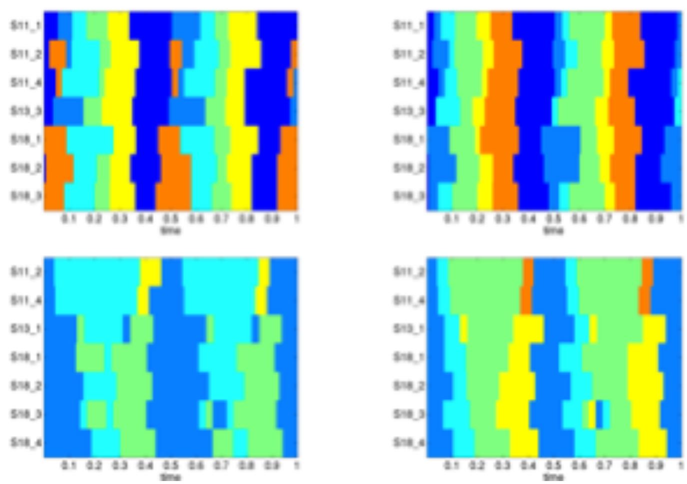
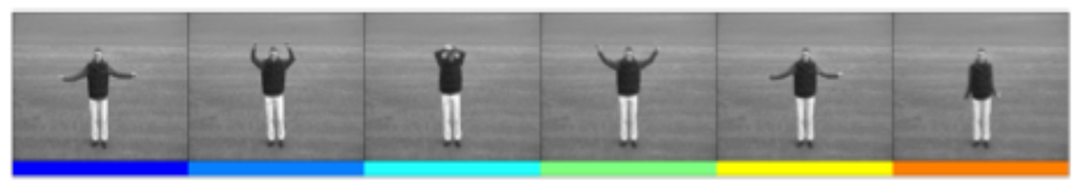
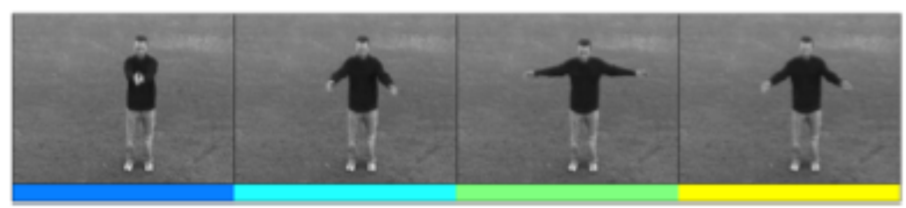
Nonparametric Discovery of Activity Patterns from Video Collections

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

Sequence Dynamics Global Dynamics







Activity Recognition Group

Colin Lea - June 11, 2013

Datasets

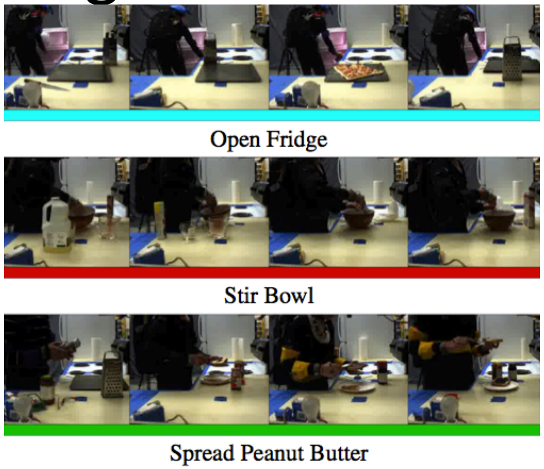
KTH – Simple actions

378 videos {clap, wave, jog}

CMU Kitchen – Long sequences of making food

10 people, 3 tasks

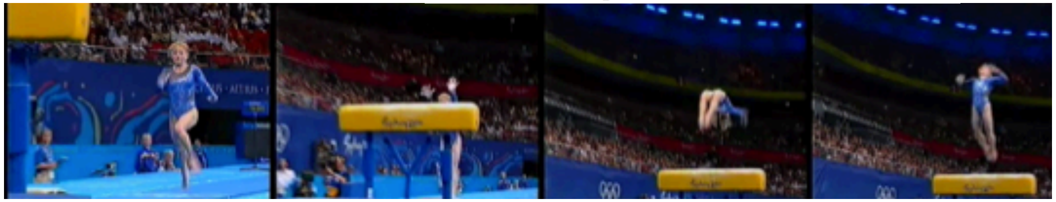
{sandwich, pizza, brownies}



Youtube (Olympic Sports) – Complex

Train: 640 vids x16 actions

Testing: 132 videos



<u>Simple</u>: hand wave, <u>Complex</u>: gymnastics vault routine

Overview

Goals:

1) Discover common <u>behaviors</u> in video collections

2) Segment/classify new videos into set of behaviors

Extends **Beta Process-HMM**:

1) Data-driven MCMC (very helpful)

2) Global sharing of behavior (helpful on KTH)

3) Application on video datasets

Benefits: Segmentation, action retrieval

Visualization of shared dynamical structure

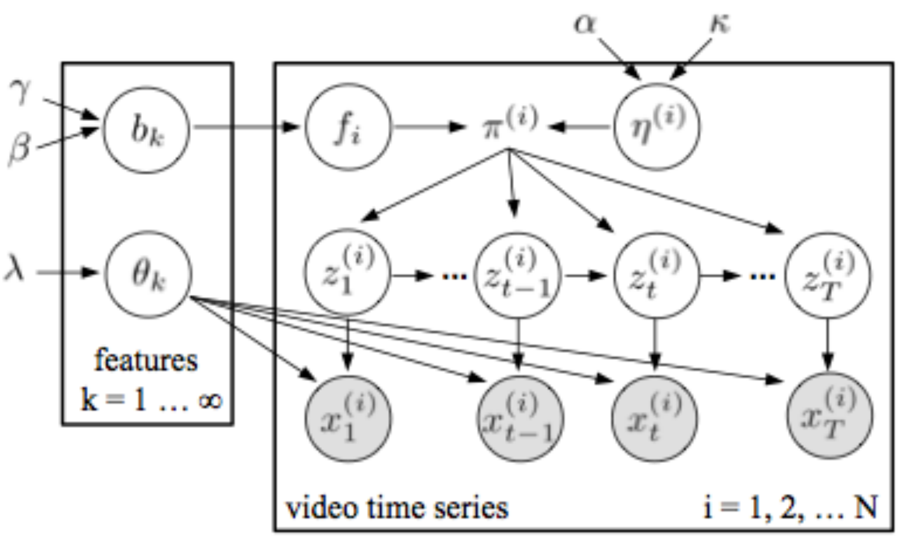
BP-HMM vs HDP-HMM

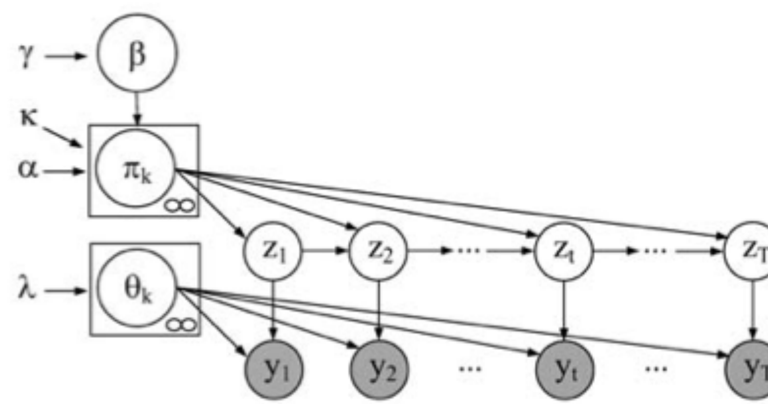
Features : Sparse binary behaviors | All behaviors have P(.) > 0

Learn/Infer: Data Driven MCMC | Reversible Jump MCMC

Indian Buffet Process | Chinese Restaurant Process

Shared Params: Global+Sequence | Global





Terminology

Features := atomic *behaviors* from global set.

Characterized by distribution on the set of STIPs

$f_i = [f_{i1}, f_{i2}, \ldots]$ (sparse, binary)

Activity := Collection of features/behaviors

Visual features

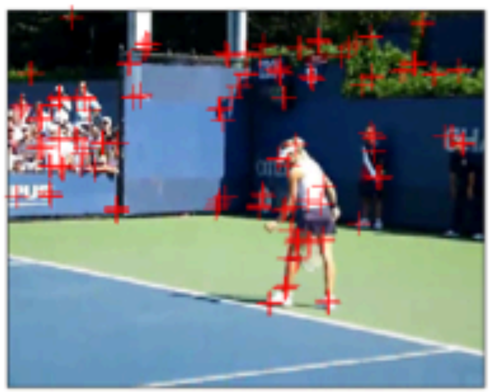
Frames:

2, 4, 15

Bin video into *w* second sequences (w={0.08, 0.16, 0.5})

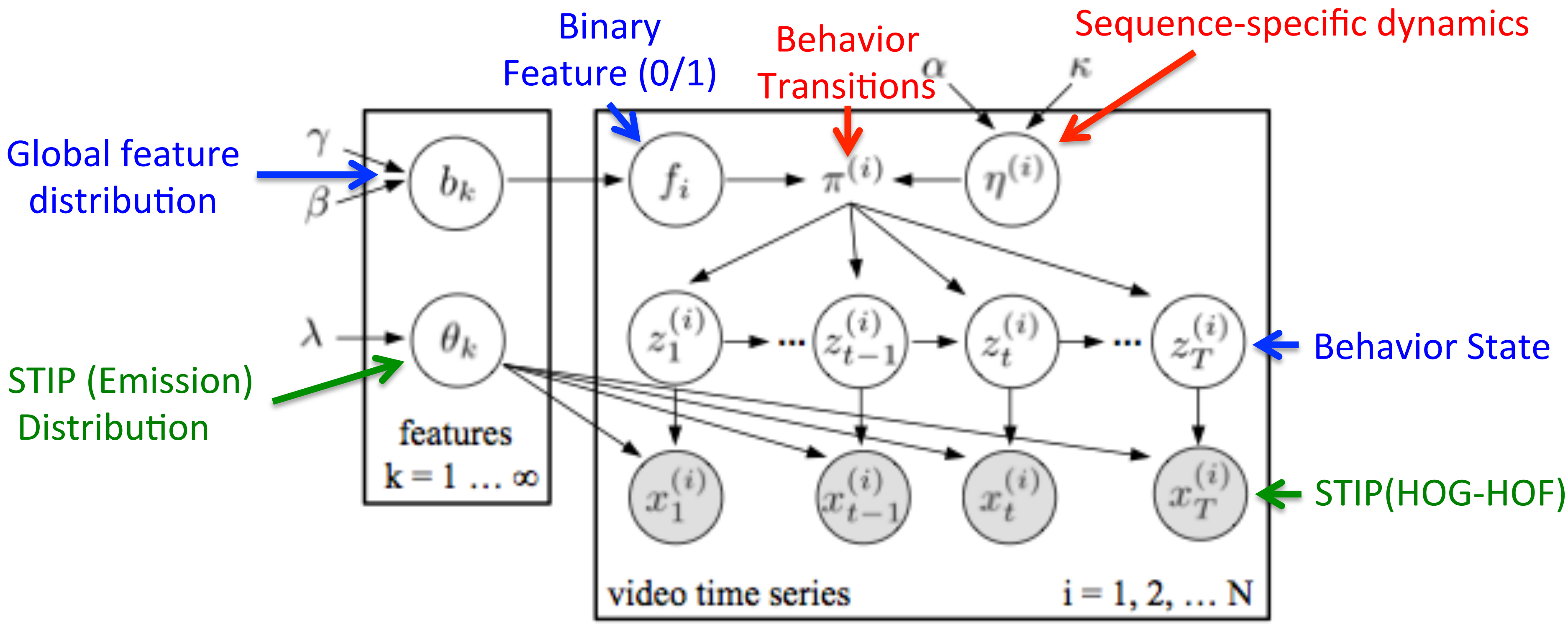
Spatio-Temporal Interest Points (STIPs)

HOG-HOF



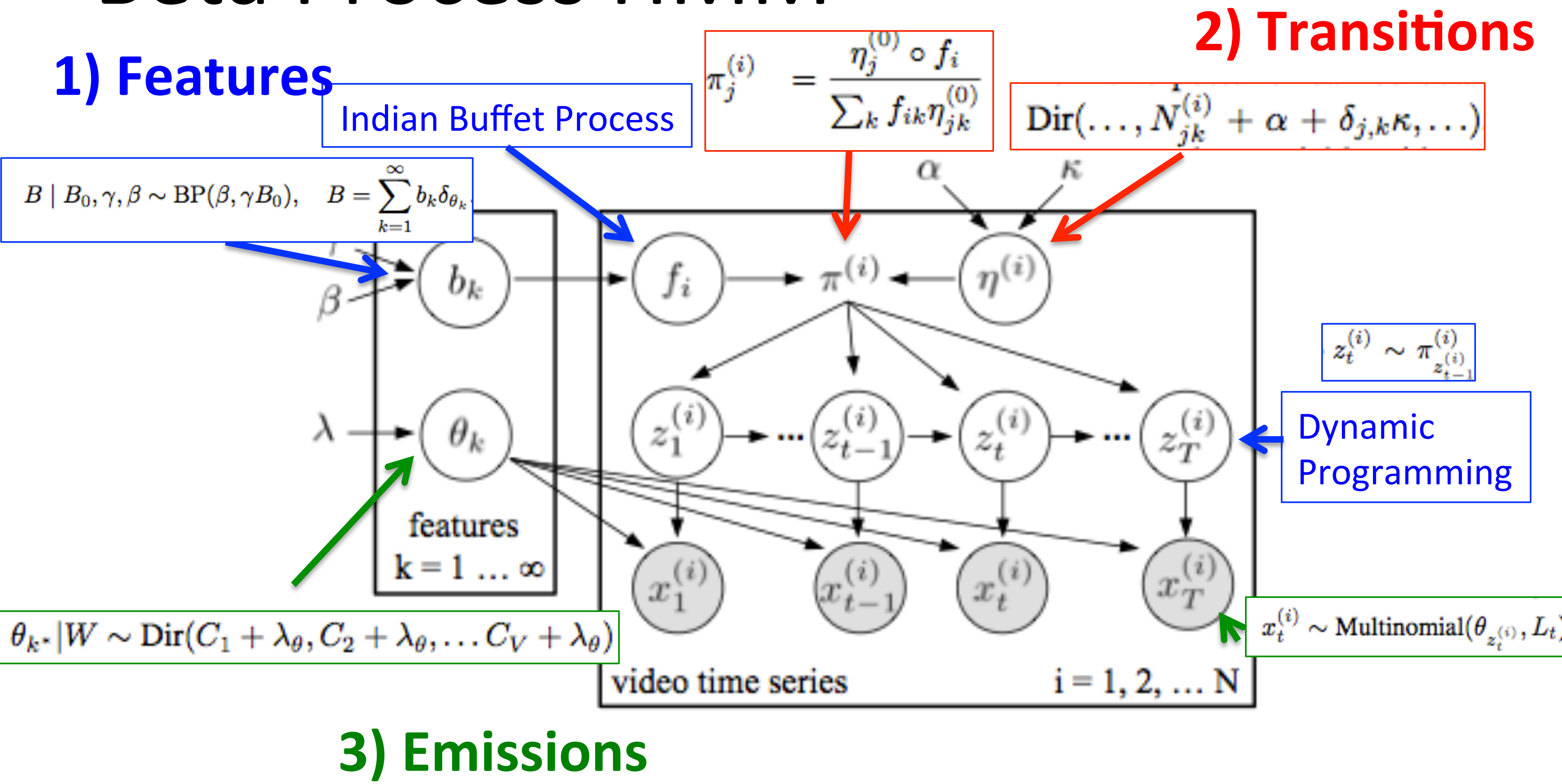
1000 word dictionary per dataset

Beta Process HMM



HMM: pi = Transition matrix theta = Emission distribution

Beta Process HMM



Learning/Inference: MCMC

Collapsed sampler [Fox NIPS'10]:

- Marginalize over global distribution (b) and behavior (z)
- Iterative conditional updates to:
  - Features (F), emissions (theta), transition weights (eta)

Feature sampling:

- Shared: flip value, accept/reject (Metropolis Hastings)
- Sequence(V1): reversible jump MCMC (birth/death) [Fox'10]
  - Vague prior! Low acceptance rates and slow exploration
- Sequence(V2): Data-driven proposal from posterior (theta)

Review: MCMC

Generate new sample from data w/o known density
P = Proposal distribution/matrix, x=state

Algorithm:
Choose proposal distribution $P(x_{t+1} \mid x_t)$

Choice:
Guassian?
Random Walk? Gibbs!

For i = 1…N:
- sample $x^* \sim P(x_{t+1} \mid x_t)$
- sample $u \sim \text{Uniform}(0,1)$
- if $u <= \min\{1, \frac{f(x^*)P(x_{i-1}|x^*)}{f(x_{i-1})P(x^*|x_{i-1})}\}$
  - $x_{t+1} = x^*$
- else:
  - $x_{t+1} = x_t$

$$\frac{f(x^*)}{f(x_{i-1})}$$

>1 if new sample has higher probability than previous sample

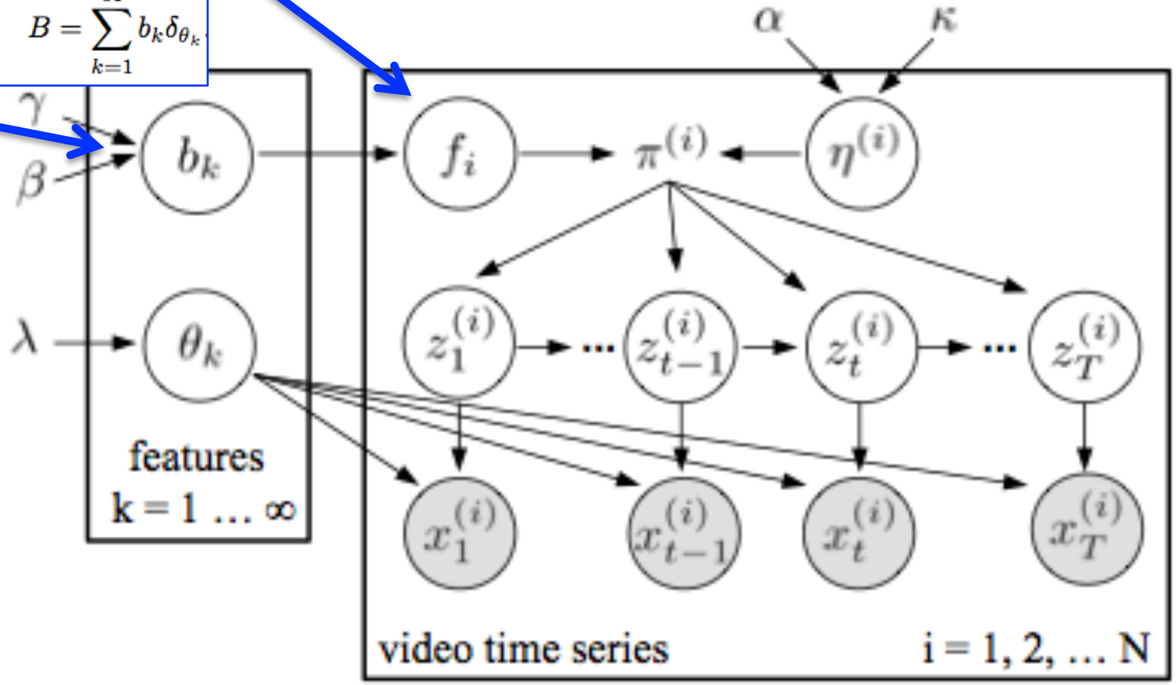
$$\frac{P(x_{i-1}|x^*)}{P(x^*|x_{i-1})}$$

>1 if transition from x_t to x^* is greater than x^* to x_t

1) Features

Indian Buffet Process

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



1) Features

f_i = video = [f_{i1} , f_{i2} , . . ., f_{T2}] (sparse, binary)

F = Corpus

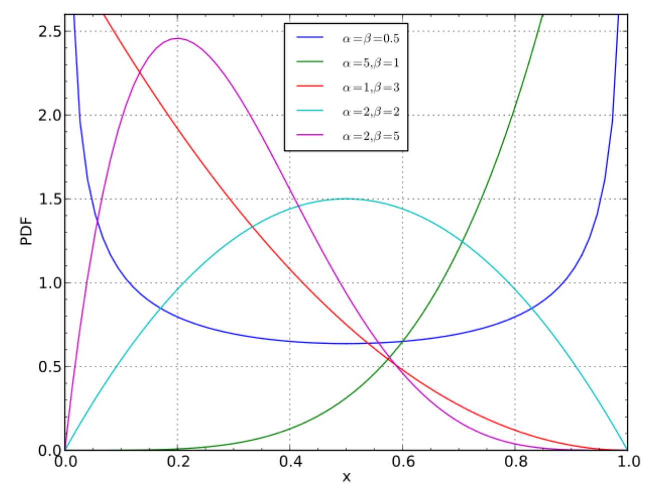
b_k = corpus frequency of feature k

θ_k= STIP distribution parameter

$$\theta_k \sim B_0$$

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

Degree to which features are shared between videos



Beta Distribution

1) Features: **Indian Buffet Process (IBP)**

- Visualize feature assignment as a sequential process of customers sampling dishes from an (infinitely long) buffet:

customers ⟷ *observed data to be modeled* ← **Videos**

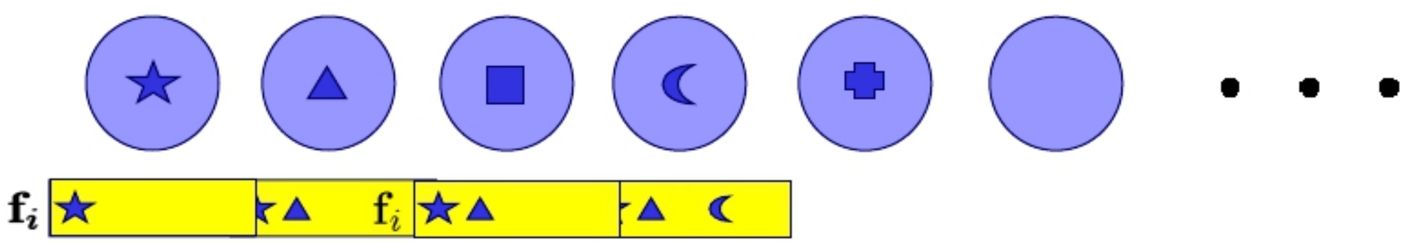
dishes ⟷ *binary features to be selected* ← **Behaviors**

- The first customer chooses Poisson(α) dishes, $\alpha > 0$
- Subsequent customer i randomly samples each previously

$$f_{ik} \sim \mathrm{Ber}\left(\frac{m_k}{i}\right)$$

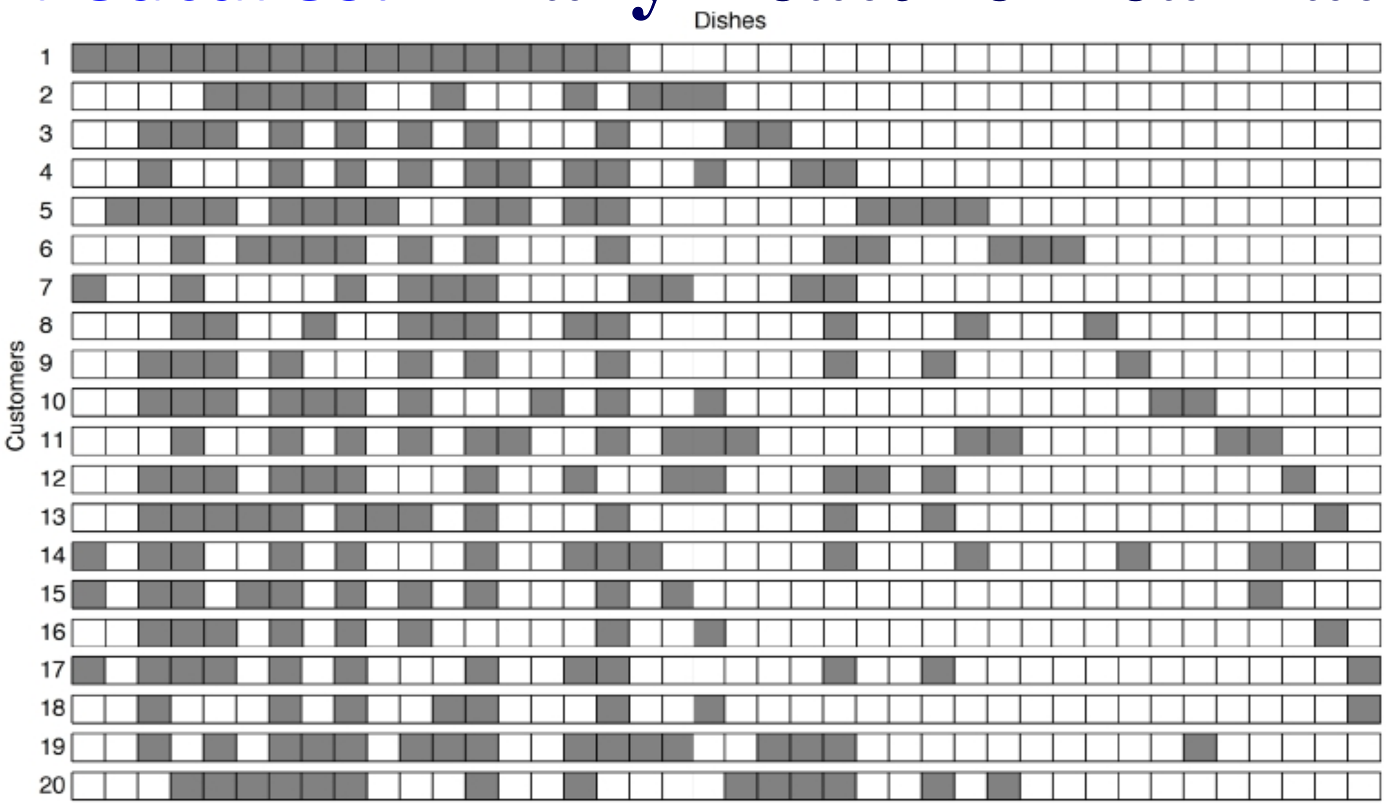
$m_k \longrightarrow$ number of previous customers to sample dish k

- That customer also samples Poisson(α/i) new dishes



[Sudderth NP Bayes Tutorial CVPR'12]

1) Features: **Binary Feature Realizations**



Videos

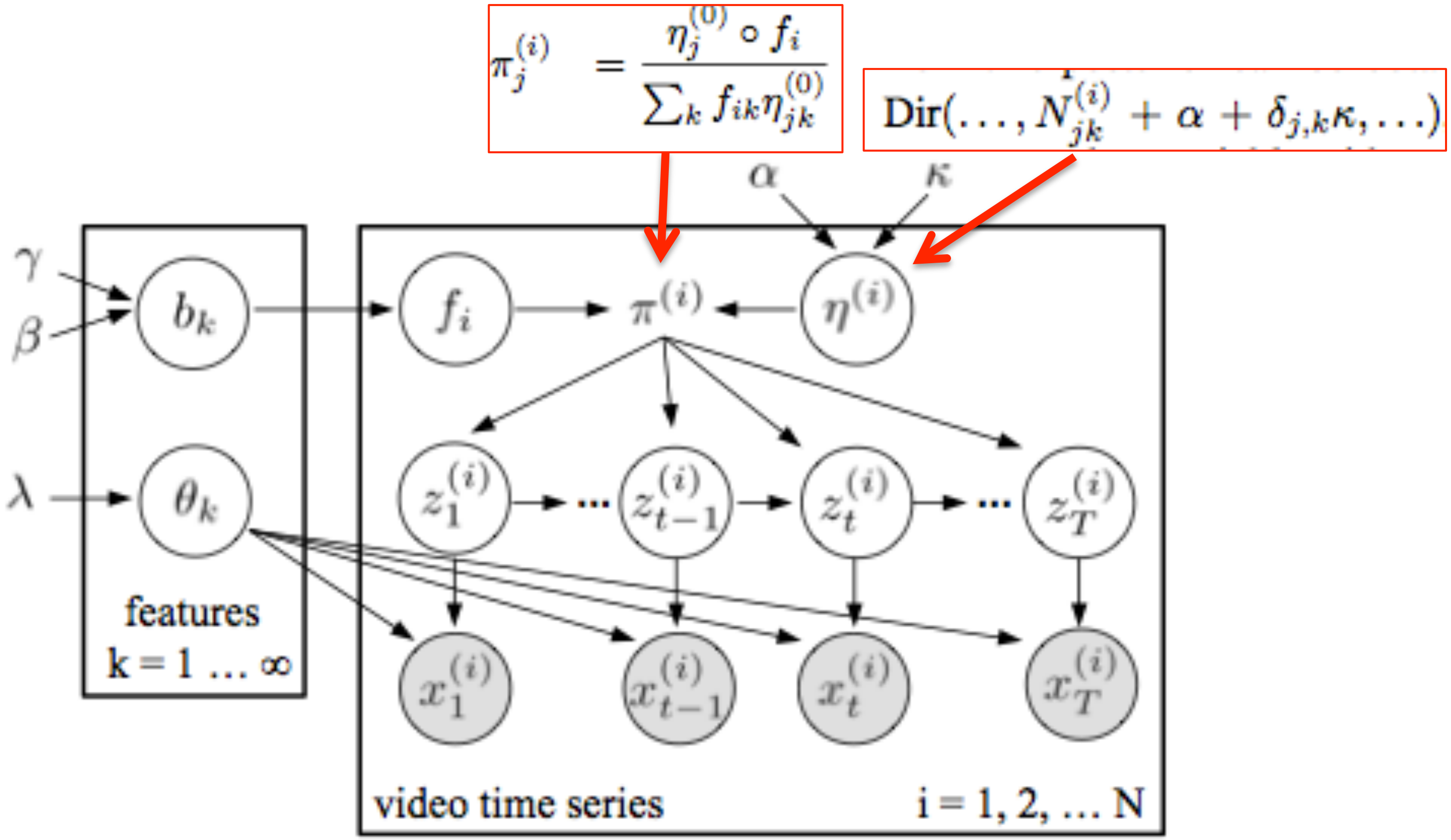
Behaviors

Ghahramani, BNP 2009

- IBP is *exchangeable*, up to a permutation of the order with which dishes are listed in the binary feature matrix
- Clustering models like the DP have one "feature" per customer
- The number of features sampled at least once is $O(\alpha \log N)$

[Sudderth NP Bayes Tutorial CVPR'12]

2) Transitions



2) Transitions

V1: Independent transition dynamics per video *[Fox NIPS'10]*

Transition dynamics: $\eta_{jk}^{(i)} \sim \mathrm{Gam}(\alpha + \kappa\delta_{j,k}, 1)$, $\delta_{j,k} = \{ \begin{matrix} 1 & j = k \\ 0 & j! = k \end{matrix}$

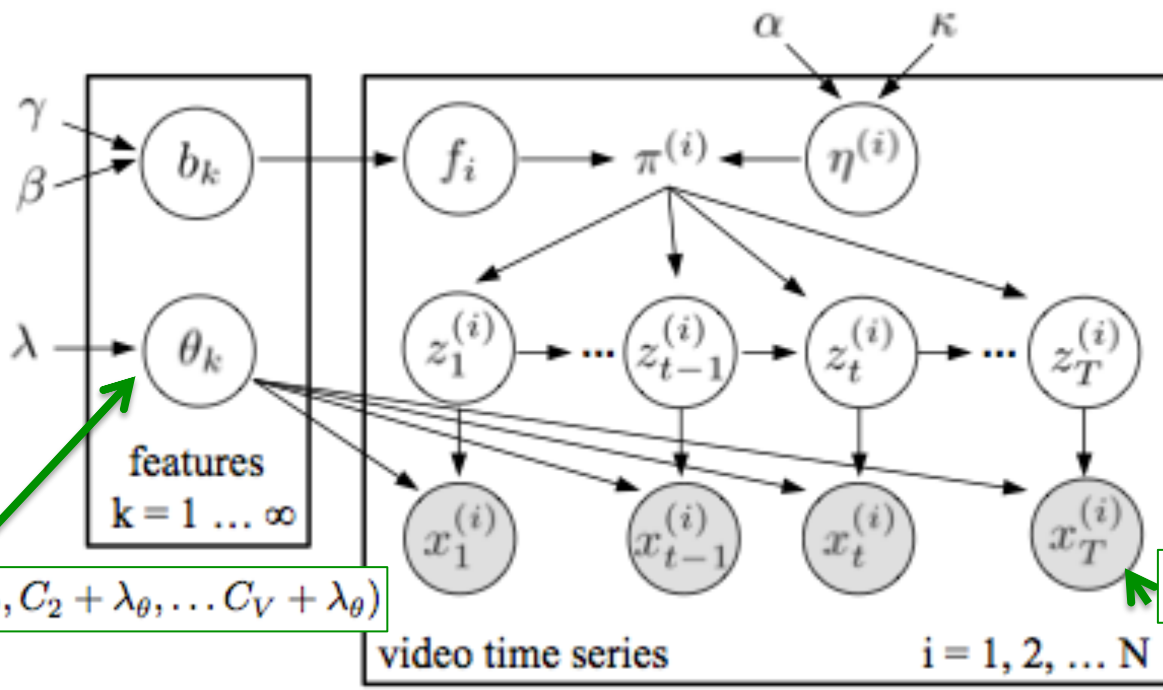
Transition matrix: $\pi_j^{(i)} = \frac{\eta_j^{(i)} \circ f_i}{\sum_k f_{ik} \eta_{jk}^{(i)}}$

V2: Global dynamics

Transition dynamics: $\eta_{jk}^{(0)} \sim \mathrm{Gam}(\alpha + \kappa\delta_{j,k}, 1)$, $\delta_{j,k} = \{ \begin{matrix} 1 & j = k \\ 0 & j! = k \end{matrix}$

Transition matrix: $\pi_j^{(i)} = \frac{\eta_j^{(0)} \circ f_i}{\sum_k f_{ik} \eta_{jk}^{(0)}}$

3) Emissions



$\theta_{k\cdot}|W \sim \mathrm{Dir}(C_1 + \lambda_\theta, C_2 + \lambda_\theta, \ldots C_V + \lambda_\theta)$

$x_t^{(i)} \sim \mathrm{Multinomial}(\theta_{z_t^{(i)}}, L_t$

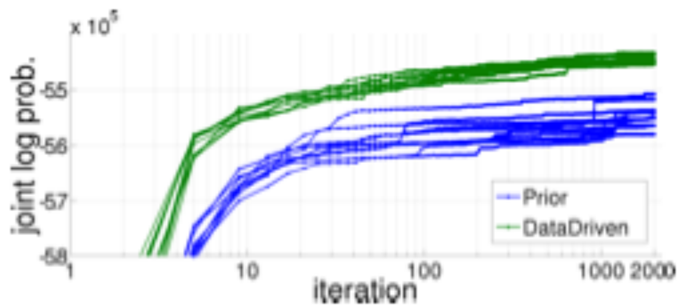
3) Emissions: Data-driven proposal distribution

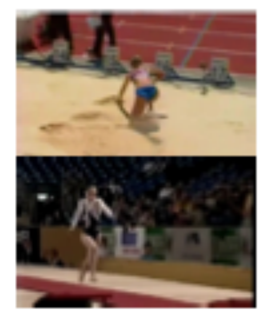
Generate new emissions using codeword count in window

W= window $\theta_{k\cdot}|W \sim \mathrm{Dir}(C_1 + \lambda_\theta, C_2 + \lambda_\theta, \ldots C_V + \lambda_\theta)$

Ci = #counts codeword i in window W

Lambda adds sparsity (lambda=0.75)





(a)

(b)

Differentiation: Log probability of Vault and Triple Jump videos (10 runs each)

(V1) All prior distributions assigned (a) and (b) to same behavior

(V2) 5 of 10 data-driven runs discover different behaviors

Discovery:

(V1) 25 behaviors

(V2) 50 behaviors

Resampling HMM

N_{jk}=#transitions j->k

Emission (theta): sampled from conjugate posterior

$$\theta_{k\cdot} | W \sim \mathrm{Dir}(C_1 + \lambda_\theta, C_2 + \lambda_\theta, \ldots C_V + \lambda_\theta) \qquad x_t^{(i)} \sim \mathrm{Multinomial}(\theta_{z_t^{(i)}}, L_t)$$

Transition Dynamics (V1) $p(\eta_{jk}^{(i)} \mid \mathbf{z}_i, f_{ik} = 1) \propto \dfrac{(\eta_{j,k}^{(i)})^{N_{jk}^{(i)} + \alpha + \delta_{j,k}\kappa - 1} e^{-\eta_{jk}^{(i)}}}{\left[\sum_\ell f_{i\ell} \eta_{j\ell}^{(i)}\right]^{N_{j\cdot}^{(i)}}}$

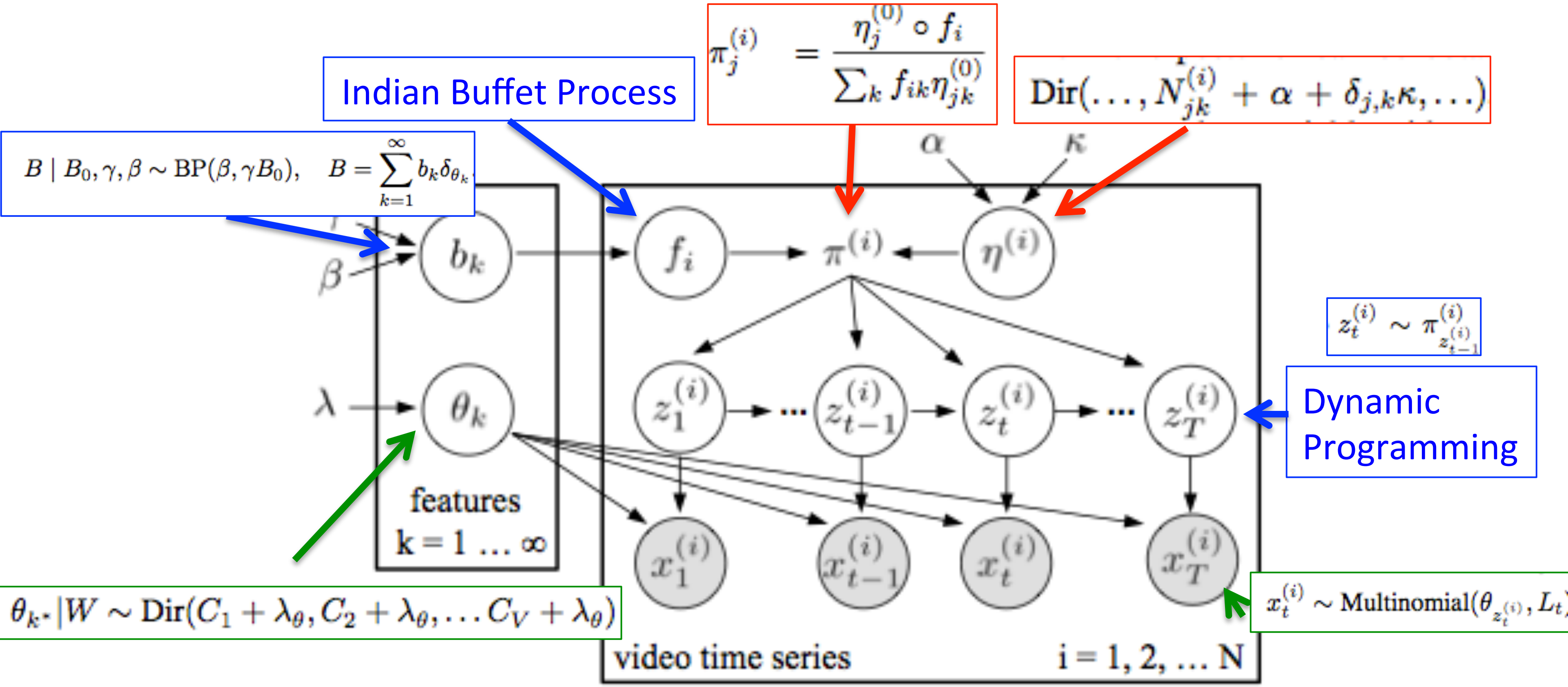
eta ~ $\mathrm{Dir}(\ldots, N_{jk}^{(i)} + \alpha + \delta_{j,k}\kappa, \ldots)$ Local

Transition Dynamics (V2): Similar but Metropolis-Hastings on Gamma random walk (mean=current, var=10) Global

Behavior State: dynamic programming

$$z_t^{(i)} \in \{k \mid f_{ik} = 1\} \text{ according to } z_t^{(i)} \sim \pi_{z_{t-1}^{(i)}}^{(i)}$$

Beta Process HMM



Experiments

For all datasets except KTH, we use sequence-specific dynamics, as global sharing is likely not beneficial when temporal variability is significant.

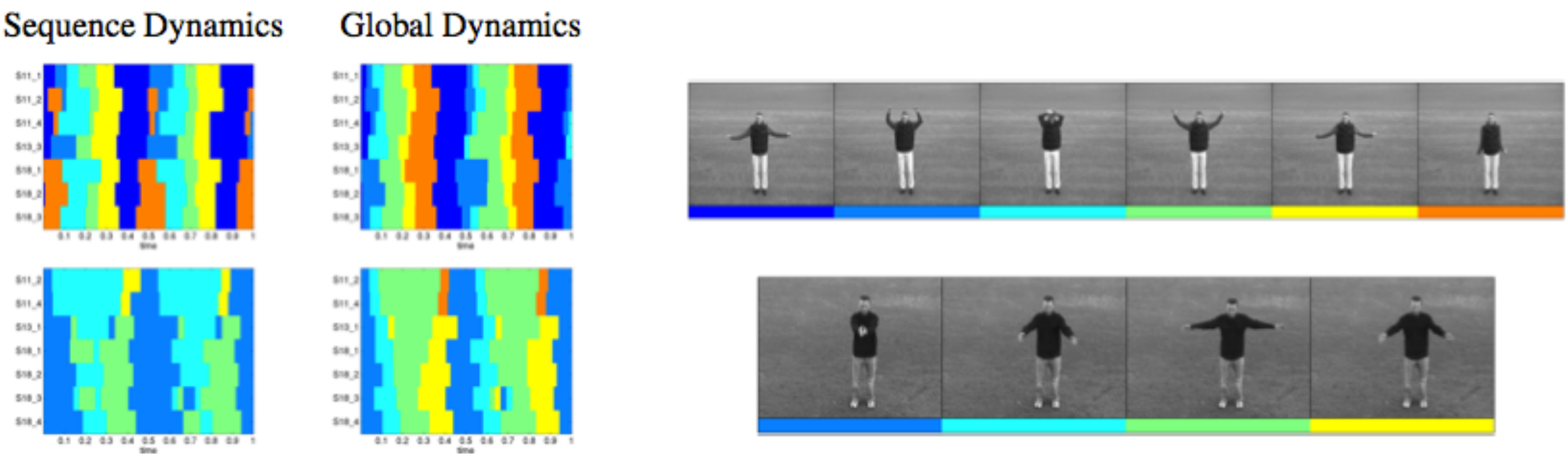
alpha= 2 and kappa = 10*alpha, BP mass Beta0 = 1 (as in the conventional IBP [6])

Experiment 1a: KTH

Hypothesis: Global shared dynamics is better

Training: 378 videos {clap, wave, jog}

Features: Only HOF, w=0.08 (2 frames)



Results: shared has more detailed segments

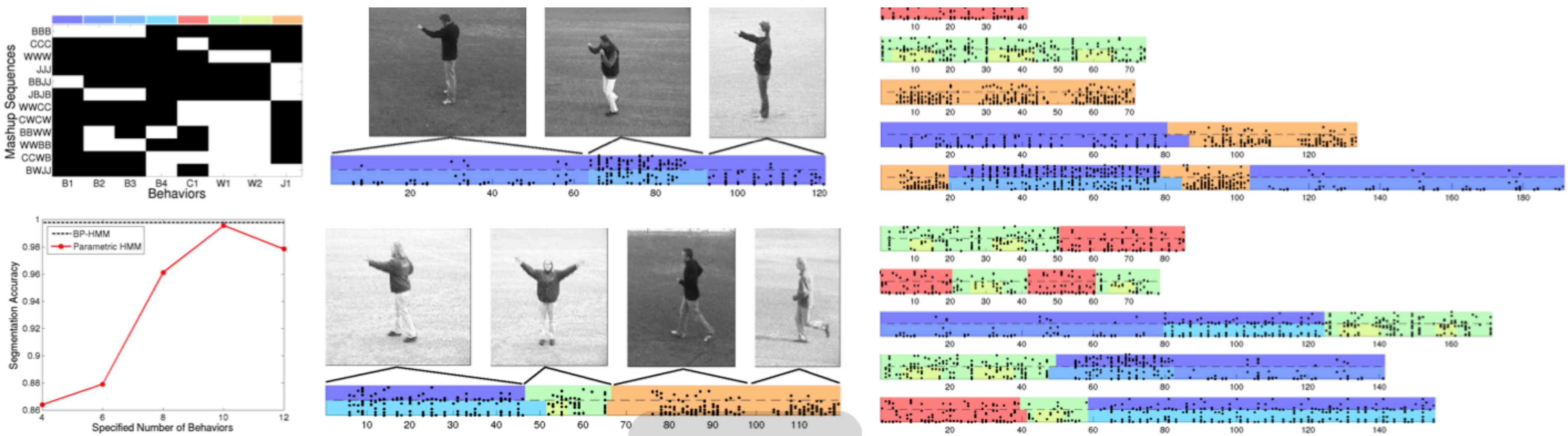
Note: Global sharing "only beneficial on KTH"

Experiment 1b: KTH Mashup

Hypothesis: BP-HMM recovers meaningful segments

Training: 12 sequences w/ mashups {box, clap, wave, jog}

Features: Only HOF, w=0.08 (2 frames)

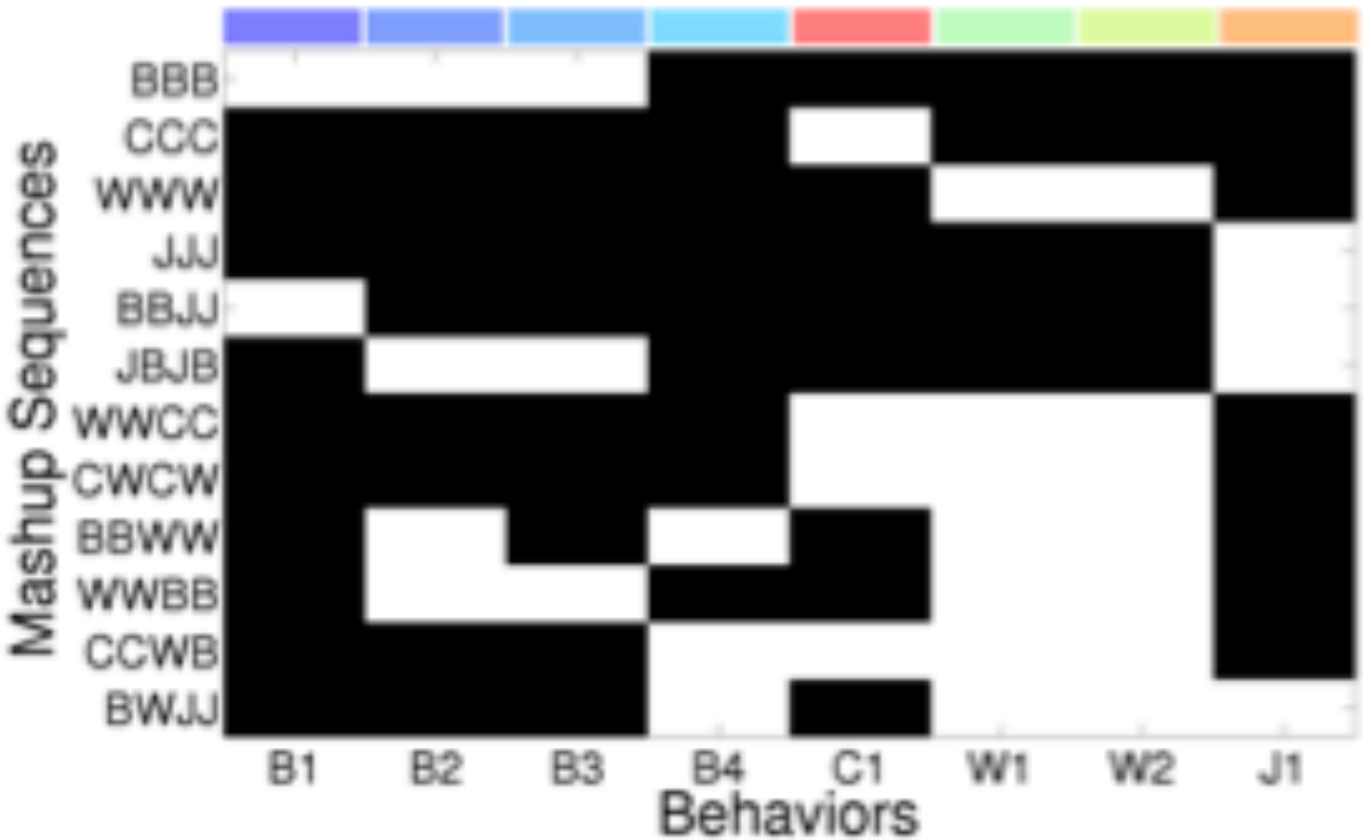


Results: 4 of 9 behaviors for boxing

1 behavior for clapping + jogging

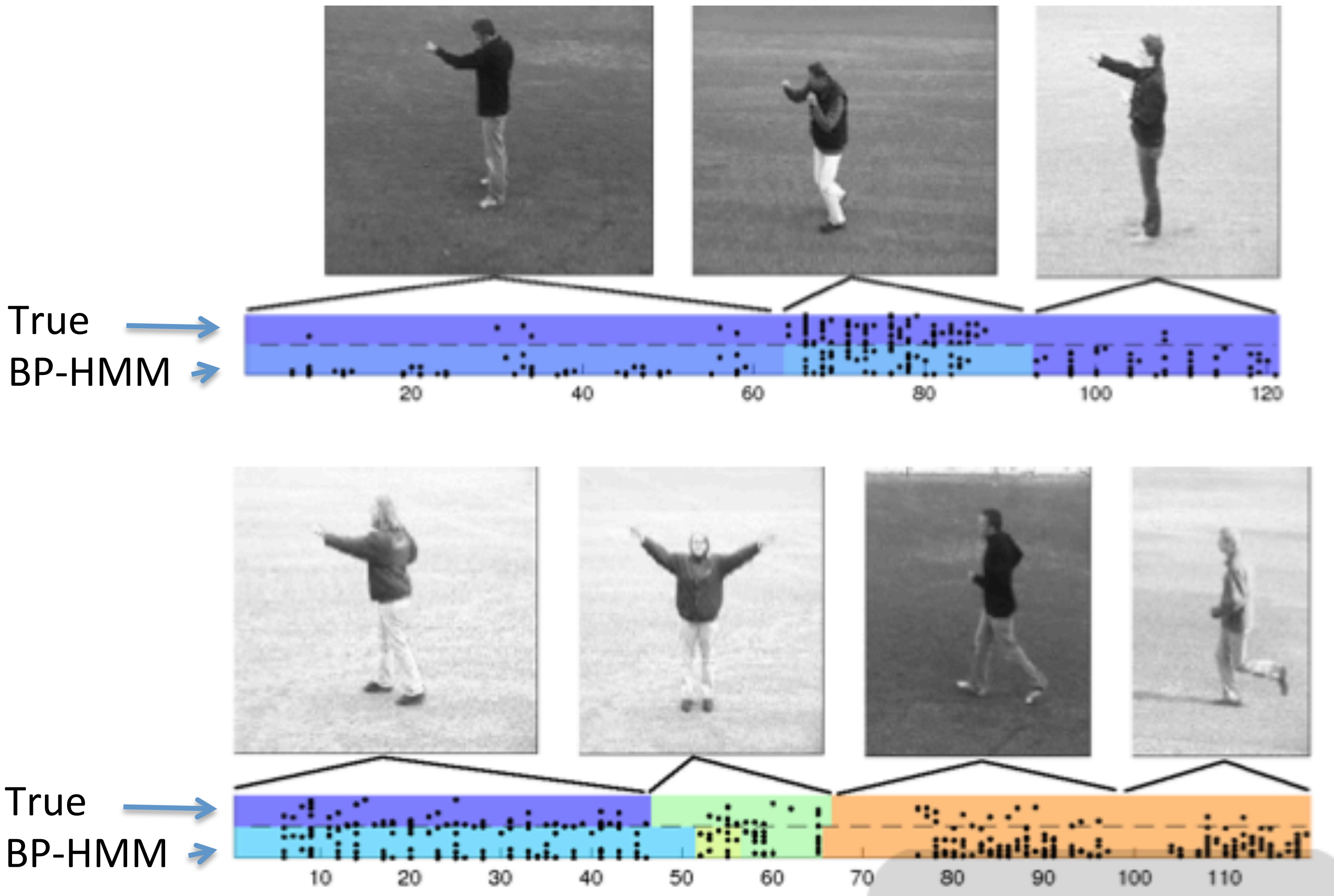
2 behaviors for waving (up, down)

Experiment 1b: KTH Mashup

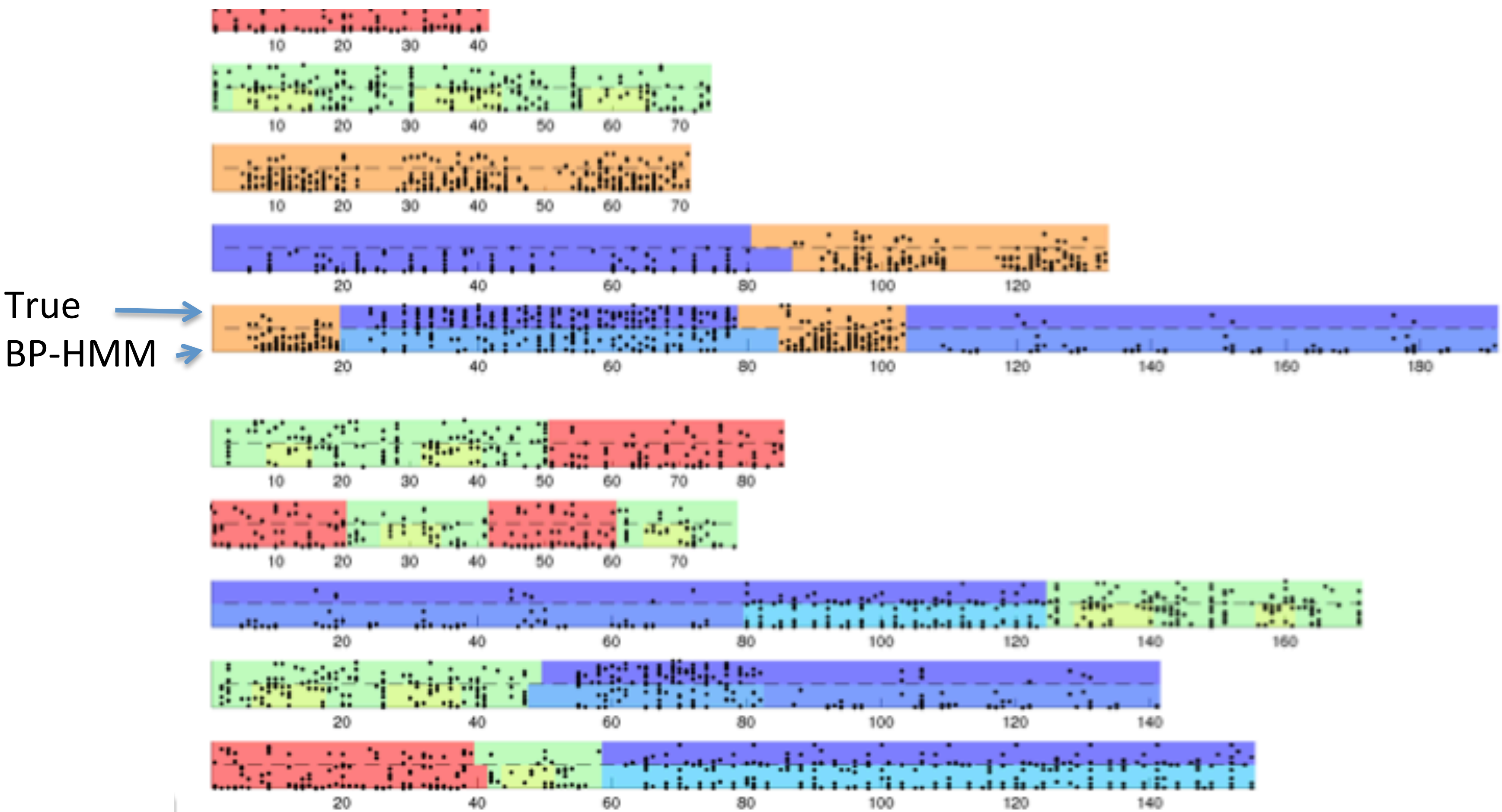


BP-HMM Binary Feature Vector

Experiment 1b: KTH Mashup



Experiment 1b: KTH Mashup



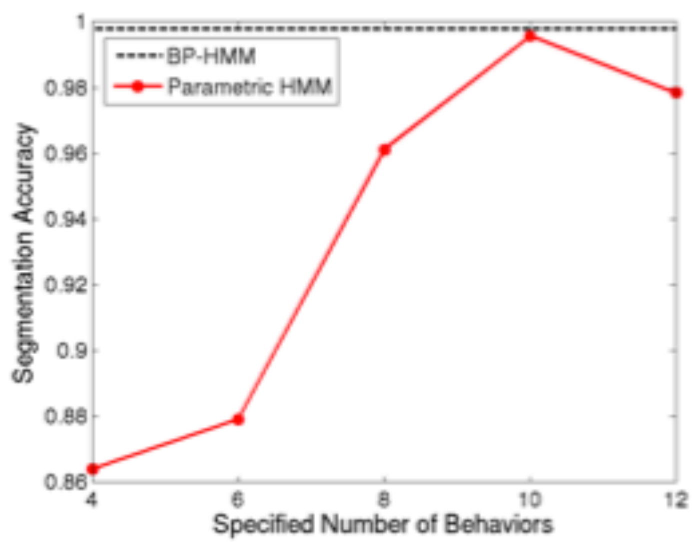
Experiment 1c: KTH Mashup

Hypothesis: BP-HMM segments better than HMM

Training: 12 sequences w/ mashups {box, clap, wave, jog}

Features: Only HOF, w=0.08 (2 frames)

Metric: Map estimated state to its closest true label, and then compute the number of timesteps where this relabeled estimate matches ground truth across all 12 sequences.



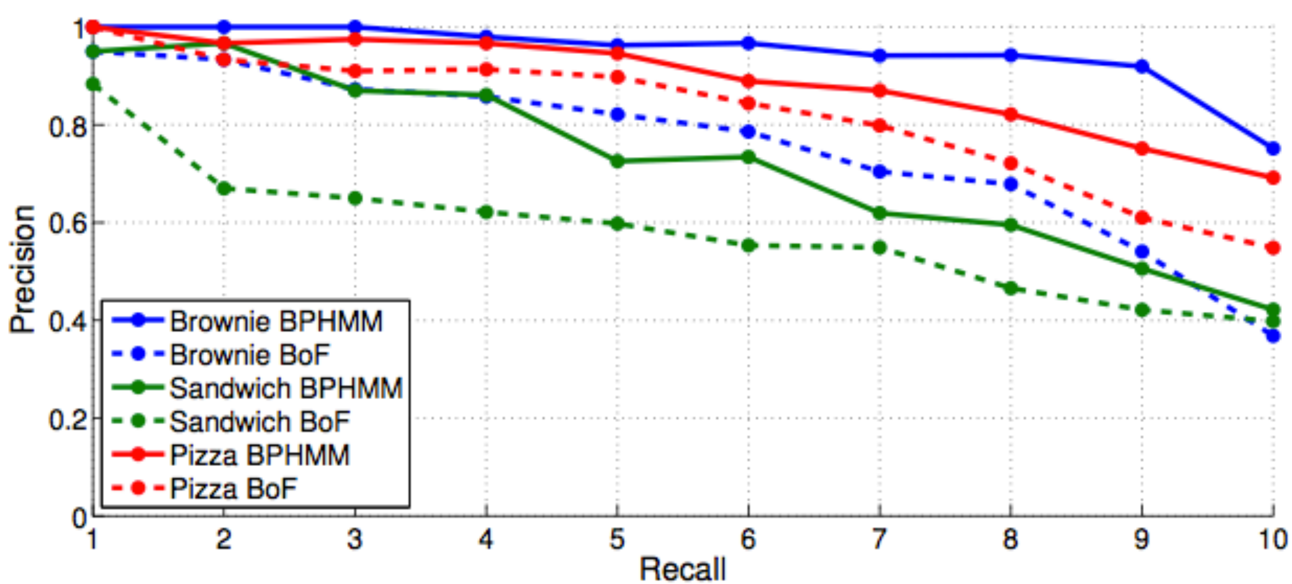
Experiment 2a: CMUKitchen

Hypothesis: (Retrieval) rank similarity of new video

Training: 10x3 sequences {Sandwich, Pizza, Brownie}

Features: HOG-HOF, w=0.5 (15 frames)

Summarize using histogram of timesteps per behavior



<u>F-score</u>

BP-HMM: 0.804

BOF: 0.703

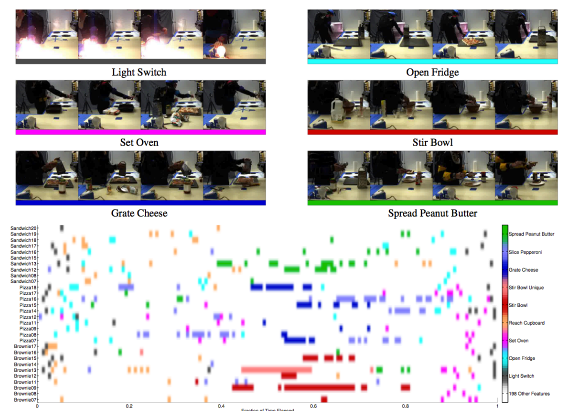
Experiment 2b: CMUKitchen

Hypothesis: BP-HMM can discovery intuitive behaviors

Training: 10x3 sequences {Sandwich, Pizza, Brownie}

Features: HOG-HOF, w=0.5 (15 frames)

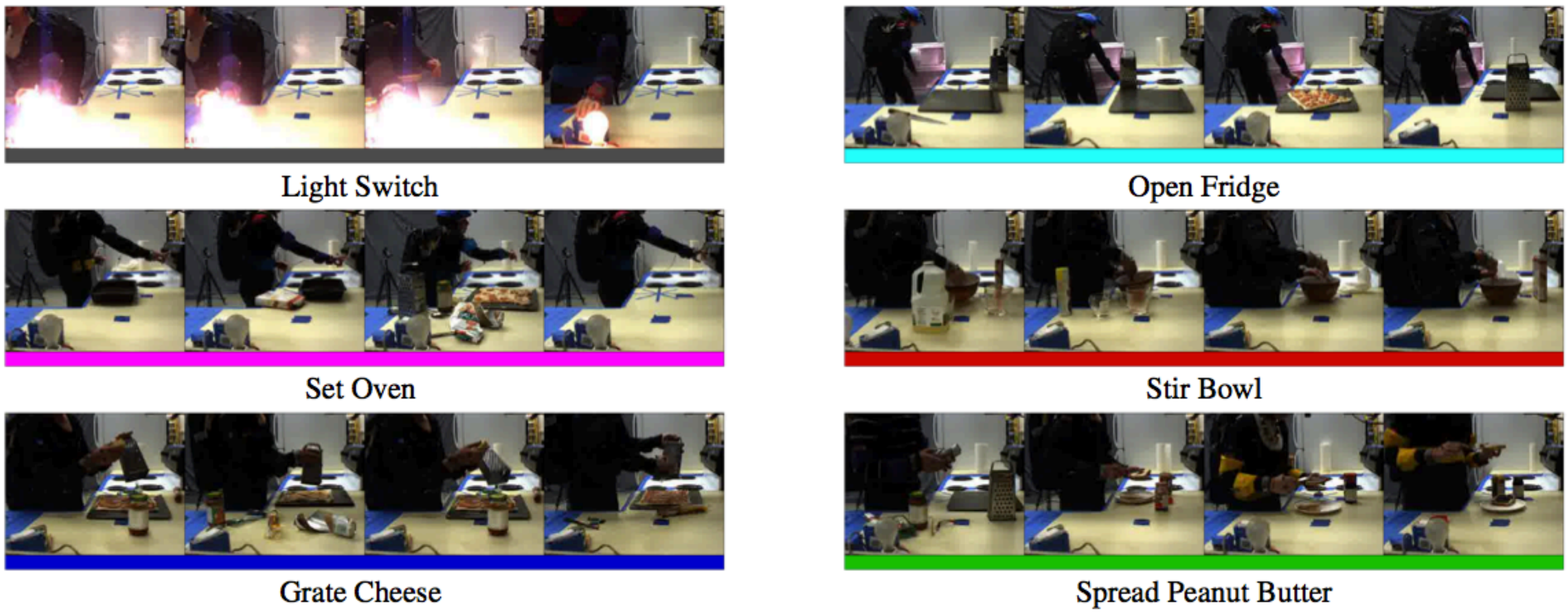
Summarize using histogram of timesteps per behavior

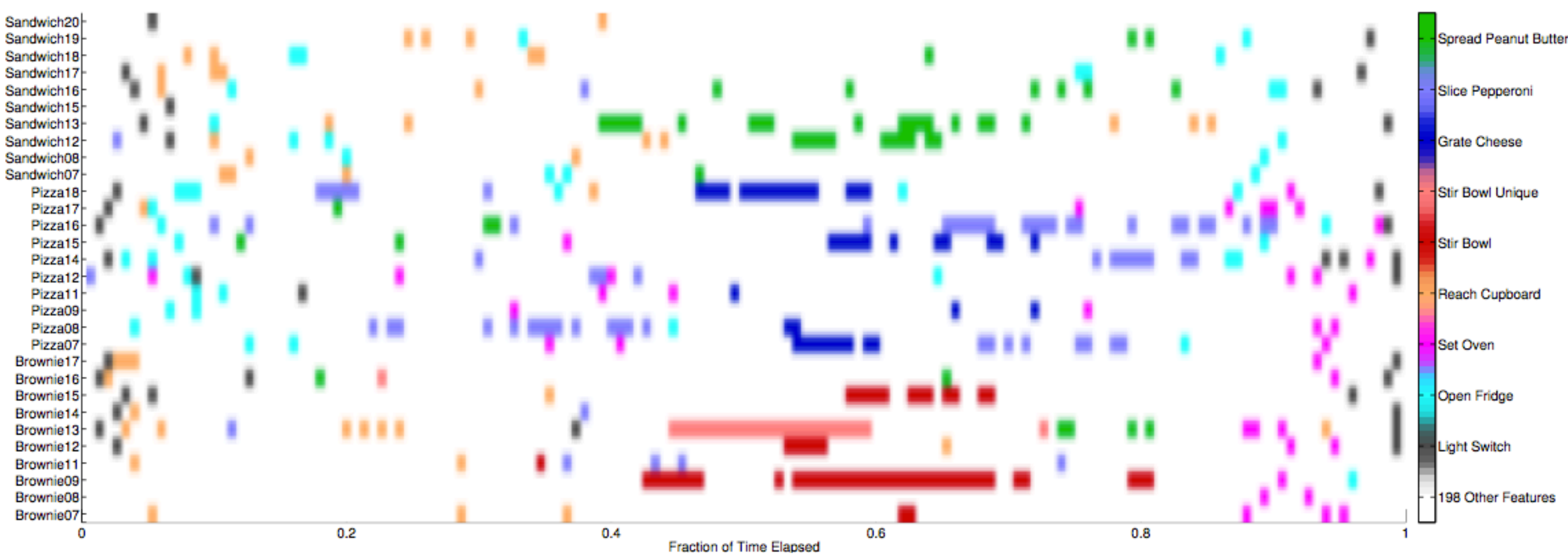


Results: Multiple features that correspond to a single behavior (e.g. stirring ingredients in a bowl)

(People can do actions in different styles!)

Experiment 2b: CMUKitchen





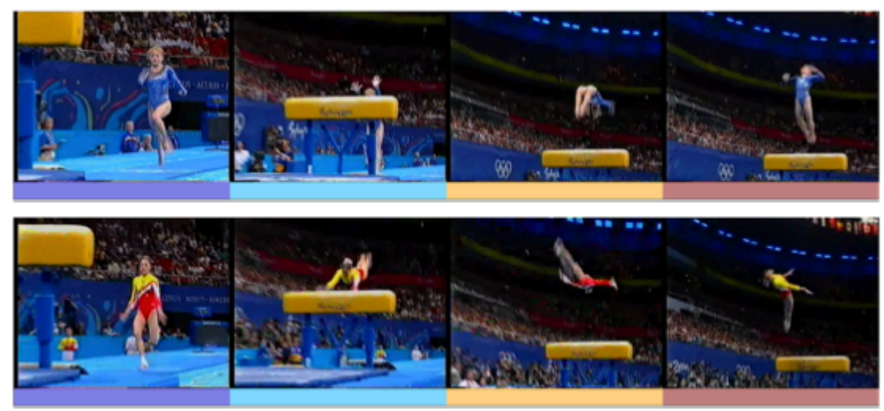
Experiment 3: Olympic Sports

Hypothesis: (Retrieval) rank similarity of new video

Training: 640 seq. x 16 actions **Testing:** 132 seq.

Features: Only HOF, w=0.16 (4 frames)

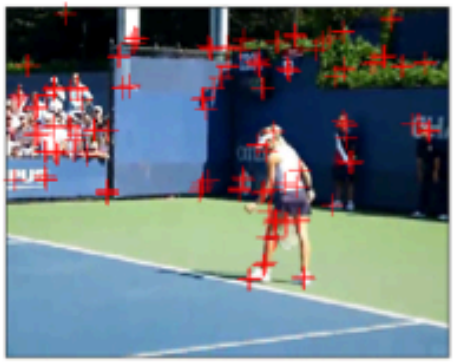
Train BP-HMM per-activity (b/c complexity)



F-score: BP-HMM=0.25, BoF=0.32

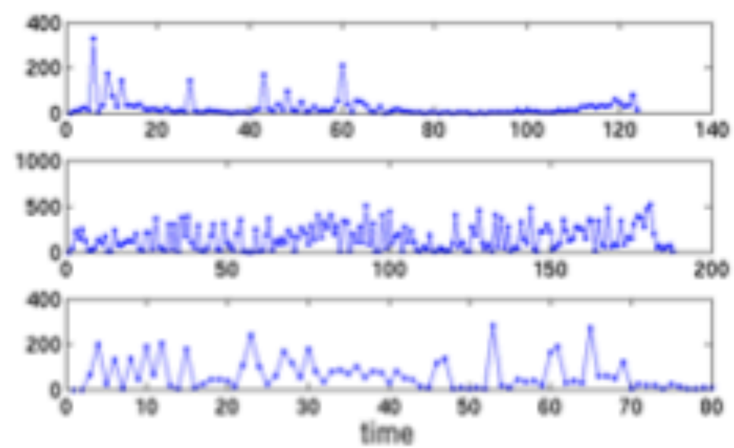
Results: Too much noise from features!

Experiment 3: Olympic Sports



Red crosses: STIP pts

Bad features!



STIPs per window

Takeaways

Good:

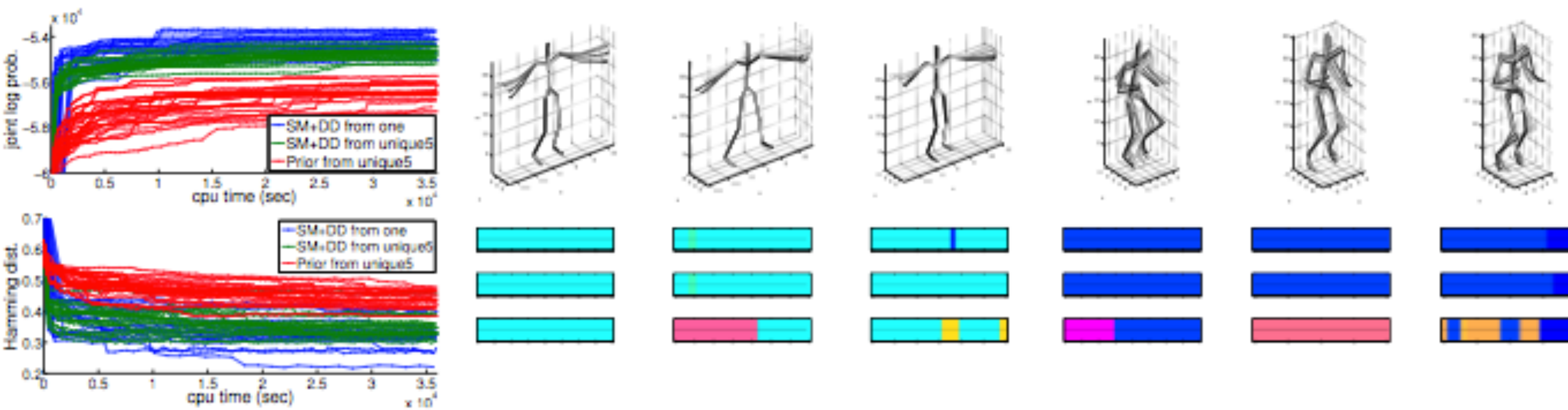
- Unsupervised discovery of intuitive behaviors
- Visualization of behaviors
- Shows shared dynamics is not helpful

Bad:

- Slow! (+Lack of discussion on complexity)

Subsequent paper: *Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data* [NIPS'12]

Data-driven Reversible MCMC -- Much faster!



Helpful resources

- Sudderth: NP Bayes CVPR’12 Tutorial
- Teh: Modern Bayes NP NIPS’11 Tutorial
- Dr. Nonparametric Bayes tutorial