A Non-parametric Hierarchical Model to Discover Behavior Dynamics from Tracks

Julian F. P. Kooij, Gwenn Englebienne, and Dariu M. Gavrila

Intelligent Systems Laboratory, University of Amsterdam, The Netherlands
{J.F.P.Kooij,G.Englebienne,D.M.Gavrila}@uva.nl

Abstract. We present a novel non-parametric Bayesian model to jointly discover the dynamics of low-level actions and high-level behaviors of tracked people in open environments. Our model represents behaviors as Markov chains of actions which capture high-level temporal dynamics. Actions may be shared by various behaviors and represent spatially localized occurrences of a person's low-level motion dynamics using Switching Linear Dynamics Systems. Since the model handles real-valued features directly, we do not lose information by quantizing measurements to 'visual words' and can thus discover variations in standing, walking and running without discrete thresholds. We describe inference using Gibbs sampling and validate our approach on several artificial and real-world tracking datasets. We show that our model can distinguish relevant behavior patterns that an existing state-of-the-art method for hierarchical clustering cannot.

1 Introduction

Computer vision and machine learning techniques can provide valuable tools for visual surveillance, aiding human operators in their task to monitor many video streams and focus their attention on possible incidents to make public spaces safer. In fixed-camera video surveillance, unsupervised learning techniques can be employed for anomaly detection by learning normative behavior from training data and detecting deviations thereof. A few issues arise when modeling behavior from observed low-level features. First, how is high-level behavior composed from low-level actions, second, where does specific behavior occur, and third, how can temporal dynamics of behavior be exploited. Ideally, action decomposition, spatial context, and temporal dynamics are jointly inferred from the training data. Related work has modeled behavior at the image level to capture patterns that govern the whole scene (e.g. monitoring traffic flow at junctions). We however target individual behavior patterns of people in open spaces, where execution of the same action may have large spatial and kinematic variations. In contrast to cars in driving lanes for instance, people can walk to the same destination along different parallel routes and move at specific or varying speeds such as standing, walking, or running. Existing methods do not properly account for such aspects, and rely on appropriate quantization of the feature space to distinguish or generalize over such variations. Further, unlike cars in traffic scenes, people in open environments generally behave independent of each other, thus instead of modeling all behaviors jointly at the image level we model people individually, using external tracker results.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 270–283, 2012.

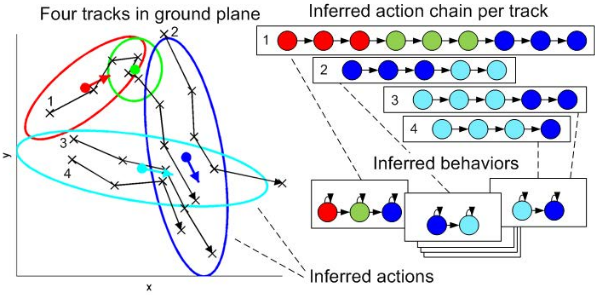


Fig. 1. Inferring a Mixture of Switching Linear Dynamic Systems from tracks, black crosses being observations. Trajectories are segmented into actions, each action being a semantic region (2D Gaussian) with a Linear Dynamic System for motion dynamics (mean motion shown as arrow). Behaviors cluster tracks with similar action chains. In this example four actions and three behaviors are inferred. The red and green actions distinguish walking and standing in track 1. Tracks 3 and 4 are clustered in the same behavior, but for track 2 another behavior is found with a different action order.

In this paper we propose a Mixture of Switching Linear Dynamic Systems to discover normative actions and their temporal relations at the object level. Actions describe low-level motion dynamics occurring in a semantic region using tracked person locations as features. As Figure 1 illustrates, our unsupervised approach segments tracks into sequences of common actions and jointly clusters the action sequences into distinct behavior classes. The number of actions and the number of behaviors are not fixed but discovered from the data itself. Key differences with previous approaches are that our hierarchical Bayesian model infers low-level actions and their temporal order within high-level behaviors directly from tracks, and that we use continuous distributions in the feature space to capture variance in action execution.

2 Previous Work

This section starts with an overview of recent developments in *topic models*, and then continues with their application for unusual behavior detection in video.

Latent Dirichlet Allocation (LDA) is a popular method for unsupervised discovery of topics in word corpora using a bag-of-word representation of documents [1]. LDA represents documents as mixtures over common topics, where each topic is a distribution over words. While LDA requires the number of topics to be known in advance, the Hierarchical Dirichlet Process (HDP) can be used instead to model an *infinite* number of topics [2], though only a finite number will be inferred. Dirichlet Processes (DPs) achieve such clustering into a finite amount of mixture components by using a *stick-breaking* construction (and a 'base' distribution over mixture components) [2]. Inference in HDPs is commonly achieved using Markov Chain Monte Carlo methods such as Gibbs sampling. The HDP can also be used to learn Infinite Hidden Markov Models (HDP-HMMs) to model HMMs where the number of states is inferred from the data itself [2, 3]. Just as HMMs can be extended to Switching Linear Dynamic System (SLDS), the HDP-HMM may be extended to HDP-SLDS [4]. A SLDS contains a top-level discrete Markov chain (the switching states), which determines the system dynamics and noise in the underlying LDS. Exact inference in a regular SLDS is intractable, but marginal distributions for Gibbs sampling of the switching states can be computed in linear time with information filters [5].

Topic models have been applied to visual behavior modeling by using quantized image features, such as optical flow, as ‘visual words’. Different techniques have been suggested to include temporal dynamics for video analysis to the bag-of-words approach. In [6] dynamics are modeled using a single Markov chain on top of LDA. The state of the chain determines the topic distribution at each frame. This approach is used to learn models for traffic junctions where the Markov chain captures the dynamics of traffic flow. This approach is extended by [7] to an infinite mixture of infinite Markov chains using a HDP, giving more flexibility than a single chain would. In [8] a combination is presented of HDP that finds, again, common optical flow topics and Probabilistic Latent Sequential Motifs [9], to represent actions as sequential patterns of topics up to a fixed length.

Unlike the previous methods which extract features at the image level, trajectory-based approaches use features of tracked objects instead. They do not assume that the joint object dynamics can reasonably be modeled at the image level. One approach is to use standard clustering methods with pair-wise distance measures on trajectories (such as Euclidean [10] or Hausdorff [11] distance) or Dynamic Time Warping [12]. The drawback is that these methods are not probabilistic, and the complexity of clustering N trajectories is $O(N^2)$ (c.f. [13]). Others try to segment the scene into semantically significant regions [14, 15], such as entry and exit points [11], or other regions where specific behavior can be observed. Semantic regions are useful to reduce the state space for modeling and classifying actions. The regions inferred by [16] describe optical flow motion dynamics using Lie algebra with Gaussian process and observation noise. However, their approach does not model long-term dependencies between low-level actions as the other models do [4, 6–8, 13].

Dual-HDP [13] extends HDP to hierarchically cluster bag-of-words representations of observed tracks, the words being quantized position / motion pairs. Jointly, words are clustered into semantic regions where common motion is found, and tracks are clustered into common mixtures of these regions. As a consequence of the bag-of-words approach, the temporal order of observations is not represented and feature quantization makes prior assumptions on what bins in the feature space are informative. If the binning resolution is too low details of the data are lost, but if the codebook size is too large then small variations in the tracks result in big variations in the bag-of-word representations. Since this trade-off is not explicitly represented in Dual-HDP, it can only be tackled by an extra external model selection procedure.

3 Model

We target scenarios with multiple people walking and waiting in open spaces. People may enter and exit the scene at different locations, though the system has no prior knowledge about these. Track data is obtained from an external person tracker, where each track is an ordered list of 2D positions on the ground plane. In our unsupervised approach, person tracks are clustered into *behaviors*, each behavior defining transition probabilities between *actions*, which we refer to in this paper as *topics*. Each topic describes for a common action the spatial location, and the low-level motion dynamics with an LDS. Topics can be shared among behaviors, thus multiple behaviors may contain the same topic but use different topic transition probabilities. Since each behavior is a SLDS, the full model forms a Mixture of SLDS. The temporal dynamics are especially helpful to distinguish behaviors with spatially overlapping actions. In Figure 1 for instance, tracks 2 and 3 have the same actions but different behaviors. Topic duration is captured by the behavior specific self-transition probability.

3.1 Contributions

Compared to previous work, our main contributions are (1) a hierarchical model to jointly infer low-level actions and high-level person behavior, inferring the number of actions and behaviors from the data itself, where (2) actions capture intuitive patterns and their variance directly in the continuous feature space, and (3) our model discriminates behaviors with different temporal action orders.

Dual-HDP [13] discovers hierarchical mixtures from quantized track features without capturing temporal order (i.e. bag-of-words). If high variance is present in the data, it also requires large amounts of data to avoid sparse bins. We instead infer the mean and variance of location and motion directly in the feature space with SLDS and Gaussian distributions. Further, while recent work in machine learning describes combining SLDS with HDP [4], the combination of SLDS with hierarchical track clustering is novel. In our model, behaviors induce higher-order dependencies between actions, as opposed to a single SLDS that only models first-order dependencies. Alternatively, one could infer behavior in ‘stages’, discovering actions first [4] and clustering them into behaviors later, but early commitment to estimated actions may lead to sub-optimal results. The benefits of Bayesian joint hierarchical clustering are well established and have popularized this approach for hierarchical activity learning [6–9, 11], as it can deal better with limited data, include priors, is robust against overfitting, and the high-level behavior clustering can inform the action clustering process. In the Supplemental Material we show that joint inference can use information from the high-level behaviors to find actions that explain the data better than those found by a single SLDS [4]. Such feedback during inference is not available in the ‘stages’ approach.

3.2 Hierarchical Clustering

This section describes hierarchical clustering in our model without the low-level motion dynamics. In Section 3.3 the model description will be extended to include these dynamics. The data consists of J tracks each being a sequence of N_j observations x_{ji}, with j the track index and i the time index. In our model the indicator variable $z_{ji} = k$ indicates that observation x_{ji} is sampled from topic k. To simplify notation we define $x_j = \{x_{j1}, ..., x_{jN_j}\}$, $z_j = \{z_{j1}, ..., z_{jN_j}\}$, and we denote the suffix $-i$ in z_j^{-i} to indicate all z_{ji} of track j except i. Each topic k defines of a probability distribution over the location on the ground plane (i.e. a semantic region) as a 2D Gaussian with parameters θ_k. Further, each track is assigned to a behavior, indexed by c_j, where a behavior c defines the topic transition probability $p(z_{ji}|z_{ji-1}) = \tilde{\pi}_c^{z_{ji-1}}$. The multinomial topic distributions $\tilde{\pi}_c^{z_{ji-1}}$ are sampled from a DP over a behavior-specific topic distribution π_c. The various π_c are sampled from a DP over π_0, which is the global topic distribution shared by all behaviors. The distribution π_0 follows a stick-breaking construction (Equation 1) and thus represents a multinomial distribution over infinite topics, although at any time during inference only some K topics will actually be used. Notice that in this multi-level hierarchical DP the distributions π_c assign non-zero probability to a subset of the topics represented in π_0, and the transition matrices $\tilde{\pi}_c$ are constrained to use those topics from π_c. The hierarchical clustering approach can be summarized as

$$\pi_0 \mid \delta \sim \text{Stick}(\delta) \qquad \pi_c \mid \pi_0, \alpha \sim \text{DP}(\alpha, \pi_0) \tag{1}$$

$$\tilde{\pi}_c^k \mid \pi_c, \beta \sim \text{DP}(\beta, \pi_c) \qquad z_{ji} \mid z_{ji-1}, c_j, \{\tilde{\pi}_c^k\} \sim \text{Mult}(\tilde{\pi}_{c_j}^{z_{ji-1}}) \tag{2}$$

$$x_{ji} \mid z_{ji}, \{\theta_k\} \sim \mathcal{N}(\theta_{z_{ji}}) \qquad \theta_k \mid \xi^{\Theta} \sim \mathcal{NW}^{-1}(\xi^{\Theta}) \tag{3}$$

where $\xi^{\Theta} = (\mu_0, \kappa, \nu, T)$ are the hyper-parameters for the Normal-Inverse-Wishart distribution.[1] Note how the above distributions define for each behavior c an HDP-HMM [2] with Gaussian observation likelihoods and the corresponding $\{\tilde{\pi}_c^k\}$ forming rows of the $K \times K$ transition matrix. Behavior labels c_j are sampled from the prior μ, a multinomial over the infinite number of behaviors which is also defined as a stick-breaking construction:

$$\mu \mid \gamma \sim \text{Stick}(\gamma), \quad c_j \mid \mu \sim \text{Mult}(\mu). \tag{4}$$

3.3 Low-Level Dynamics

The hierarchical model is extended with a SLDS on the labels z_j by introducing latent 2D-position variables y_{ji} for each observed position x_{ji}. Topics now not only define a distribution over the 2D space, but also the low-level dynamics of the position sequence. In fact, topic labels z_j form a Markov chain of switch variables which select the stochastic state dynamics and observation noise. The resulting SLDS is a Switching Kalman Smoother [17]:

$$y_{ji} = Ay_{ji-1} + q_{ji} \qquad q_{ji} \sim \mathcal{N}(m_{z_{ji}}, Q_{z_{ji}}) \tag{5}$$

$$x_{ji} = Cy_{ji} + r_{ji} \qquad r_{ji} \sim \mathcal{N}(0, R_{z_{ji}}) \tag{6}$$

Matrices A and C are fixed and determine the type of kinematics used. In our experiments we set A and C to identity, resulting in a fixed-velocity model where the learned velocity is captured in the mean of the process noise, $m_{z_{ji}}$. Using appropriate priors on the noise components, the updated model becomes

$$y_{ji} \mid y_{ji-1}, z_{ji}, \{m_k, Q_k\} \sim \mathcal{N}(Ay_{ji-1} + m_{z_{ji}}, Q_{z_{ji}}) \tag{7}$$

$$x_{ji} \mid y_{ji}, z_{ji}, \{R_k\}, \{\theta_k\} \sim \mathcal{N}(Cy_{ji}, R_{z_{ji}})\mathcal{N}(\theta_{z_{ji}}) \tag{8}$$

$$m_k, Q_k \sim \mathcal{NW}^{-1}(\xi^Q), \qquad R_k \sim \mathcal{W}^{-1}(\xi^R) \tag{9}$$

[1] The Normal-Inverse-Wishart is the conjugate prior for the mean and covariance parameters of a multivariate Normal distribution.

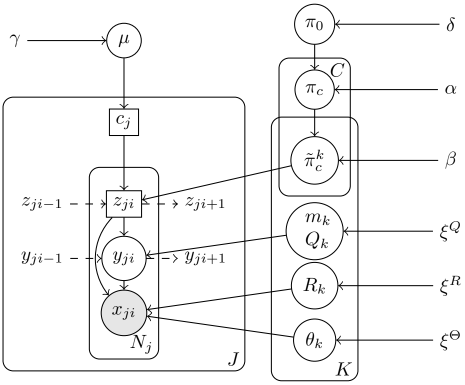


Fig. 2. Graphical model of our Mixture of SLDS. Square nodes are discrete, dashed arrows are temporal dependencies. Any time during inference K topics and C behaviors are represented. Track j consists of N_j observations x_{ji}, latent positions y_{ji}, topic label z_{ji} and has behavior label c_j drawn from behavior distribution μ. π_c is the topic distribution in behavior c, and $\tilde{\pi}_c^k$ are topic transition probabilities. Behaviors sample topics from the global distribution π_0. m_k, Q_k, R_k are the LDS and θ_k the (Gaussian) semantic region parameters of topic k.

with ξ^Q and ξ^R the hyper-parameters for a Normal-Inverse-Wishart and Inverse-Wishart distribution respectively. In Equation 8 the distribution over x_{ji} is factorized into a distribution over motion and location. The graphical model is represented in Figure 2.

4 Inference

Inference is achieved by defining the marginal distributions of each variable given its Markov Blanket (see Fig.2), and applying Gibbs sampling. Detailed derivations are given in the Supplemental Material. For convenience, let $m_j^{k'k}$ be the number of transitions from a state $z_{ji-1} = k'$ to $z_{ji} = k$ in track j, and $m_c^{k'k} = \sum_{j|c_j=c} m_j^{k'k}$ the total transition counts of behavior c. Also, we use z_c to denote all $z_{j'}$ with $c_{j'} = c$, and x^{-j}, z^{-j}, c^{-j} as respectively all observations, topic labels and behavior labels except those of track j. Distributions over μ, $\{y_{ji}\}$ and $\{\tilde{\pi}_c\}$ can be integrated out analytically during sampling, as we will see.

The posterior of each π_c is computed from the counts $m_c^{k'k}$ using the auxiliary variable sampling scheme [2]. The same scheme can then be applied to sample π_0.

Labels c_j are sampled one by one from the posterior, where the multinomial distributions μ and $\tilde{\pi}_c^k$ can be integrated out analytically such that

$$p(c_j|z_j, z^{-j}, \{\pi_c\}, c^{-j}) \quad \propto \quad p(z_j|z^{-j}, \{\pi_c\}, c^{-j}, c_j)p(c_j|c^{-j}). \tag{10}$$

Let n_c^{-j} be the occurrence count of behavior c in c^{-j}, then due to the stick-breaking prior (Equation 4),

$$p(c_j = c|c^{-j}) = n_c^{-j}/(J - 1 - \gamma), \tag{11}$$

and the likelihood term for $c_j = c$ in Equation 10 is computed as

$$p(z_j|z_c^{-j}, \pi_c, c_j = c) = \frac{p(z_j, z_c^{-j}|\pi_c, c^{-j}, c_j = c)}{p(z_c^{-j}|\pi_c, c^{-j}, c_j = c)}. \tag{12}$$

The terms in this fraction have the same form, and are obtained by integrating over $\tilde{\pi}_c^k$:

$$\begin{aligned} p(z_c|\pi_c) &= \int p(z_c|\tilde{\pi}_c^{k'k})p(\tilde{\pi}_c^{k'k}|\pi_c)\, d\tilde{\pi}_c^{k'k} \\ &= \prod_{k'} \frac{\Gamma(\beta)}{\Gamma(\beta + \sum_k m_c^{k'k})} \prod_k \frac{\Gamma(\beta\pi_c^k + m_c^{k'k})}{\Gamma(\beta\pi_c^k)}. \end{aligned} \tag{13}$$

The likelihood of assigning j to a unrepresented *new* behavior, c_{new}, is obtained by substituting n_c^{-j} with γ in Equation 11, and using the track likelihood of Equation 12:

$$p(z_j|z_{c_{\text{new}}}^{-j}, \pi_0, c_j = c_{\text{new}}) = \int \frac{p(z_j, z_{c_{\text{new}}}^{-j}|\pi_{c_{\text{new}}}, c_j = c_{\text{new}})}{p(z_{c_{\text{new}}}^{-j}|\pi_{c_{\text{new}}})} p(\pi_{c_{\text{new}}}|\pi_0)\, d\pi_{c_{\text{new}}}. \tag{14}$$

We find that sometimes assigning several similar tracks to a new behavior class would have high likelihood, but creating a new behavior first for a single track has low likelihood. Applying Blocking Gibbs sampling [18] to the relevant tracks would solve the problem, but determining which tracks to select is intractable. Instead, we approximate this effect by computing the probability of a new behavior using the likelihood of having the same track twice in the new behavior. We interpret Equation 14 therefore as if behavior c_{new} already contains track $\widehat{j}$ similar to j, thus $z_{c_{\text{new}}}^{-j} = z_{\widehat{j}}$, where the integral over $\pi_{c_{\text{new}}}$ is estimated by importance sampling:

$$p(z_j|z_{\widehat{j}}, \pi_0, c_j = c_{\text{new}}) = \int \frac{p(z_j|z_{\widehat{j}}, \pi_{c_{\text{new}}}, c_j = c_{\text{new}})}{p(z_{\widehat{j}}|\pi_{c_{\text{new}}})} p(\pi_{c_{\text{new}}}|\pi_0)\, d\pi_{c_{\text{new}}}. \tag{15}$$

Next, the topic labels z_{ji} are sampled with all c_j and K fixed. This is done efficiently in $O(N_j K)$ using forward and backward information filters for SLDS [5]. For the likelihood of a new topic we estimate the integral over the topic parameters by importance sampling. The information filters also give the distribution $p(y_{ji}|x_j^{-i}, z_j)$ as $\mathcal{N}(y_{ji}|\mu_{ji}, \Sigma_{ji})$ with parameters μ_{ji}, Σ_{ji}. Instead of sampling values y_{ji}, their posteriors are used directly to compute the posteriors of m_k, Q_k, R_k and θ_k. For instance, the prior of R_k is $\mathcal{W}^{-1}(R_k|\xi^R) = \mathcal{W}^{-1}(R_k|\nu^R, T^R)$, and the likelihood is $\mathcal{N}(x_{ji}|y_{ji}, R_k)$, thus the posterior will also be an Inverse-Wishart distribution,

$$p(R_k|x_{ji}, y_{ji}) = \mathcal{W}^{-1}(R_k|\nu^R + n, T^R + \widehat{S_R}) \tag{16}$$

where n is the number of observations with topic k, and $\widehat{S_R} = \sum_i (x_{ji} - y_{ji})(x_{ji} - y_{ji})^\top$ is N_j times the estimated sample covariance matrix (i.e. the *scatter matrix*). With Equation 6 and integrating out y_{ji}, it then follows that $x_{ji} \sim \mathcal{N}(\mu_{ji}, R_k + \Sigma_{ji})$, thus for the posterior of R_k we estimate $\widehat{S_R}$ as

$$\widehat{S_R} = \sum_i \left[(x_{ji} - \mu_{ji})(x_{ji} - \mu_{ji})^\top - \Sigma_{ji}\right]. \tag{17}$$

While this approach avoids additional sampling of all y_{ji}, the matrix $\widehat{S_R}$ can be an inaccurate estimate of the noise covariance when few observations are assigned to topic k, and it may happen that $\widehat{S_R}$ is not positive-semidefinite as required. To resolve these rare cases we do an Eigen-decomposition $\widehat{S_R} = W \Lambda W^{-1}$, set any negative Eigenvalue in Λ to zero, and recompose $\widehat{S_R}$ again. In the worst case all elements in $\widehat{S_R}$ will be set to zero, and sampling from the posterior of R_k reduces to sampling from its prior only.

The full complexity of sampling is $O(NK + CK^2) = O(N)$, N being the total number of observations. Given normative training data x^{-j}, anomaly detection requires a measure of 'normality' for unseen tracks x_j to rank these tracks or to set a threshold to isolate unusual from normal tracks. One possible measure is the normalized (to compensate for the track length) log-likelihood [13]: $\log(p(x_j|x^{-j}))/N_j$. We propose $\log \mathbb{E}_{c_j}[\min_i p(x_{ji}|x^{-j})]$, a different measure which penalizes unusual temporal transitions more. Both measures are estimated from Gibbs samples of the model posterior.

5 Experiments

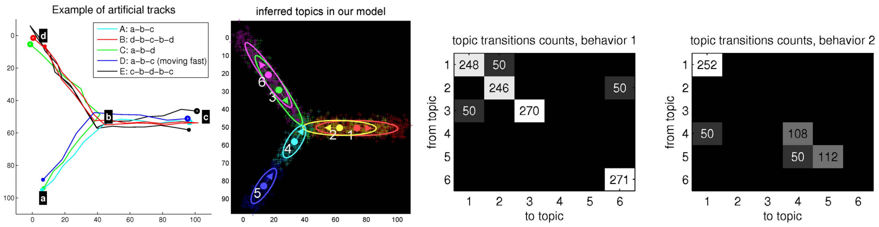
We compare our model with Dual-HDP [13] on an artificial dataset, a novel real-world indoor surveillance dataset,[2] and an existing publicly available pedestrian dataset. Like our method, Dual-HDP is a state-of-the-art hierarchical Bayesian non-parametric model for *track* data, and is therefore best suited for comparison. On the challenging surveillance dataset we also compare to Dynamic Time Warping (DTW) [12] for anomaly detection, as it is commonly used to compare tracks with varying action durations.

5.1 Artificial Dataset

Artificial track data is created by defining four waypoints in the 2D ground plane, and five behavior classes (labeled A to E) as ordered lists of these waypoints, see Fig. 3(a). For a behavior class tracks are generated by first sampling observations at the waypoints with added Gaussian noise ($\Sigma = \left[\begin{smallmatrix} 10 & 0 \\ 0 & 10 \end{smallmatrix}\right]$). Then, intermediate observations are created along this track at a speed of 10 units per time step (15 for behavior D), adding again Gaussian noise at each location ($\Sigma = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$). We generate a training dataset containing only 50 tracks from A and 50 from B. Then a test set is constructed with 20 tracks from each of the five behaviors. Tracks in C coincide partially with both A and B, those in D follow the same route as A but move faster. Tracks in E move in reverse direction of B.

For Dual-HDP we quantize observations into a 20 × 20 position grid and 4 directions, as in [13]. We generated 500 samples with each model, keeping every 25th sample (after 700 burn-in iterations for our model, and 2000 for Dual-HDP). Both models infer two behavior classes from the training data, corresponding to the tracks from A and B. Fig. 3(b) shows the topics from our model, and Fig. 3(c) the topic transition counts in the two found behaviors. Table 1 shows the normality ratings on the test data per behavior class. Our model assigns high likelihood to the novel tracks from the normative classes A and B, but anomalous tracks from C, D and E have low likelihood and could

[2] The dataset and other supplementary material is made available for non-commercial research purposes. Please follow the links from `http://isla.science.uva.nl/` or contact the authors.



(a) (b) (c)

Fig. 3. (a) Example tracks generated from the five different behaviors in the artificial dataset. The lower-case letters in the plot indicate the waypoints that define the behaviors. (b) Topics found by our model in training data of A and B. For each topic the assigned topic label, Gaussian semantic area, and mean system motion (the arrows) are shown in random colors. (c) The corresponding topic transition counts for the two behaviors our model found, e.g. in inferred behavior 1 there are 248 self transitions of topic 1 and 50 transitions to topic 2. Behavior 1 corresponds to topic chain 3-1-2-6 (matches B), behavior 2 to the chain 5-4-1 (matches A).

be separated by a threshold. Dual-HDP can also distinguish A and B from C, but not A and B from D and E, for the following reasons: first, as motion is only quantized into discrete directions [13], the speed differences between D and A are not recognized. This could be improved by quantizing motions at different speeds too, but this introduces again the problem of determining appropriate bins and increases the codebook size. Second, with the bag-of-words approach the unusual temporal order of E cannot be distinguished from B. We also apply both models to the combined tracks of all five behaviors. In addition to inferring new behavior for C, our model also separates tracks B and E into distinct behaviors. Depending on the prior on system motion, different topics (spatially coinciding but with different LDSs) can be found for A and D. Thus they are assigned to a single behavior group with topics allowing varying speeds, or to distinct behavior groups with topics for more specific speeds. Dual-HDP, however, finds only three behaviors: one corresponding to A+D, one for B+E, and a third for C.

5.2 Real-World Indoor Surveillance Dataset

We recorded a novel dataset at the central hall of a large building, using actors to perform various roles that commonly occur there at a normal working day (Fig. 4(a)). A recording session was held to capture normative behavior, which includes employees entering at the main entrance and walking to one of the exits, and visitors that register at the reception and wait to be picked up by an employee. Then an abnormal scenario was recorded which mostly contains normative behavior with some exceptions: in the scene a ‘terrorist’ scouts the environment, walking in and out of view. Later, a second ‘terrorist’ joins him and mixes with the visitors. When a security guard confronts the first ‘terrorist’, the second one shoots the guard, and bystanders run for safety.

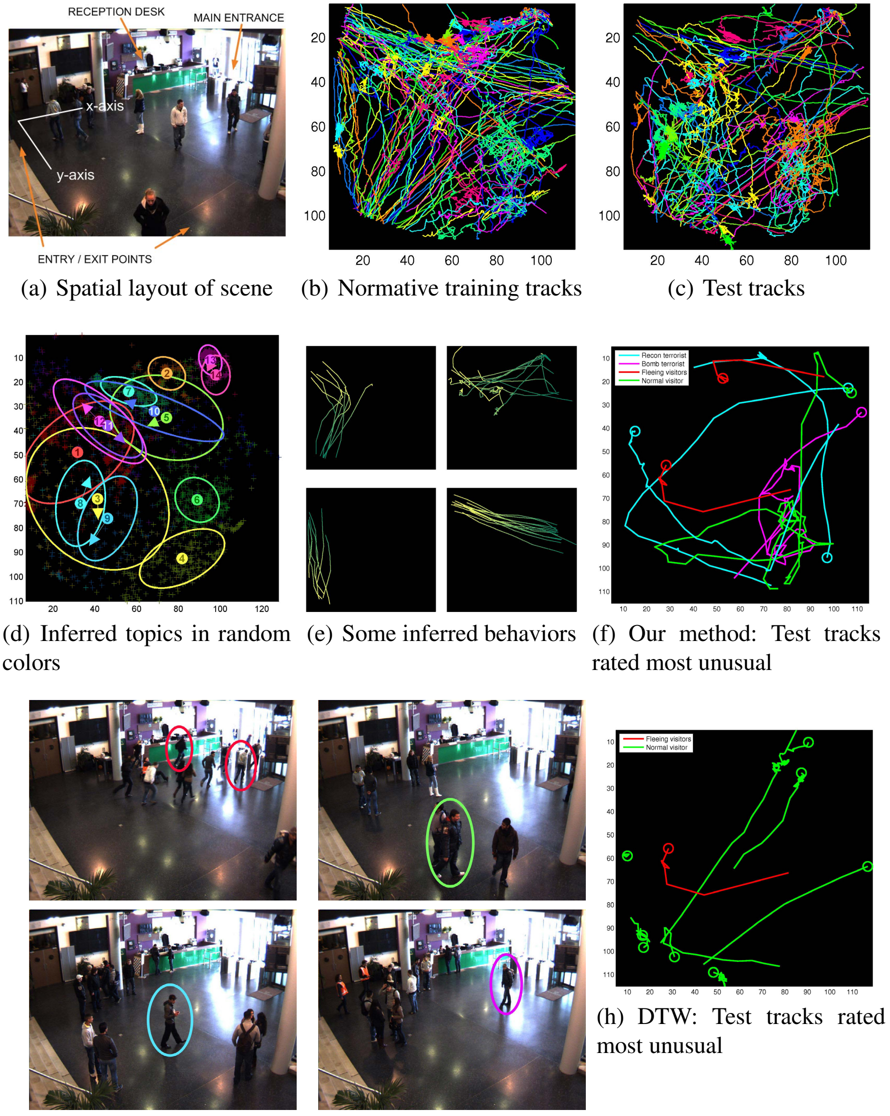
A multi-view tracker [19] is used to automatically detect and track people in the scene, but with many people in the scene not everyone is correctly tracked all the time. In a pre-processing stage we subsample tracks to one observations per second. We noticed

Table 1. Track ratings per ground truth behavior, after training on only tracks from A and B (see Figure 3). The test set contained 20 random tracks per behavior. Note that while both models can distinguish C as unusual, our model assigns low likelihood to tracks from D and E as well.

	$\log(p(x_j\|x^{-j}))/N_j$				$\log \mathbb{E}_{c_j}[\min_i p(x_{ji}\|x^{-j})]$			
	Mean	Var	Min	Max	Mean	Var	Min	Max
	Our model							
A	-10.4	0.2	-11.4	-9.7	-13.8	2.5	-17.1	-12.0
B	-10.6	0.1	-11.2	-10.0	-15.0	2.1	-17.8	-12.8
C	-14.1	0.2	-15.3	-13.6	-24.1	0.5	-25.8	-22.9
D	-16.9	2.4	-20.7	-14.4	-25.6	22.9	-37.7	-20.6
E	-10.9	0.3	-12.5	-10.3	-19.8	1.9	-23.5	-18.4
	Dual-HDP							
A	-4.8	0.2	-5.9	-4.2	-11.6	4.1	-15.5	-9.7
B	-5.2	0.1	-6.3	-4.8	-13.3	3.5	-16.3	-10.8
C	-7.4	0.2	-8.8	-6.9	-24.5	18.4	-36.4	-18.2
D	-4.8	0.1	-5.4	-4.5	-10.4	1.2	-13.5	-9.4
E	-5.5	0.3	-7.3	-4.9	-14.9	2.7	-17.1	-11.8

empirically that, due to wrong initializations in this challenging environment, the tracker created a number of short tracks that did not correspond to actual people. To remove such noise all tracks shorter than 5 seconds were discarded and we kept only the longer tracks (though still many are incomplete or truncated). The resulting training data of 118 tracks is shown in Fig. 4(b), the test data in 4(c) contains 64 tracks.

We apply our model on the training data and discover various actions, shown in Fig. 4(d). For instance, we interpret topic 13 as the action 'waiting at reception desk', and topic 9 as 'walking to lower exit'. In Fig. 4(e) tracks from 4 of the 24 inferred behaviors are shown (10 behaviors contain 75% of the tracks), corresponding to workers picking up visitors, visitors entering at the reception and waiting in the hall, and workers walking straight from the main entrance to an exit. The model is then used to rank the test tracks, the lowest ranked ones are shown in Fig. 4(f). The two most unusual tracks are of visitors fleeing to the main entrance after the gunshot. Two other tracks belong to the first 'terrorist', walking through common areas in an unusual order, and the second 'terrorist' who walks directly from the main entrance to a visitors area. A false positive occurs as one of the normal visitors walks to a different area to chat with others, something which did not occur in the training data. Fig. 4(g) shows several screenshots corresponding to these events. We also applied Dual-HDP [13] on this dataset, trying both high and low binning resolutions, but could not reliably distinguish normal test tracks from the anomalous ones. As it turns out, even if a low resolution of 10×10 spatial and 5 motion bins (4 directions + no motion) is used, 33 out of 64 test tracks contain words not seen in the training data. While in this particular scenario the Dual-HDP approach might be more successful if more training data were available, it again shows the problem of quantizing tracks with some variability. For additional comparison of anomaly detection on this dataset, we also applied a standard implementation of Dynamic Time Warping (DTW) [12]. Recall from Section 2 that DTW can be used to create a pair-wise distance measure for tracks that is robust against difference in



(a) Spatial layout of scene (b) Normative training tracks (c) Test tracks

(d) Inferred topics in random colors (e) Some inferred behaviors (f) Our method: Test tracks rated most unusual

(h) DTW: Test tracks rated most unusual

(g) Our method: screenshots from unusual test tracks

Fig. 4. (a) Screenshot from the real-world tracking dataset. (b) Tracks (in random colors) in the normative training data. When people stand still their tracks form dense spots. (d) Topics found by our method (in random colors). Shown are the topic label, Gaussian semantic region, and mean motion. (e) Several found behaviors (tracks start green, end yellow). (c) Tracks in the test data, containing normal people, suspicious individuals, and people running after a gunshot. (f) The most unusual tracks in the test data, circles mark the start of each track. (g) Screenshots of unusual tracks. Top: fleeing visitors; wandering visitor (false positive). Bottom: ‘terrorists’ exploring the space. (h) Dynamic Time Warping: most unusual tracks in the test data. The 9 most unusual tracks are false positives, the 10th track is a fleeing visitor.

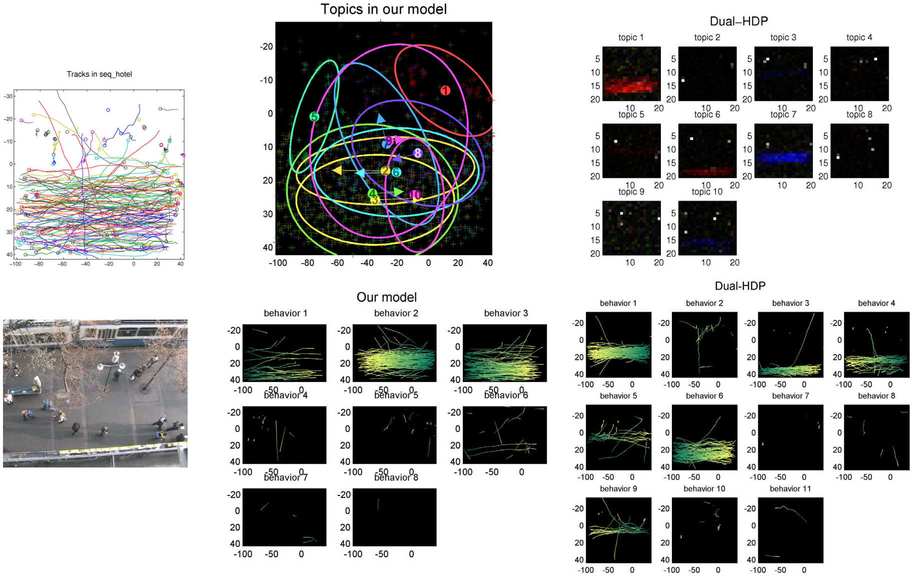


Fig. 5. (Left column) Tracks (in random colors) and screenshot from the BIWI dataset seq_hotel [20]. (Center column) Topics (in random colors) and behaviors found by our model. (Right column) Topics and behaviors found by Dual-HDP. In the topic images motion is color coded: → red, ↓ green, ← blue, ↑ yellow, 'no motion' white. The tracks in the behavior classes start green and end yellow to illustrate motion.

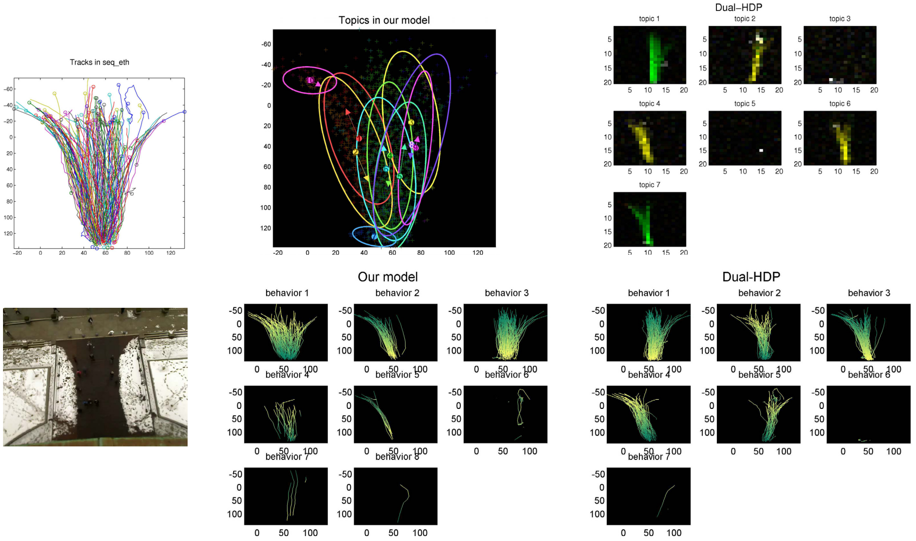


Fig. 6. (Left column) Tracks (in random colors) and a screenshot from the BIWI dataset seq_eth [20]. (Center column) Topics (in random colors) and behaviors found by our model. (Right column) Topics and behaviors found by Dual-HDP. In the topic images motion is color coded: → red, ↓ green, ← blue, ↑ yellow, 'no motion' white. The tracks in the behavior classes start green and end yellow to illustrate motion.

execution speed. The distance measure we use is the root-mean-square (RMS). Using DTW we compute for all test tracks the distance to each train track and rank the test tracks by the lowest distance. In Fig. 4(h) the 10 most unusual tracks according to the DTW ranking are shown. The first 9 most unusual tracks are all false positives, due to partial tracks and spatial offsets not found in the training data.

5.3 BIWI Walking Pedestrians Dataset

We have also made a qualitative comparison of our model and Dual-HDP [13] on the publicly available *BIWI Walking Pedestrians dataset* [20]. The dataset contains tracks from two top-down video sequences, namely seq_hotel, containing pedestrians walking along a sidewalk and some waiting for and entering a tram, and seq_eth which shows people entering and exiting the building from where the video was shot. We applied Dual-HDP and our own model on both datasets. For Dual-HDP the space is quantized into 20×20 grid cells, and motion into 5 cells (four directions, as in [13], and an extra 'no motion' bin, for people standing still). Inferred behaviors and topics by both models are shown in Fig. 5 (seq_hotel) and Fig. 6 (seq_eth). As we can expect from the videos, we see that both methods discover topics and behaviors that correspond to people walking in straight lines or standing and waiting. Accordingly, on these datasets the benefit of capturing action dynamics within behaviors (as our model does) is minimal, which explains why similar behavior classes are found in both models.

We do nevertheless observe some clear differences between the topics of both methods. First, in Dual-HDP, waiting people are represented in the topics as 'no motion' at one or few spatial cells, as can be seen by the white dots in Fig. 5 (top right). These topics thus capture waiting at exactly those spatial positions, but do not generalize over people waiting in the near vicinity. Our model on the other hand infers waiting areas as 2D Gaussian distribution over space, such as topics 1 and 5 for seq_hotel (Fig. 5, top center), and topics 3 and 7 for seq_eth (Fig. 6, top center). Second, in seq_hotel our model found spatially overlapping topics for people walking at different speeds. In Figure 5 (bottom left) topics 3 and 4 in our model both describe people walking from left-to-right, but with different velocities. The same can be said of topics 6 and 2 for people walking right-to-left. These different typical speeds were found due to the prior distribution over the mean and variance of an action's low-level motion, not due to specifying a speed threshold for 'slow' or 'fast' movement. To summarize, in the quantized feature space of Dual-HDP the spatial variance of waiting people results in sparse spatial bins, and different motions can only be distinguished if the binning thresholds are set appropriately in advance, while our model infers mean and variance of location and motion in the continuous feature space.

6 Conclusions

We have presented a novel model that uses a Mixture of SLDS to jointly infer common actions and behaviors from track data. The results show that our approach is capable of inferring informative clusters even when limited data is available as it does not rely on feature quantization. Although we have focused on video surveillance scenarios, other applications outside the computer vision domain may benefit from our method too.

Acknowledgments. This research has received funding from the European Community's Seventh Framework Programme under grant agreement number 218197, the ADABTS project.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. of Machine Learning Research 3, 993–1022 (2003)
2. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. The Journal of the Acoustical Society of America 101(476), 1566–1581 (2006)
3. Beal, M.J., Ghahramani, Z., Rasmussen, C.E.: The infinite hidden Markov model. In: Advances in Neural Information Processing Systems, vol. 1, pp. 577–584 (2002)
4. Fox, E., Sudderth, E., Jordan, M., Willsky, A.: Bayesian nonparametric inference of switching dynamic linear models. IEEE Trans. on Signal Processing 59(4), 1569–1585 (2011)
5. Rosti, A.V., Gales, M.J.F.: Rao-Blackwellised Gibbs sampling for switching linear dynamical systems. In: Proc. of the ICASSP, vol. 1, p. I–809 (2004)
6. Hospedales, T., Gong, S., Xiang, T.: A Markov clustering topic model for mining behaviour in video. In: Proc. of the IEEE ICCV, pp. 1165–1172 (2009)
7. Kuettel, D., Breitenstein, M., Van Gool, L., Ferrari, V.: What's going on? Discovering spatiotemporal dependencies in dynamic scenes. In: Proc. of the IEEE CVPR, pp. 1951–1958 (2010)
8. Emonet, R., Varadarajan, J., Odobez, J.M.: Multi-camera open space human activity discovery for anomaly detection. In: Proc. of the IEEE AVSS, p. 6 (August 2011)
9. Varadarajan, J., Emonet, R., Odobez, J.M.: Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In: Proc. of the BMVC (2010)
10. Fu, Z., Hu, W., Tan, T.: Similarity based vehicle trajectory clustering and anomaly detection. In: Proc. of the ICIP, vol. 2, p. II–602 (2005)
11. Wang, X., Tieu, K., Grimson, E.: Learning Semantic Scene Models by Trajectory Analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part III. LNCS, vol. 3953, pp. 110–123. Springer, Heidelberg (2006)
12. Keogh, E., Pazzani, M.: Scaling up dynamic time warping for datamining applications. In: Proc. of the ACM SIGKDD, pp. 285–289 (2000)
13. Wang, X., Ma, K.T., Ng, G.W., Grimson, W.E.: Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In: Proc. of the IEEE CVPR, pp. 1–8 (2008)
14. Fernyhough, J., Cohn, A., Hogg, D.: Generation of Semantic Regions From Image Sequences. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, vol. 1065, pp. 475–484. Springer, Heidelberg (1996)
15. Makris, D., Ellis, T.: Automatic learning of an activity-based semantic scene model. In: Proc. of the IEEE AVSS, pp. 183–188 (2003)
16. Lin, D., Grimson, E., Fisher, J.: Learning visual flows: A Lie algebraic approach. In: Proc. of the IEEE CVPR, pp. 747–754 (2009)
17. Rauch, H., Tung, F., Striebel, C.: Maximum likelihood estimates of linear dynamic systems. AIAA Journal 3(8), 1445–1450 (1965)
18. Jensen, C., Kjærulff, U., Kong, A.: Blocking Gibbs sampling in very large probabilistic expert systems. International J. of Human Computer Studies 42(6), 647–666 (1995)
19. Liem, M., Gavrila, D.M.: Multi-person Localization and Track Assignment in Overlapping Camera Views. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 173–183. Springer, Heidelberg (2011)
20. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Proc. of the IEEE ICCV, pp. 261–268 (2009)