Nonparametric Discovery of Activity Patterns from Video Collections

Michael C. Hughes
Brown University
mhughes@cs.brown.edu

Erik Sudderth
Brown University
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 using a hierarchical *beta process HMM (HBP-HMM) to improve behavior sharing among videos with the same category label. We illustrate discovery of intuitive and useful dynamical structure, at various temporal scales, from videos of simple exercises, Olympic sporting events, and recipe preparation. Video retrieval experiments show that our approach leads to quantitative improvements over conventional bag-of-feature representations.*

1. Introduction

We consider the problem of understanding the temporal structure within a video corpus. Our goal is to uncover patterns of unfolding events within individual video sequences, identify how these patterns are shared across videos, and potentially use our inferred representation to retrieve similar videos. We present a new Bayesian nonparametric generative model, the *hierarchical beta process hidden Markov model* (HBP-HMM), which captures this structure via an inferred set of latent dynamical behaviors. Each video is represented as a discrete time series of vector-quantized spatio-temporal interest points, which are in turn generated by latent states with Markovian dynamics.

The earlier BP-HMM [6] provides a powerful model for generic sequential data. This model defines an unbounded global library of local behaviors (emission distributions) and describes each sequence with a finite subset of these behaviors (one per time-step). We adapt the BP-HMM to the video domain and introduce the HBP-HMM, an extension which adapts the hierarchical beta process [18] to obtain category-specific biases in the frequencies with which behaviors occur. Our usage of Bayesian nonparametric priors allows the number of behaviors, and thus the model's internal structure, to grow and adapt as new data is observed. Importantly, the HBP-HMM can discover multiple distinct patterns within each category, and thus model activities with significant within-class variability.

1.1. Previous Work

Video understanding, and in particular supervised activity recognition, is a widely studied area [21], [1]. Many contemporary approaches begin by extracting descriptors of local spatio-temporal interest points, which are then vector quantized into a "bag of words" [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 understanding differences between complicated activities (e.g., the long jump and triple jump). Simple extensions have built independent 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 [24]. However, this rigid structure cannot learn behaviors which are shared across categories, or model detailed behavior patterns which distinguish examples of the same category.

Among action models that capture some notion of temporal evolution, many presume external, expert knowledge of the activity domain, either by specifying the action semantics [10] or even predefining motion-capture templates for every possible action [12]. Linear dynamical systems have been used for unsupervised temporal learning [22], but without notions of discrete behaviors or shared structure among multiple videos. Variants of hidden markov models have been explored [11] for tasks like sign language recognition, but these are limited to single video tasks and require manual specification of the behavior set. More recently, Niebles *et al*. [14] proposed a discriminative recog-

nition 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. Furthermore, the number of classifiers must be manually specified.

1.2. Contributions

As one of the first applications of Bayesian nonparametrics to video analysis, we make several important contributions. 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. 4) to the more structured patterns of sporting events (e.g., a gymnast's vault routine contains running, vaulting, in-air acrobatics, and landing, as in Fig. 1). 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 within a reversible jump MCMC algorithm for posterior inference. Previously, Fox *et al*. employed simple proposals from the prior, and their experiments with motion capture data train a BP-HMM for at most 6 sequences at once. With our novel data-driven proposals, we can rapidly explore the parameter space of hundreds of videos simultaneously. Importantly, our data-driven moves are general purpose and can generalize to other applications of the (H)BP-HMM outside the video domain.

We begin in Sec. 2 by describing the BP-HMM and HBP-HMM models, as well as the interest point representation they build on. Sec. 3 then derives MCMC methods for learning and inference, with a focus on our data-driven proposals. Sec. 4 demonstrates recovery of interesting dynamical structure from three datasets: KTH actions [16], Olympic sports [14], and CMU kitchen activities [8]. Compared to a bag of words baseline, this learned structure also leads to superior video retrieval performance.

2. Beta Processes for Video Analysis

After describing our video representation (Sec. 2.1), we review existing Bayesian nonparametric binary featural models (Sec. 2.2), and the beta process HMM (Sec. 2.3). We then describe our primary technical contribution, the hierarchical BP-HMM (Sec. 2.4), and discuss related nonparametric models (Sec. 2.5).

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. In particular, we randomly subsample to obtain 100,000 descriptors, use the *K-means++* initialization [2] for robustness, and choose the most accurate codebook from 10 random initializations. Each STIP is then mapped to the nearest codeword, providing a standard "bag of words" representation [23].

To represent videos as discrete time series, we choose a temporal bin-width w (in seconds for invariance to framerate), 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 natural and intuitive descriptions of the high-level actions found in any video corpus. We assume there exists a global set of possible atomic actions corresponding to short-term movements, which we will call behaviors or *features*[1]. These features are characterized by distributions on the set of STIP codewords, and are linked over time to create semantically meaningful, long-term *activities*. Each video sequence in the corpus exhibits a sparse subset of the global features (e.g., a clip might contain running and jumping, but not diving or 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 of the unbounded collection of features. Global feature k is represented by its probability of inclusion b_k, and the parameters θ_k that relate that behavior to observed STIPs. These global variables are determined by an underlying stochastic process, the *beta process*:

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

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

[1] This terminology comes from the machine learning community, where the beta process (and Indian Buffet Process) are used for latent feature models. This is not to be confused with the common computer vision use of "feature" as a reference to a descriptor (e.g. SIFT).

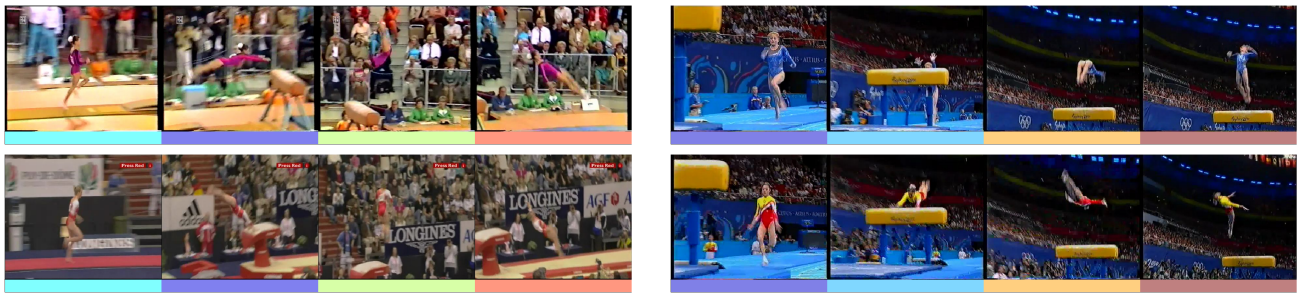

Figure 1. State sequences and associated example frames recovered by BP-HMM for several `vault` videos in OlympicSports. The BP-HMM assigns an appropriate subset of global features to explain each video. The behavior (state) at each frame is indicated by the colored bar below it. Here, videos on the left (side views of the gymnast) exhibit a similar sequence of behaviors, while the videos on the right (oncoming views) share a completely different sequence of behaviors. This intuitive distinction is entirely driven by the observed motion statistics, showcasing the flexibility of the BP-HMM as a model for video collections.

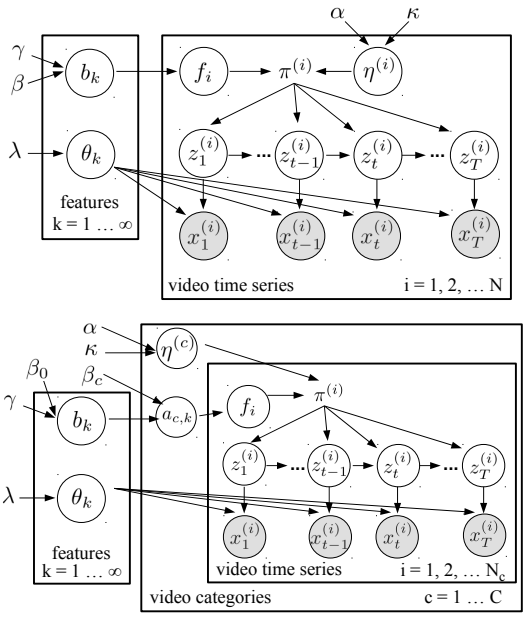

Figure 2. Representation of the BP-HMM (top) and HBP-HMM (bottom) as directed graphical models. Note that we illustrate the HBP-HMM with *category* specific dynamics for the HMM transition parameter η, but in fact the model can also be used with independent dynamics for each video. Likewise the BP-HMM could be altered to have category-specific or even universally-shared dynamics. See Sec. 2 for details.

rule for f_i known as the *Indian buffet process* (IBP) [7]. In this analogy, objects (videos) are customers, and features (behaviors) are dishes in a restaurant. The first customer (video) samples $\text{Poisson}(\gamma)$ unique dishes (behaviors). Successive customer i takes each previously sampled dish with probability $\frac{m_k}{i}$ proportional to the number of previous videos to sample it, and also samples $\text{Poisson}(\frac{\gamma}{i})$ new behaviors. While the IBP representation is intuitive and plays a key role in learning algorithms, the beta process representation is crucial to our later hierarchical extension.

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 [6] shown in Fig. 2. As above we define binary features f_i indicating the behaviors observed in video sequence i, which are coupled by a global feature distribution $B \sim \text{BP}(\beta, \gamma B_0)$. 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 for these parameters is a symmetric Dirichlet, with scalar mass λ_θ and uniform mean $\mathbf{u}_V$:

$$\theta_k \mid B_0 \sim \text{Dirichlet}(\lambda_\theta \mathbf{u}_V) \tag{2}$$

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

Our baseline model, proposed by Fox *et al.* [6], 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:

$$\eta_{jk}^{(i)} \sim \text{Gamma}(\alpha + \kappa\delta_{j,k}, 1) \tag{3}$$

$$\pi_j^{(i)} = \frac{\eta_j^{(i)} \circ f_i}{\sum_{k:f_{i,k}} \eta_{jk}^{(i)}} \tag{4}$$

Here, $\circ$ denotes the element-wise vector product, and $\delta_{j,k}$ the Kronecker delta function. This definition of $\pi_j^{(i)}$ assigns positive transition probability only to those features k present in f_i. The *sticky* parameter κ places extra expected mass on a self-transition in the HMM [5], encouraging the model to learn state sequences with the temporal persistence characteristic of real activities.

The transition matrix $\pi^{(i)}$ and emission distributions θ fully parameterize the HMM which generates the observed STIPs. For each time bin t, we draw its feature assignment

$z_t^{(i)} \in \{k \mid f_{ik} = 1\}$ according to

$$z_t^{(i)} \sim \pi^{(i)}_{z_{t-1}^{(i)}} \tag{5}$$

The L_t spatio-temporal codewords in bin t, whose histogram we denote by $x_t^{(i)}$, are then emitted according to

$$x_t^{(i)} \sim \text{Multinomial}(\theta_{z_t^{(i)}}, L_t) \tag{6}$$

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 state transition dynamics is flexible, in many situations we expect there to be stronger relationships among different video sequences. We thus also consider the following, alternative prior:

$$\eta_{jk}^{(0)} \sim \text{Gamma}(\alpha + \kappa\delta_{j,k}, 1) \tag{7}$$

$$\pi_j^{(i)} = \frac{\eta_j^{(0)} \circ f_i}{\sum_{k:f_{i,k}} \eta_{jk}^{(0)}} \tag{8}$$

Here, a single *common* set of weights is normalized by sequence-specific feature activations. Alternatively, we also define category-specific transition weights $\eta_{jk}^{(c)}$ as in Eq. (7), which are normalized as in Eq. (8). The overall model is summarized by the directed graphical model of Fig. 2.

Note that there can still be significant variability in multiple state sequences $z_t^{(i)}$ sampled from common Markov dynamics.

2.4. Hierarchical BP-HMM

While the BP-HMM can flexibly describe complex relationships among video sequences, it does not model known video groupings or categories. We certainly expect videos depicting the same activity to have more similar dynamics, and to exhibit more highly overlapping subsets of the global behaviors. To model these additional relationships, we adapt the hierarchical beta process (HBP) [18].

For each of C activity categories, we begin by defining category-specific feature inclusion probabilities according to the following hierarchy of Beta processes:

$$B \sim \text{BP}(\beta_0, \gamma B_0), \tag{9}$$

$$A_c \sim \text{BP}(\beta, B), \quad c = 1, \dots, C. \tag{10}$$

Each category has its own feature inclusion probabilities $A_c = [a_{c1}, a_{c2}, \dots]$ for a common set of global behaviors, defined by B. The concentration parameter β determines the variability across categories.

A primary advantage of this hierarchical structure is that for categories with few exemplars, estimates will be robustly shrunk towards information learned from other categories. If video i is a member of category c_i, its active features are sampled according to $f_{ik} \sim \text{Ber}(a_{c_i k})$.

2.5. Related Work

Applications of Bayesian nonparametric models to the video domain are limited. For general nonparametric modeling of sequential data, alternatives to the BP-HMM include the earlier infinite HMM [3] and the hierarchical Dirichlet process HMM [17]. The latter has been used in the video domain to perform unusual event detection [15]. In far-field static camera surveillance, [25] present a model for finding instantaneous rules via the HDP. Later work by [4] present a dependent Dirichlet process DDP-HMM that uncovers temporal rules for traffic motion patterns in a single scene. Their model has both unbounded state space and unbounded number of Markov chains for describing activities in the scene, improving [19] which uses only one chain and require manually specifying the model size.

Our model is novel in its emphasis on understanding *collections* of videos rather than simply individual scenes from one camera angle, and in its featural representation of behavior-video relationships via the beta process. Using a dirichlet process would force all videos to have positive probability of displaying all behaviors, while the beta process elegantly allows a video to contain only a sparse subset of relevant behaviors. The beta process has been used for vision tasks such as image denoising [26], but thus far has not been applied to time-series modeling for video.

3. MCMC Methods for Posterior Inference

The BP-HMM and HBP-HMM have significant combinatorial latent variable structure. To perform inference, we appeal to Markov Chain Monte Carlo methods. We base our algorithms on the efficient and exact MCMC procedure in Fox *et al*. [6]. We briefly overview the relevant updates here. For complete algorithms consult the supplementary material.

Fox *et al*. present a collapsed sampler for the BP-HMM, which marginalizes over feature inclusion parameters $\mathbf{b}$ as well as the state assignments $\mathbf{z}$ and leads to faster mixing. Conditional updates to the remaining variables – feature matrix F, the instantiated emissions distributions θ_k, and transition weights $\eta^{(i)}$ – proceed in an iterative fashion.

Sampling F – We iterate over each time series i and sample its features f_i in two parts: features shared by some other time series, and features unique to time series i. The shared update proposes flipping binary entries in f_i one at a time, accepting via Metroplolis-Hastings. This step is the computational bottleneck of our inference. With K_i assigned features and T_i discrete time steps, each update of a shared feature has cost $O(K_i^2 T_i)$ due to the dynamic programming necessary to compute the likelihood. Profiling indicates over 80% of computation time is spent here.

The unique features update proposes the birth or death of some feature in a Reversible Jump MCMC fashion. We im-

prove upon the original procedure presented in Fox *et al*. by developing novel *data-driven* proposals for the emissions parameters θ^* of a newborn feature, inspired by [20]. Intuitively, draws from the prior as used in [6] will rarely explain the sparse distributions of observed codewords found in realistic data, resulting in low acceptance rates and slow exploration of feature space. Instead, we suggest drawing emissions parameters from a mixture of the prior and the empirical distribution of codewords observed in the sequence. We further extend this idea by customizing to *subwindows* of the sequence in question. In this proposal, we choose a subwindow of the current time series at random, build its empirical distribution, and draw a new θ^* from a mixture of this and the prior.

To validate this contribution, we examine all three proposals in figure 3. Starting all chains from a poor configuration of just 5 states shared among all `triple-jump` videos, we see that chains with subwindow proposals rapidly achieve significantly higher log-likelihood than either alternatives due to their ability to many more create new states relevant to explaining the data at hand.

Sampling HMM parameters η, θ – Given fixed F and η, we instantiate state sequences $\mathbf{z}$ as auxiliary variables and then achieve closed form conjugate posterior updates for θ. Similar updates exist for η given $\mathbf{z}$ and relevant hyperparameters, though we note importantly that the update equations given in [6] are slightly incorrect. The correct posterior for $\eta_{j,k}|\mathbf{z_i}, \alpha, \kappa$ is given up to a proportionality constant as

$$p(\eta_{j,k}|\mathbf{z}_i, f_{i,k}=1) \propto \frac{\eta_{j,k}^{N_{j,k}^{(i)}+\alpha+\delta_{j,k}\kappa-1} e^{-\eta_{j,k}}}{\left[\sum_{k':f_{i,k'}=1} \eta_{j,k'}\right]^{N_j^{(i)}}} \qquad (11)$$

where $N_{j,k}^{(i)}$ counts the number of transitions from state j to k in sequence z_i, and $N_j^{(i)} = \sum_{k:f_{i,k}=1} N_{j,k}^{(i)}$. Note that draws from this posterior can be obtained by sampling a Dir($N_{j,k}^{(i)}+\alpha+\delta_{j,k}\kappa$) vector and then scaling it by a Gamma random variable with parameter $K\alpha + \kappa$. This process inverts the usual Gamma to Dirichlet scaling transformation.

HBP-HMM – We follow [18]'s scheme for HBP inference, marginalizing out category-level feature-inclusion parameters $\mathbf{a}_c$ and updating top-level parameters $\mathbf{b}$ via rejection sampling. Sampling feature assignments F proceeds as above, except the prior terms in the acceptance ratio now depend on $\mathbf{b}$. Updates for the category-wide transition weights $\eta^{(c)}$ are no longer conjugate and require Metropolis Hastings updates.

Hyperparameters – We treat all hyperparameters as fixed constants. We set the transition weight parameters $\alpha = 2$ and the self-transition weight $\kappa = 10\alpha$. We fix BP mass $\gamma = 2$ and concentration $\beta_0 = 1$. For the HBP-HMM, we let $\beta_c = \frac{1}{2}$ to promote within-category sharing. Finally, we set λ to 500 on KTH and 750 for the other datasets, encouraging moderate sparsity in emissions.

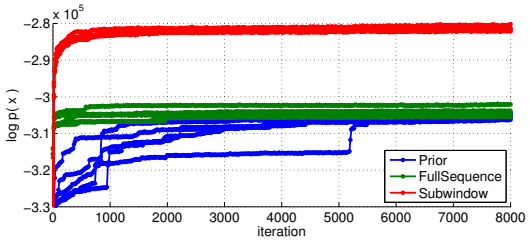

Figure 3. Comparison of different proposal schemes for the birth move of RJMCMC inference across 5 chains. We plot the log likelihood as a function of iteration, which strongly improves as the model discovers more activities. The data-driven subwindow proposals quickly reach a high-likelihood configuration preferred by the model, while the rarely-accepted prior proposals result in mode jumping between local optima. Data: 21 `triple-jump` videos from Olympic Sports.

4. Experiments

We investigate the capabilities of the BP-HMM and HBP-HMM by applying our models to three video datasets. Our goals here are twofold. First, we illustrate the useful chronological structure our approach recovers. Second, we demonstrate that our representations produce improved performance over bag-of-words for measuring video similarity.

To assess this second goal, we propose an information retrieval task where the goal is to identify videos in a held-out test set similar to a query video from training. We prefer this retrieval evaluation to a recognition framework, because our model's strengths lie in capturing unique temporal structure for individual videos rather than entire categories.

For the quantitative retrieval task, we fit trained models to held-out test data as follows. First, we fix feature parameters $\hat{\theta}$ and training assignments $\hat{F}$ after sufficently many iterations of MCMC. Next, for each video in the corpus with observed time series $\mathbf{x}^*$, we sample features and state assignments from $p(f^*, z^*|x^*, \hat{F}, \hat{\theta})$ via MCMC, skipping updates to $\hat{\theta}$ and sampling f^* conditioned on fixed F. We thus obtain S samples $\{z^{*,1}, \dots z^{*,S}\}$ to compute a summary descriptor ϕ^* for the retrieval task.

We use a very simple descriptor: a histogram counting the fraction of time each video spent in each of the K fixed states defined by $\hat{\theta}$. For each video in the corpus, we compute $\phi_k^* = \sum_s \sum_{t=1}^{T^*} \delta_{z_t^{(*,s)},k}$ and then normalize.

To compare videos we employ the exponentiated χ^2 kernel, as done in [23] for BoF representations. We define similarity between the histograms of videos i and i' as

$$k(i,i') = \exp\left[-\frac{1}{A}\chi^2(\phi_i, \phi_{i'})\right] \qquad (12)$$

where $\chi^2(\cdot)$ is the chi-square distance and A is the average

pairwise distance in the training set.

Using this similarity measure, for each action category we search the test set for videos that match each positive example in training, and average over these to create precision-recall curves. We report the macro-averaged F1-score as the summary statistic across all categories.

4.1. KTH Dataset

The KTH actions dataset contains 6 simple exercise actions performed by 25 actors across 2396 video sequences. 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. We use the original training-validation-test split presented in [16]. We train each BP-HMM model for 500 iterations on the entire 760 video training set.

Most instances within a KTH category have very similar temporal content, making this an interesting dataset for assessing the capabilities of the HBP-HMM. We compare three versions of the model: (1) BP-HMM with independent dynamics for each video, (2) BP-HMM with category-wide dynamics η, and (3) an HBP-HMM that shares dynamics *and* feature inclusion probabilities within a category. Figure 4 shows example state sequences recovered by each version across several videos from `jogging` and `handwaving` categories. The top 3 jogging videos have actors moving to the left, the bottom 3 jog right. We can see all models differentiate between these directions of motion as well as recover periodic structure for handwaving. Across both categories, the HBP-HMM's structure is most consistent across instances, indicating the benefits of sharing feature inclusion as well as dynamics information. Additionally, the HBP-HMM recovers finer temporal structure: it identifies 2-3 distinct gait phases of each jogging subtype, and 6 distinct handwaving phases. When category instances have highly similar chronological content, using the hierarchical approach of sharing features and dynamics via the HBP-HMM offers benefits for recovering detailed, consistent structure. However, further experiments reveal that the variability between sampler runs may be larger than the variability between these model classes.

Evaluating our best performing model (HBP-HMM) against the standard BoF representation for our new information retrieval task on KTH, we find the HBP offers noticeable improvement. BoF obtains a summary F1 score of 0.583, while the HBP-HMM obtains 0.639.

4.2. Olympic Sports Dataset

The Olympic Sports dataset, introduced by [14], 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`[2], which contains 16 action categories represented by 640 training and 132 test videos collected from YouTube. Note that this release contains only some of the sequences reported in [14].

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 to efficiently complete a sufficent number of iterations (8000).

Qualitative results from example `vault` videos are shown in Figure 1. We observe that our model identifies distinct sets of features to explain a side-view and the oncoming-view of the gymnast's vault. Our non-parametric method adapts to the data at hand without requiring an expert to define the number of different possible views or geometric reasoning capabilities.

Further qualitative results are shown in Figure 5. We recover an intuitive breakdown of the periodic wind-up an athlete performs in a `hammer throw`, swinging the hammer around his head multiple times to build momentum before the throw. In the `snatch` example, we learn features that correspond to different phases of weight-lifting motion. Finally, the `triple jump` example shows 3 distinct jumps, each broken down into a two-state pattern that corresponds to the up-down motion of the jumping athlete. These results highlight the flexibility and expressiveness our modeling framework can bring to video understanding.

In our quantative evaluations, however, we obtain an F1-score of 0.254 while BoF obtains 0.321. BP-HMM's poor performance is likely explained by poor interest point detections. While the holistic BoF approach can be robust to a fair amount of noisy detections due to background clutter or camera shake, these can dramatically alter the behavioral representations our model recovers. Removing camera shake and motion as well as isolating foreground from background activity before obtaining a video representation would improve the performance of our model considerably.

4.3. CMU Kitchen Dataset

As a final investigation, we apply the BP-HMM model to a collection of videos from the CMU Multi-Modal Activity Database [8]. Each video is several minutes long and depicts a single actor in the same kitchen creating a prescribed dish from start to finish. We chose three 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 in time as actors exhibit noticeable variability in ordering of parts of a recipe, making this data well-suited to the BP-HMM with individualized dynamics. We also inspected results from the HBP-HMM, but observed similar behaviors and sharing patterns with no notable differences, as we might expect given the small corpus size and the huge variation in dynamics between actors.

We complete two evaluations: a quantiative retrieval

[2] http://vision.stanford.edu/Datasets/OlympicSports/

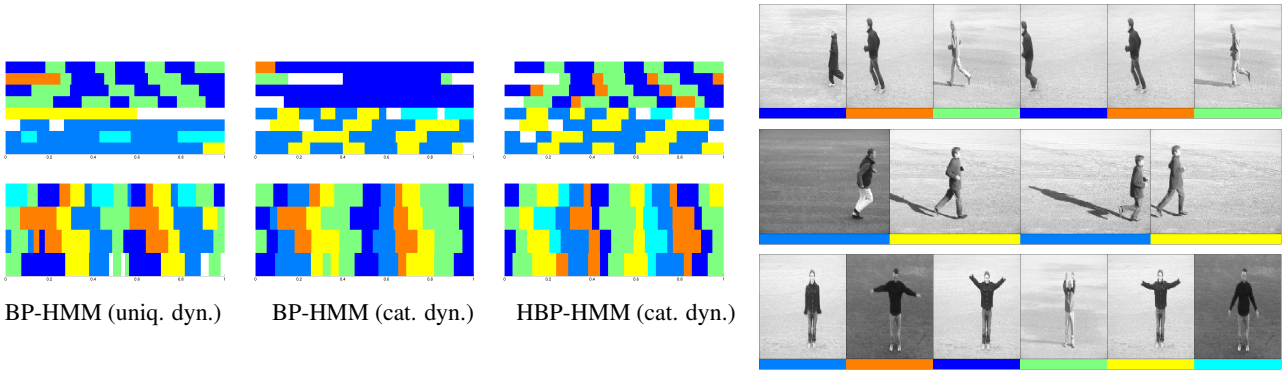

Figure 4. **Left**: Comparison of state seq. $\mathbf{z}_i$ recovered by different models for example videos from `jogging` (top) and `handwaving` (bottom) KTH videos. We compare models with dynamics parameters η unique to each video (uniq. dyn.) with models shared across videos of the same category (cat. dyn.). Each row shows feature assignments over time in single example video, same colors denote same features within each plot. For the shown sampler runs, the HBP-HMM recovers more consistent segmentations across videos and finer detail. However, further experiments indicate this superior performance is not always consistent and highly dependent on initialization. **Right**: video frames for the repeated patterns discovered by HBP-HMM, sampled at random from example videos. Colors match to HBP-HMM plot of the same category.

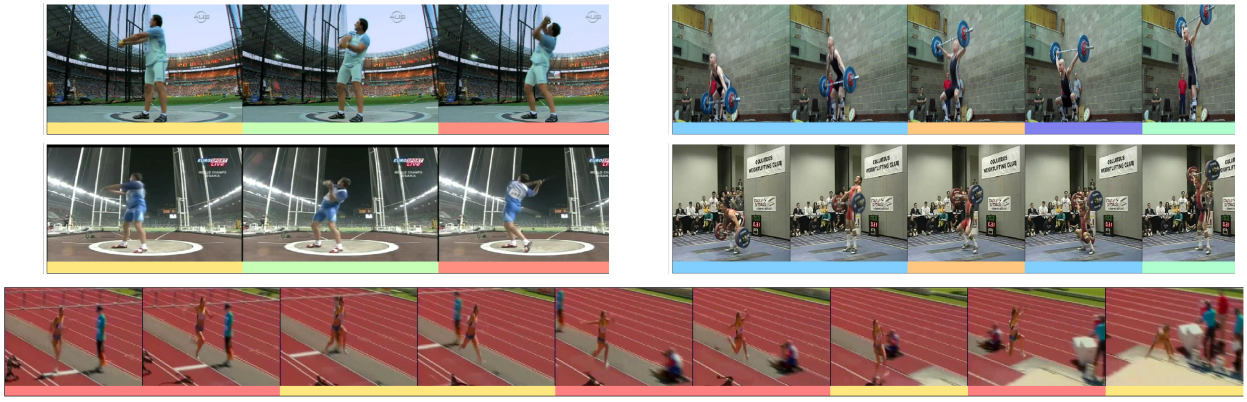

Figure 5. Example frames and associated features for patterns recovered by BP-HMM on `OlympicSports`. Bar color indicates distinct feature assigned to each frame. Colors do not correspond across categories. **Top Left**: 3 phase `hammer throw` wind-up. **Top Right**: `snatch` lift progression. **Bottom**: 3 repeats of up-down pattern discovered for a single `triple jump` video.

task, and an unsupervised exploration of the latent behaviors discovered by our model for this data. For both comparisons, we train a single BP-HMM on all 30 training videos for over 2000 iterations. We set the window size w to be 0.5 seconds (15 frames), since these videos are quite long and coarser scale behaviors are more appropriate.

4.3.1 Retrieval Evaluation

We compare the BP-HMM to the bag-of-features approach in identifying similar videos to those in training in a held-out set of 10 videos from each recipe. We obtain class-specific precision-recall curves shown in Fig. 6. 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

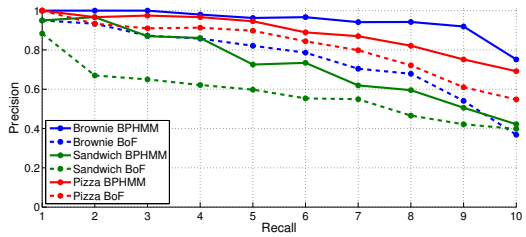

Figure 6. Comparison of BP-HMM with bag-of-features (BoF) on retrieval of 10 test videos for each `CMUKitchen` recipe.

to the rigid temporal discretization of BoF proposed by [9] with both 2 and 3 bins. Best performance (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.

4.3.2 Unsupervised Learning of Behavior Patterns

Finally, we explore BP-HMM's utility as an unsupervised *knowledge discovery* tool for the complex actions of the CMUKitchen dataset. We investigate the global behaviors it recovers and study how these behaviors are shared across videos and used over time. Fig. 7 summarizes structure discovered across all 30 videos.

For this illustration, we manually selected a handful of features that best matched meaningful behaviors for certain subjects. 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 behavior. This doesn't necessarily mean a Pizza video lacking the "GrateCheese" feature was never assigned such a behavior, rather just that the particular feature chosen was unused.

Overall, this visualization suggests the BP-HMM successfully identifies interesting behaviors and intuitive sharing patterns. 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 (though a few subjects appear to preheat it earlier on). We also discover that almost all actors switch a light on and off at the start and end of their sessions, as required by the data collection protocol, and 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 "StirBowlUnique" feature is remarkably unique to subject 13. Inspection reveals that his stirring technique is noticeably different from peers. This example highlights the ability of our model to identify idiosynchracies and unusual behaviors, which can be useful in some applications.

5. Discussion

We have presented nonparametric models that recover shared activity structure in a video collections, obtaining scalable MCMC inference with our data-driven proposals and encouraging videos within a category to share behaviors with the HBP-HMM. We expect improved interest point detections and more efficient inference methods to be fruitful avenues for building upon this work. Incorporating non-exclusive category labels, like those found in the Hollywood dataset [9], is also an open problem.

Acknowledgments The data used in this paper was obtained from `kitchen.cs.cmu.edu` and the data collection was funded in part by the National Science Foundation under 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] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *SODA*, pages 1027–1035, 2007. 2

[3] G. Z. Beal, M.J. and C. Rasmussen. The infinite hidden markov model. In *NIPS*, 2002. 4

[4] L. V. G. D. Kuettel, M. Breitenstein and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 4

[5] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 2011. 3

[6] 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, 3, 4, 5

[7] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. In *NIPS*, 2007. 3

[8] F. D. la Torre et al. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, 2009. 2, 6

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2, 7, 8

[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. Coupled hidden semi markov models for activity recognition. In *WMVC*, 2007. 1

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

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

[14] J. C. 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

[15] I. Pruteanu-Malinici and L. Carin. Infinite hidden markov models for unusual-event detection in video. *IEEE Transactions on Image Processing*, 17(5):811–822, 2008. 4

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, pages 32–36, 2004. 2, 6

[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. 4

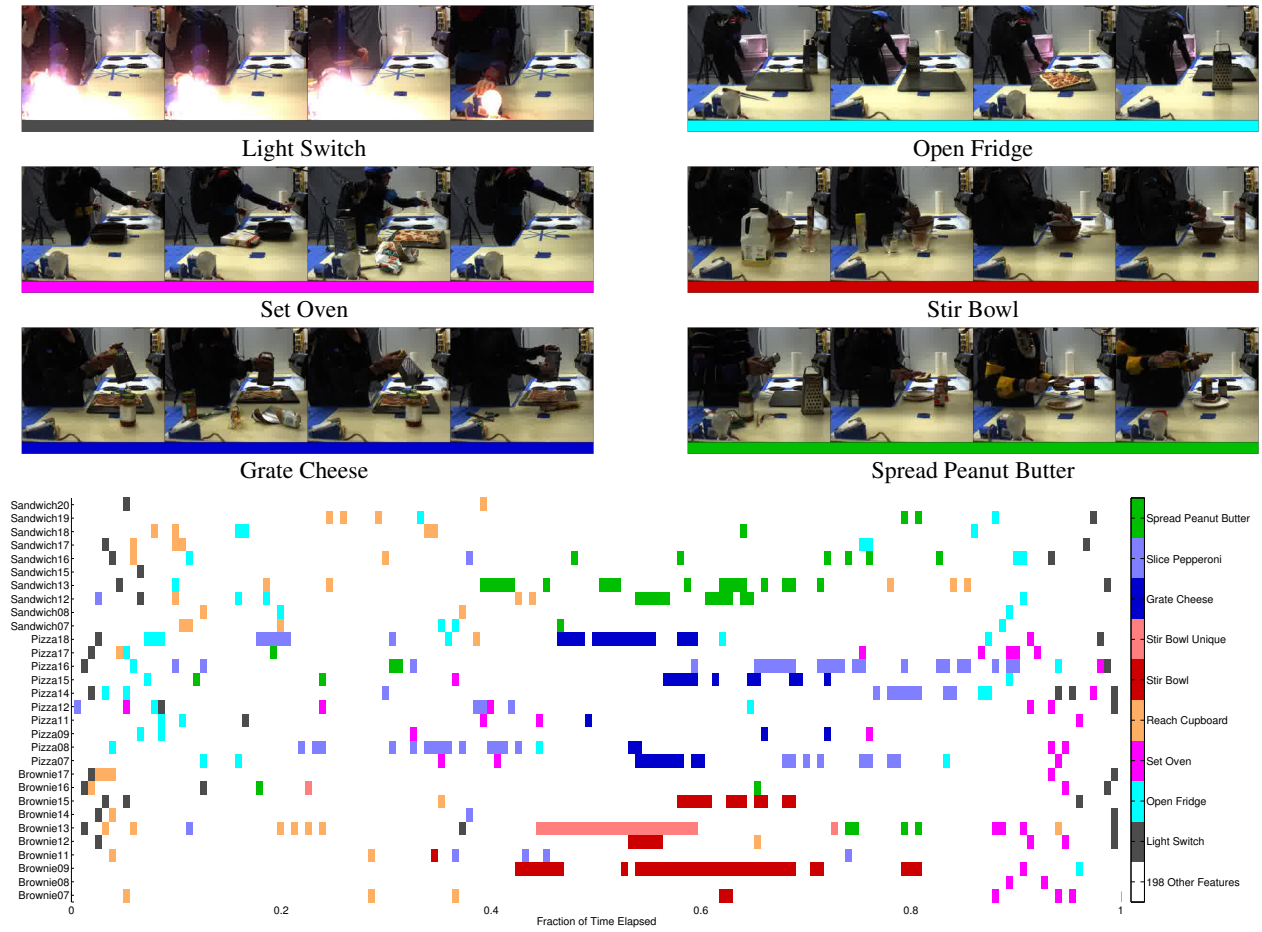

Figure 7. Feature sharing patterns recovered on the `CMUKitchen` dataset. Best viewed in color. **Top**: Example frames assigned to discovered features. **Bottom**: Assigned locations of features in time across all 30 videos in the corpus. Each row represents a single video, labeled at left by recipe type and actor ID. We show only locations where the feature is assigned to at least two time steps in a local window.

[18] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the indian buffet process. *JMLR*, 2:564–571, 2007. 1, 2, 4, 5

[19] S. G. Timothy Hospedales and T. Xiang. A markov clustering topic model for mining behaviour in video. In *ICCV*, 2009. 4

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

[21] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488, 2008. 1

[22] P. K. Turaga and A. Veeraraghavan. From videos to verbs: mining videos for activities using a cascade of dynamical systems. In *CVPR*, pages 1–8, 2007. 1

[23] 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, 5

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

[25] X. M. X. Wang and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *PAMI*, 31(3), 2009. 4

[26] M. Zhou et al. Dependent hierarchical Beta process for image interpolation and denoising. In *AISTATS*, 2011. 4