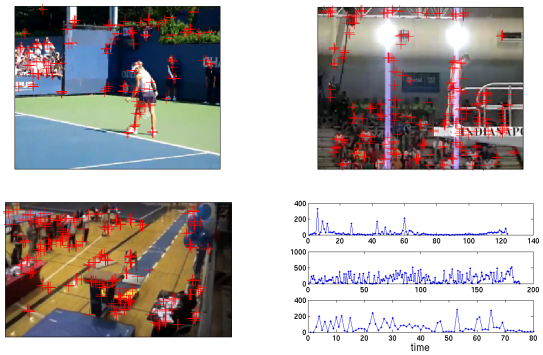

Figure 8. Key frames and noisy STIP detections (red crosses) for Olympic Sports clips with clutter, camera motion, and variable lighting. *Lower Right:* STIP counts over time for each example. Frequent spikes lead to BP-HMM oversegmentation.

plained by unreliable interest point detection. Fig. 8 shows that huge instantaneous spikes in STIP detections can occur due to background motion or camera shake. While the holistic BoF approach can be robust to such noise, spurious temporally localized STIPs can dramatically alter the input to the BP-HMM, resulting in significant oversegmentation. Removing camera motion as well as isolating foreground activity in the input video representation would likely improve the BP-HMM's performance.

5. Discussion

We have demonstrated unsupervised activity discovery in video collections via the BP-HMM. We achieve scalable MCMC inference with our novel data-driven proposals, and encourage more consistent, detailed segmentations via global sharing of dynamics parameters. We expect improved video representations and more efficient inference methods to be fruitful avenues for further work.

Acknowledgments MCH supported by an NSF graduate research fellowship. `CMUKitchen` data (`http://kitchen.cs.cmu.edu`) funded in part by NSF Grant No. EEEC-0540865.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3), 2011. 1

[2] M. Beal, Z. Ghahramani, and C. Rasmussen. The infinite hidden Markov model. In *NIPS*, 2002. 3

[3] F. De la Torre et al. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Technical Report CMU-RI-TR-08-22, CMU Robotics Institute, 2008. 2, 6

[4] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *NIPS*, 2010. 1, 2, 3, 4

[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5:1020–1056, 2011. 3

[6] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2006. 2, 4

[7] M. Hoai, Z. Lan, and F. de la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011. 1, 5

[8] D. Kuettel, M. Breitenstein, L. V. Gool, and V. Ferrari. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 3

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2, 6

[10] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *CVPR*, 2007. 1

[11] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *CVPR*, 2008. 1

[12] J. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, pages 392–405, 2010. 1, 2, 6, 7

[13] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008. 1

[14] I. Pruteanu-Malinici and L. Carin. Infinite hidden Markov models for unusual-event detection in video. *IEEE Trans. on Image Processing*, 17(5):811–822, 2008. 3

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004. 2

[16] S. L. Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *JASA*, 97(457):337–351, 2002. 4

[17] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 3

[18] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian Buffet process. In *AISTATS*, 2007. 2

[19] Z. Tu and S. C. Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *PAMI*, 24(5):657–673, 2002. 4

[20] P. K. Turaga and A. Veeraraghavan. From videos to verbs: Mining videos for activities using a cascade of dynamical systems. In *CVPR*, 2007. 1

[21] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2, 6

[22] X. Wang, X. Ma, and W. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *PAMI*, 31(3):539 – 555, 2009. 3

[23] Y. Wang and G. Mori. Human action recognition by semi-latent topic models. *PAMI*, 31(10):1762–1774, 2009. 1

[24] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin. Dependent hierarchical beta process for image interpolation and denoising. In *AISTATS*, pages 883–891, 2011. 3