Globally Continuous and Non-Markovian Activity Analysis from Videos

He Wang[1,2]*and Carol O'Sullivan[1,3]

[1] Disney Research Los Angeles**, United States of America
[2] University of Leeds, United Kingdom
realcrane@gmail.com
[3] Trinity College Dublin, Ireland
carol.osullivan@scss.tcd.ie

Abstract. Automatically recognizing activities in video is a classic problem in vision and helps to understand behaviors, describe scenes and detect anomalies. We propose an unsupervised method for such purposes. Given video data, we discover recurring activity patterns that appear, peak, wane and disappear over time. By using non-parametric Bayesian methods, we learn coupled spatial and temporal patterns with minimum prior knowledge. To model the temporal changes of patterns, previous works compute Markovian progressions or locally continuous motifs whereas we model time in a globally continuous and non-Markovian way. Visually, the patterns depict flows of major activities. Temporally, each pattern has its own unique appearance-disappearance cycles. To compute compact pattern representations, we also propose a hybrid sampling method. By combining these patterns with detailed environment information, we interpret the semantics of activities and report anomalies. Also, our method fits data better and detects anomalies that were difficult to detect previously.

1 Introduction

Understanding crowd activities from videos has been a goal in many areas [1]. In computer vision, a number of subtopics have been studied extensively, including flow estimation [2], behavior tracking [3] and activity detection [4,5]. The main problem is essentially mining recurrent patterns over time from video data. In this work, we are particularly interested in mining recurrent spatio-temporal activity patterns, i.e., recurrent motions such as pedestrians walking or cars driving. Discovering these patterns can be useful for applications such as scene summarization, event counting or unusual activity detection. On a higher level, such patterns could be used to reduce the dimensionality of the scene description for other research questions.

Pattern finding has been previously addressed [6,7,4], but only either for the spatial case, a Markovian progression or local motifs. To consider temporal information in a global non-Markovian fashion, we propose a Spatio-temporal Hierarchical Dirichlet Process (STHDP) model. STHDP leverages the power of Hierarchical Dirichlet Process

* Corresponding Author, ORCID ID:orcid.org/0000-0002-2281-5679
** This work is mostly done by the authors when they were with Disney Research Los Angeles.

(HDP) models to cluster location-velocity pairs and time simultaneously by introducing two mutually-influential HDPs. The results are presented as activity patterns and their time-varying presence (e.g. appear, peak, wane and disappear).

Combined with environment information, our approach provides enriched information for activity analysis by automatically answering questions (such as what, where, when and how important/frequent) for each activity, which facilitates activity-level and higher-level analysis. The novelty and contributions of our work are as follows:

1. We present an unsupervised method for activity analysis that requires no prior knowledge about the crowd dynamics, user labeling or predefined pattern numbers.
2. Compared to static HDP variants, we explicitly model the time-varying presence of activity patterns.
3. Complementary to other dynamic HDP variants, we model time in a globally continuous and non-Markovian way, which provides a new perspective for temporal analysis of activities.
4. We also propose a non-trivial split-merge strategy combined with Gibbs sampling to make the patterns more compact.

1.1 Related Work

Activities can be computed from different perspectives. On an individual level, tracking-based methods [8,9] and those with labeled motion features [10,11] have been successful. On a larger scale, flow fields [2,12] can be computed and segmented to extract meaningful crowd flows. However, these methods do not reveal the latent structures of the data at the flow level well where trajectory-based approaches prove to be very useful [5,13,14,15]. Trajectories can be clustered based on dynamics [5], underlying decision-making processes [14] or the environment [15,13]. However, these works need assumptions or prior knowledge of the crowd dynamics or environment. Another category of trajectory-based approaches is unsupervised clustering to reveal latent structures [7,4,16,17]. This kind of approaches assumes minimal prior knowledge about the environment or cluster number. Our method falls into this category.

Non-parametric Bayesian models have been used for clustering trajectories. Compared to the methods mentioned above, non-parametric Bayesian models have been proven successful due to minimal requirements of prior knowledge such as cluster numbers and have thus been widely used for scene classifications [18,19], object recognition [20], human action detection [21] and video analysis [22,7]. Initial efforts on using these kinds of models to cluster trajectories mainly focused on the spatial data [7]. Later on, more dynamic models have been proposed [16,4,17]. Wang et al. [16] propose a dynamic Dual-HDP model by assuming a Markovian progression of the activities and manually sliced the data into equal-length intervals. Emonet et al. [4,23] and Varadarajan et al.[17] model time as part of local spatio-temporal patterns, but no pattern progression is modeled. The former requires manual segmentation of the data and assumes the Markovian property, which does not always apply and could adversely affect detecting temporal anomalies. The latter focuses on local continuity in time and cannot learn time activities well when chunks of data are missing.

Inspired by many works in Natural Language Processing and Machine Learning [24,25,26,27,28,29], we propose a method that is complementary to the methods above

in that we model time in a globally continuous and non-Markovian way. We thus avoid manual segmentation and expose the time-varying presence of each activity. We show how our method fits data better and in general more aligned with human judgments. In addition, our method is good at detecting temporal anomalies that could be missed by previous methods.

2 Methodology

2.1 Spatio-temporal Hierarchical Dirichlet Processes

Given a video, raw trajectories can be automatically estimated by a standard tracker and clustered to show activities, with each activity represented by a trajectory cluster. One has the option of grouping trajectories in an unsupervised fashion where a distance metric needs to be defined, which is difficult due to the ambiguity of the associations between trajectories and activities across different scenarios. Another possibility is to cluster the individual observations of trajectories, such as locations, in every frame. Since observations of the same activity are more likely to co-occur, clustering co-occurring individual observations will eventually cluster their trajectories. This problem is usually converted into a data association problem, where each individual observation is associated with an activity. However, it is hard to know the number of activities in advance, so Dirichlet Processes (DPs) are used to model potentially infinite number of activities. In this way, each observation is associated with an activity and trajectories can be clustered based on a *softmax* scheme (a trajectory is assigned to the activity that gives the best likelihood on its individual observations). During the data association, DPs also automatically compute the ideal number of activities so that the co-occurrences of observations in the whole dataset can be best explained by an finite number of activities. To further capture the commonalities among the activities across different data segments, Hierarchical DPs (HDPs) are used, where one DP captures the activities in one data segment and another DP on a higher level captures all possible activities.

To cluster video data in the scheme explained above, we discretize the camera image into grids, that discretizing a trajectory into locations. We also discretize the velocity into several subdomains based on the orientation so that each location also comes with a velocity component. Finally, we can model activities as Multinomial distributions of time-stamped location-velocity pairs $\{w, t\}$, $w = (p_x, p_y, p'_x, p'_y)$ where (p_x, p_y) is the position, (p'_x, p'_y) is the velocity and each $\{w, t\}$ is an *observation*. Given multiple data segments consisting of location-velocity pairs, we can use the HDP scheme explained above to cluster trajectories. In addition, our STHDP also has a temporal part. Consider that a time data segment is formed by all the time stamps of the observations associated with one activity, then the distribution of these time stamps reflect the temporal changes of the activity. Since these time stamps might come from different periods (e.g. an activity appears/disappears multiple times), we need a multi-modal model to capture it. Again, since we do not know how many periods there are, we can use a DP to model this unknown too, which can be captured by an infinite mixture of Gaussians over the time stamps. Finally, to compute the time activities across different time data segments, we also use a HDP to model time. The whole scheme is explained by a Bayesian model shown in Figure 1.

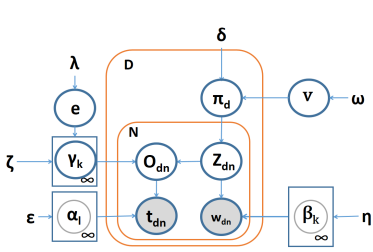

Fig. 1. STHDP model

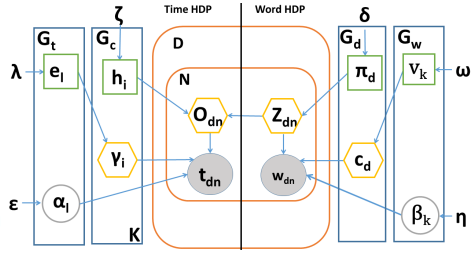

Fig. 2. Model used for sampling.

To mathematically explain our model, we first introduce some background and terminologies. In a *stick-breaking* representation [30] of a DP: $G = \sum_{k=1}^{\infty} \sigma_k(v)\beta_k$, where $\sigma_k(v)$ is *iteratively* generated from $\sigma_k(v) = v_k \prod_{j=1}^{k-1}(1 - v_j)$, $\sum_{k=1}^{\infty} \sigma_k(v) = 1$, $v \sim Beta(1, \omega)$ and $\beta_k \sim H(\eta)$. β_k are DP *atoms* drawn from some base distribution H and $\sigma_k(v)$ are *stick proportions*. We refer to the iterative generation of sticks $\sigma_k(v)$ from v as $\sigma \sim GEM(v)$, as in [24]. Following the convention of topics models, we refer to a location-velocity pair as a ***word***, its time stamp as a ***time word***, activity patterns as ***word topics*** and time activities as ***time topics***. A data segment is called a ***document*** and a time stamp data segment is called a ***time document***. The whole dataset is called a ***corpus***. The overall activities and time activities we are aiming for are the corpus-level word topics and time topics.

Figure 1 depicts two HDPs: a word HDP and a time HDP, respectively modeling the spatial and temporal data as described above. The word HDP starts with a DP over corpus-level word topics $v \sim GEM(\omega)$. In each document, there exists a DP $\pi_d \sim DP(v, \sigma)$ governing the document-level topics. For each word, a topic indicator is sampled by $Z_{d_n} \sim \pi_d$ and the word is generated from $w_{dn}|\beta_{Z_{dn}} \sim Mult(\beta_{Z_{dn}})$. The time HDP models how word topics evolve. Unlike previous models, it captures two aspects of time: continuity and multi-modality. Continuity is straightforward. Multi-modality means a word topic can appear/disappear several times. Imagine all the time words associated with the words under one word topic. The word topic could peak multiple times which means its time words are mainly aggregated within a number of time intervals. Meanwhile, there can be infinitely many word topics and some of their time words share some time intervals. Finally, there can be infinitely many such shared time intervals or time topics, which are modeled by an infinite mixture of Gaussians, of which each component is a time topic. A global DP $e \sim GEM(\lambda)$ governs all possible time topics. Then, for each corpus-level word topic k, a time DP $\gamma_k \sim DP(\zeta, e)$ is drawn. Finally, when a specific time word is needed, its Z_{d_n} indicates its word topic based on which we draw a time word indicator $O_{d_n} \sim \gamma_{Z_{d_n}}$ and a time word is generated from $t_{dn}|\alpha_{O_{d_n}} \sim Normal(\alpha_{O_{d_n}})$. In this way, each word topic corresponds to a subset of time topics with different weights. Thus a Gaussian Mixture Model (GMM) is naturally used for every word topic. Due to the space limit, the generative process of Figure 1 is explained in the supplementary material.

2.2 Posterior by Sampling

To compute the word and time topics we need to compute the posterior of STHDP. Both sampling and variational inference have been used for computing the posterior of hierarchical models [24,31]. After our first attempt at variational inference, we found that it suffers from the sub-optimal local minima because of the word-level coupling between HDPs. Naturally, we resort to sampling. Many sampling algorithms have been proposed for such purposes [25,32,33]. However, due to the structure of STHDP, we found that it is difficult to derive a parallel sampling scheme such as the one in [32]. Finally, we employ a hybrid approach that combines Gibbs and Metropolis-Hasting (MH) sampling based on the stick-breaking model shown in Figure 2, the latter being the split-merge (SM) operation. For Gibbs sampling, we use both Chinese Restaurant Franchise (CRF) [24] and modified Chinese Restaurant Franchise (mCRF) [25]. As there are two HDPs in STHDP, we fix the word HDP when sampling the time HDP which is a standard two-level HDP, so we run CRF sampling [24] on it. For the word HDP, we run mCRF. Please refer to the supplementary material for details.

HDPs suffer from difficulties when two topics are similar, as the sampling needs to go through a low probability area to merge them [34]. This is particular problematic in our case because each observation is pulled by two HDPs. Split-merge (SM) methods have been proposed [34,35] for Dirichlet Processes Mixture Models, but they do not handle HDPs. Wang et al. [26] proposes an SM method for HDP, but only for one HDP, whereas STHDP has two entwined HDPs. We propose a Metropolis-Hasting (MH) sampling scheme to perform SM operations. In our version of the CRF metaphor, word topics and time topics are called ***word dishes*** and ***time dishes***. Word documents are called ***restaurants*** and time documents are called ***time restaurants***. Some variables are given in Table 1. Similar to [26], we also only do split-merge on the word dish level. We start with the SM operations for the word HDP. In each operation, we randomly choose two word tables, indexed by i and j. If they serve the same dish, we try to split this dish into two, and otherwise merge these two dishes. Since the merge is just the opposite operation of split, we only explain the split strategy here.

Table 1. Variables in CRF

v_w	a word in the vocabulary
V_w	the size of the vocabulary
n_{jik}	the number of words in restaurant j at table i serving dish k
z_{ji}	the table indicator of the ith word in restaurant j
m_{jk}	the number of word tables in restaurant j serving dish k
$m_{j\cdot}$	the number of word tables in restaurant j
K	the number of word dishes

Following [34], the MH sampling acceptance ratio is computed by:

$$a(c^*, c) = min\{1, \frac{q(c|c^*)}{q(c^*|c)} \frac{P(c^*)}{P(c)} \frac{L(c^*|y)}{L(c|y)}\} \tag{1}$$

where c^* and c are states (table and dish indicators) after and before split and $q(c^*|c)$ is the split transition probability. The merge transition probability $q(c|c^*) = 1$ because there is only one way to merge. P is the prior probability, y are the observations, so $L(c^*|y)$ and $L(c|y)$ are the likelihoods of the two states. The split process of MH is: sample a dish, split it into two according to some process, and compute the acceptance probability $a(c^*, c)$. Finally, sample a probability $\phi \sim Uniform(0, 1)$. If $\phi > a(c^*, c)$, it is accepted, and rejected otherwise. The whole process is done only within the sampled dish and two new dishes. All the remaining variables are fixed.

Now we derive every term in Equation 1. The state c consists of the table and dish indicators. Because the time HDP needs to be considered when sampling the word HDP, the prior of table indicators is:

$$p(\mathbf{z}_j) = \frac{\delta^{m_{j\cdot}} \prod_{t=1}^{m_{j\cdot}} (n_{jt} p(t|\bullet) - 1)!}{\prod_{i=1}^{n_{j\cdot}} (i + \delta - 1)} \tag{2}$$

where $p(t|\bullet)$ represents the marginal likelihood of all time words involved. Similarly, for word dish indicators:

$$p(\mathbf{k}) = \frac{\omega^K \prod_{k=1}^{K} (m_{\cdot k} p(t|\bullet) - 1)!}{\prod_{i=1}^{m_{\cdot\cdot}} (i + \omega - 1)} \tag{3}$$

Now we have the prior for p(c):

$$p(c) = p(\mathbf{k}) \prod_{j=1}^{D} p(\mathbf{z}_j) \tag{4}$$

where D is the number of restaurants; $p(c^*)$ can be similarly computed.

Now we derive $q(c^*|c)$. Assume that tables i and j both serve dish k. We denote S as the set of indices of all tables also serving dish k excluding i and j. In the split state, k is split into k_1 and k_2. We denote S_1 and S_2 as the sets of indices of tables serving dishes k_1 and k_2. We first assign table i to k_1 and j to k_2, then allocate all tables indexed by S into either k_1 or k_2 by *sequential allocation restricted Gibbs sampling* [35]:

$$p(SK = k_j|S_1, S_2) \propto m_{\cdot k_j} f_{k_j}(\mathbf{w}_{SK}) p(\mathbf{t}_{SK}|\bullet) \tag{5}$$

where j = 1 or 2, $SK \in S$, $\mathbf{w}_{SK}$ is all the words at table SK and $m_{\cdot k_j}$ is the total number of tables assigned to k_j. All the tables in S are assigned to either k_1 or k_2. We still approximate $p(\mathbf{t}_{SK}|\bullet)$ by $\hat{p}(\mathbf{t}_{SK}|\bullet)$ as we do for Gibbs sampling (cf. supplementary material). Note that during the process, the sizes of S1 and S2 constantly change. Finally, we compute $q(c^*|c)$ by Equation 6:

$$q(c^*|c) = \prod_{i \in S} p(k^i = k|S1, S2) \tag{6}$$

Finally, the likelihoods are:

$$\frac{L(c^*|y)}{L(c|y)} = \frac{f_{k_1}^{lik}(\mathbf{w}_{k_1}, \mathbf{t}_{k_1}|c^*) f_{k_2}^{lik}(\mathbf{w}_{k_2}, \mathbf{t}_{k_1}|c^*)}{f_{k}^{lik}(\mathbf{w}_{k}, \mathbf{t}_{k}|c)} \tag{7}$$

where

$$f^{lik}(w,t|c) = \frac{\Gamma(V_w\eta)}{\Gamma(n_{..k}+V_w\eta)} \frac{\prod_{v_w}\Gamma(n_{..k}^{v_w}+\eta)}{\Gamma^{V_w}(\eta)} p(t|\bullet) \quad (8)$$

Γ is the gamma function, $n_{..k}$ is the number of words in topic k, $n_{..k}^{v_w}$ is the number of words v_w assigned to topic k, and $p(t|\bullet)$ is the likelihood of the time words involved.

Now we have fully defined our split-merge operations. Whenever an SM operation is executed, the time HDP needs to be updated. During the experiments, we do one iteration of SM after a certain number of Gibbs sampling iterations.

We found it is unnecessary to do SM on the time HDP for two reasons. First, we already implicitly consider the time HDP here through Equations 2 - 8 and an SM operation on the word HDP will affects the time HDP. Second, we want the word HDP to be the dominating force over the time HDP in SM. A merge operation on the word topics will cause a merge operation on the time HDP, which makes the patterns more compact. The reverse is not ideal because it can merge different word topics. However, this does not mean that the time HDP is always dominated. Its impact in the Gibbs sampling plays a strong role in clustering samples that are temporally close together while separating samples that are not.

3 Experiments

Empirically, we use a standard set of parameters for all our experiments. The prior $Dirichlet(\eta)$ is a symmetric Dirichlet where η is initialized to 0.5. For all GEM weights, we put a vague $Gamma$ prior, $Gamma(0.1, 0.1)$, on their $Beta$ distribution parameters, which are updated in the same way as [24]. The last is the Normal-Inverse-Gamma prior, $NIG(\mu, \lambda, \sigma_1, \sigma_2)$, where μ is the mean, λ is the variance scalar, and σ_1 and σ_2 are the shape and scale. For our datasets, we set μ to the sample mean, λ = 0.01, σ_1 = 0.3 and $\sigma_2 = 1$. Because both the Gamma and NIG priors are very vague here, we find that the performance is not much affected by different values, so we fix the NIG parameters. For simplicity, we henceforth refer to all activity Patterns with the letter P.

3.1 Synthetic Data

A straightforward way to show the effectiveness of our method is to use synthetic data where we know the ground truth so that we can compare learned results with the ground truth. Similar to [7], we use grid image data where the ground truth patterns (Figure 3 P1 and P2) are used to generate synthetic document data. Different from [7], to show the ability of our model in capturing temporal information, we generate data for 4 periods where the two ground truth patterns are combined in different ways for every period. Documents are randomly sampled from the combined pattern for each period. Figure 3 Right shows the learned patterns and their time activities from HDP [24] and STHDP. The time activities for HDP are represented by a GMM over the time words associated with the activity pattern. For STHDP, a GMM naturally arises for each pattern by combining the time patterns and their respective weights.

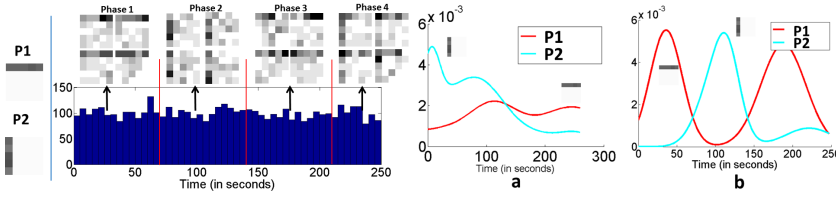

Fig. 3. Left: Two ground truth patterns on a 5 × 5 grid (one distributed over a horizontal bar, the other over a vertical bar) and generated data over four time periods (P1 is used for generating data in phase 1 and phase 3, P2 is used for data in phase 2 and both are used for phase 4. Some document examples and the histogram of observation numbers are shown for each phase). Right: Learned patterns and their time activities by (a) HDP [24] and (b) STHDP.

Both HDP and STHDP learn the correct spatial patterns. However, HDP assumes exchangeability of all data points thus its time information is not meaningful. In contrast, STHDP not only learns the correct patterns, but also learns a multi-modal representation of its temporal information which reveals three types of information. First, all GMMs are scaled proportionally to the number of associated data samples, so their integrals indicate their relative importance. In Figure 3 Left, the number of data samples generated from P1 is roughly twice as big as that from P2. This is reflected in Figure 3 Right (b) (The area under the red curve is roughly twice as big as that under the blue curve). Second, each activity pattern has its own GMM to show its presence over time. The small bump of the blue curve in (b) shows that there is a relatively small number of data samples from P2 beyond the 210^{th} second. It is how we generated data for phase 4. Finally, different activity patterns have different weights across all the time topics. Conversely, at any time, the data can be explained by a weighted combination of all activity patterns. Our method provides an enriched temporal model that can be used for analysis in many ways.

3.2 Real Data

In this section, we test our model on the Edinburgh dataset [36], the MIT Carpark database [16] and New York Central Terminal [13], referred to as **Forum**, **Carpark** and **TrainStation** respectively. They are widely used to test activity analysis methods [16,13,7,37,38]. Each dataset demonstrates different strengths of our method. Forum consists of indoor video data with the environment information available for semantic interpretation of the activities. Carpark is an outdoor scene consisting of periodic video data that serves as a good example to show the multi-modality of our time modeling. TrainStation is a good example of large scenes with complex traffic. All patterns are shown by representative (high probability) trajectories.

Forum Dataset The forum dataset is recorded by a bird's eye static camera installed on the ceiling above an open area in a school building. 664 trajectories have been extracted as described in [36], starting from 14:19:28 GMT, 24 August 2009 and lasting for 4.68 hours. The detailed environment is shown in Figure 4 (left). We discretize the 640 *

480 camera image into 50×50 pixel grids and the velocity direction into 4 cardinal subdomains and then run a burn-in 50 iterations of Gibbs sampling. For the first 500 iterations, MH sampling is done after every 10 Gibbs sampling iterations. Then we continue to run it for another 1500 iterations.

Nine patterns are shown in Figure 4, where the semantics can be derived by also considering the environment information in the form of Zones Z1: Stairs, Z2-Z7: Doors, Z8: Conference room, Z9: a floating staircase that blocks the camera view but has no semantic effect here. P3 and P4 are two groups of opposite trajectories connecting Z1 and Z2. We observe many more trajectories in P3 than P4. From the detailed environment information, we know that the side door outside of Z2 is a security door. This door can be opened from the inside, but people need to swipe their cards to open it from outside, which could explain why there are more exiting than entering activities through Z2. P2 is the major flow when people come down the stairs and go to the front entrance. P1 has a relatively small number of trajectories from Z6 to Z7, i.e., leaving through the front entrance. From the temporal point of view, the two major incoming flows can be seen in Figure 4 P4 and Figure 4 P5, spanning the first half of the data. We also spot a pattern with a high peak at the beginning (around 2:34pm), shown by Figure 4 P7, which connects the second part of the area and the conference room. We therefore speculate that there may have been a big meeting around that time.

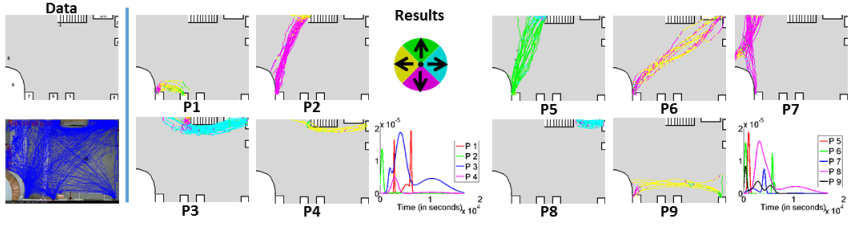

Fig. 4. Top Left: Environment of Edinburgh dataset. Bottom Left: Trajectories overlaid on the environment. Right: Some activities shown by representative trajectories and their respective time activities. Colors indicate orientations described by the legend in the middle.

Carpark Dataset The Carpark dataset was recorded by a far-distance static camera over a week and 1000 trajectories were randomly sampled as shown in Figure 5 Left. Since this dataset is periodic, it demonstrates the multi-modality of our time modeling. We run the sampling in the same way as in the Forum experiment.

Four top activity patterns and their respective time presence are shown in Figure 5. P1 is the major flow of in-coming cars, P2 is an out-going flow, and P3 and P4 are two opposite flows. Unfortunately, we do not have detailed environment information as we do from Forum for further semantic interpretations. The temporal information shows how all peaks are captured by our method, but different patterns have different weights in different periods.

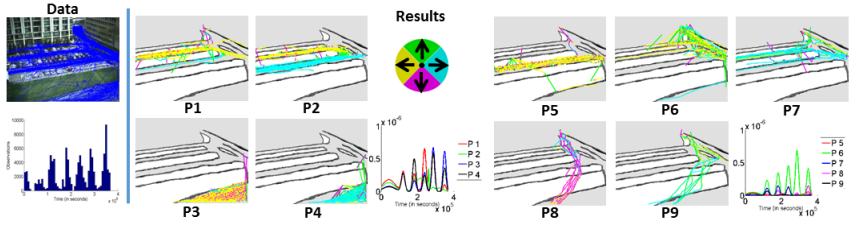

Fig. 5. Top Left: Environment of the car park. Bottom Left: Observation numbers over time. Right: Some activities shown by representative trajectories and their respective time activities. Colors indicate orientations described by the legend in the middle.

TrainStation Dataset The TrainStation dataset was recorded from a large environment and 1000 trajectories were randomly selected for the experiment. The data and activities are shown in Figure 6.

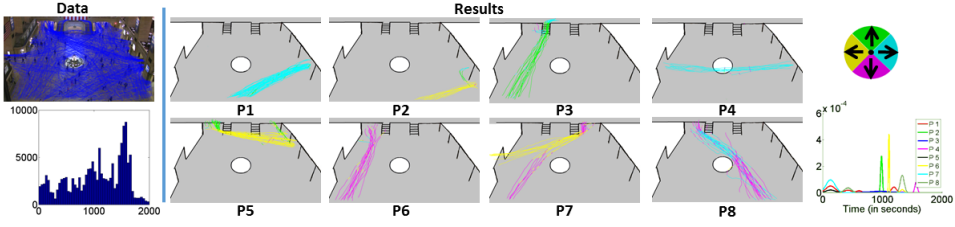

Fig. 6. Top Left: Environment of the New York Central Terminal. Bottom Left: Observation numbers over time. Right: Some activities shown by representative trajectories and their respective time activities. Colors indicate orientations described by the legend on the right.

3.3 Split and Merge

We test the effectiveness of split-merge (SM) by the per-word log likelihood by:

$$P_{per-word} = \frac{\sum_{n=1}^{N} P(w_n, t_n|\beta, v, \alpha, e)}{N} = \frac{\sum_{n=1}^{N}(\sum_{k=1}^{K} P(w_n|\beta_k, v_k)\sum_{l=1}^{L} P(t_n|\alpha_l, e_l, \bullet))}{N} \tag{9}$$

where N, K and L are the number of observations, learned spatial activities and time activities respectively. β and v are the spatial activities and their weights, α and e are the time activities and their weights, $\bullet$ represents all the other factors. In general, we found that SM increases the likelihood thus improves the model fitness on the data. Also, we found that MH sampling is more likely to pick a merge operation than a split. One reason is the time HDP tends to separate data samples that are temporally far away

from each other, thus causing similar patterns to appear at different times. A merge on those patterns has higher probability, thus is more likely to be chosen. Merging such spatially similar patterns makes each final activity unique. It is very important because not only does it make the activities more compact, it also makes sure that all the time activities for a particular spatial activity can be summarized under one pattern.

3.4 Anomaly Detection

For anomaly detection, Figure 7 shows the top outliers (i.e., unusual activities) in three datasets: (g) and (l) show trajectories crossing the lawn, which is rare in our sampled data; (i) shows a trajectory of leaving the lot then returning. In the latter case, the trajectory exited in the upper lane, whereas most activities involve entering in the upper lane and exiting in the bottom lane. The outliers in the Forum are also interesting. Figure 7 (a) shows a person entering through the front entrance, checking with the reception then going to the conference room; (b) shows a person entering through Z2 then leaving again;(d) is unexpected because visually it should be in Figure 4 (P7), but we found that the pattern peaks around 2:34pm and falls off quickly to a low probability area before 2:27:30pm whereas Figure 7 (c) occurs between 2:26:48pm-2:26:53pm. This example also demonstrates that our model identifies outliers not only on the spatial domain but also on the time domain. We also found similar cases in Figure 7 (k), (l) and (o) that are normal when only looking at the spatial activities but become anomalies when the timing is also considered.

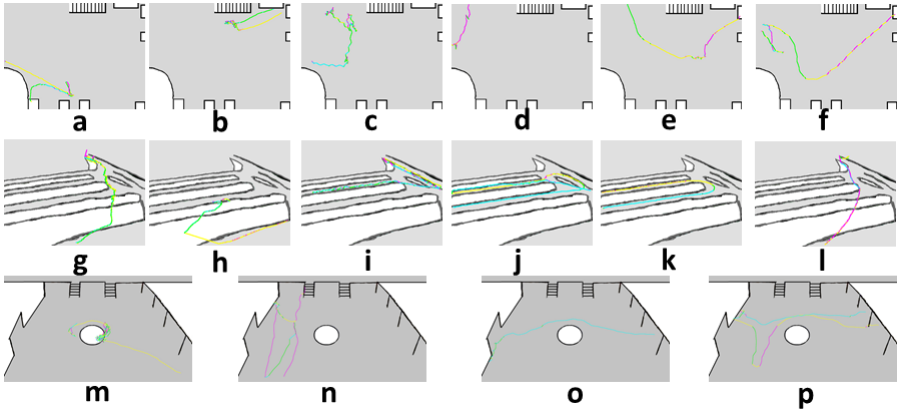

Fig. 7. Top: Outliers in Forum. Middle: Outliers in Carpark. Bottom: Outliers in TrainStation.

3.5 Comparison

Qualitative Comparison Our model is complementary to dynamic non-parametric models such as DHDP [16] and MOTIF [4]. Theoretically, our time modeling differs in

two aspects: continuity and multi-modality. Wang et.al [16] manually segment data into equal-length episodes and Emonet et.al [4] model local spatio-temporal patterns. Our method treats time as a globally continuous variable. Different time modeling affects the clustering differently. All three models are variants of HDP which assumes the *exchangeability* of data samples. This assumption is overly strict when time is involved because it requires any two samples from different time to be exchangeable. The manual segmentation [16] restricts the exchangeability within segment. Enforcing a peak-and-fall-off time prior [4] has similar effects.

For DHDP, we segment both datasets equally into 4 episodes. We run it for 5000 iterations on each episode. For MOTIF, we used the author's implementation [4]. For parameter settings, we use our best effort at picking the same parameters for three models, e.g. the Dirichlet, DP, Gamma and Beta distribution parameters on each level. Other model-specific parameters are empirically set to achieve the best performance. Also, since there is no golden rule regarding when to stop the sampling, we use the time DHDP models takes and run the other two for roughly the same period of time.

Since all three methods learn similar spatial activities, we mainly compare the temporal information. Figure 8 shows one common pattern found by all three methods. The temporal information of DHDP is simply the weight of this activity across different episodes. To get a dense distribution, smaller episodes are needed, but the ideal size is not clear. Therefore, we only plot the temporal information for MOTIF and STHDP. In Figure 8 Left, (c) is the starting time probability distribution of Figure 8 Left (b). The distribution is discrete and shows how likely it is that this pattern could start at a certain time instance, which reports quite different information from our pattern. Figure 8 Left (d) shows the time activities of Figure 4 (P3), which is continuous and shows its appearance, crescendo, multiple peaks, wane and disappearance. An interesting fact is that both methods capture this pattern within the first 8000 seconds while our model also captures a small bump beyond the first 8000 seconds. By looking at the data, we find that there are indeed a few trajectories belonging to this pattern beyond the first 8000 seconds. Figure 8 Right shows a common pattern in the Carpark dataset. Both MOTIF and STHDP capture the periodicity as seen in P3 and P4. They mainly differ at the start in that P3 captures two peaks whereas P4 captures one. Note that the two peaks that P3 captures depict how likely it is that the activity starts at those time instances, while STHDP captures the time span of that activity, which is essentially the same.

Quantitative Comparison Because all three methods model time differently, it is hard to do a fair quantitative comparison. As a general metric to evaluate a model's ability to predict, we use the per-word log likelihood (Equation 9). We hold out 10% of the data as testing data and evaluate the performance of the three models with respect to how well they can predict the testing data. We show the best per-word log likelihood of the methods after the burn-in period in Table 2.

This experiment favors DHDP and MOTIF. Because DHDP learns topics on different episodes, when computing the likelihood of a testing sample $\{w, t\}$, we only weighted-sum its likelihoods across the topics learned within the corresponding episode. So the likelihood is $p(w|t, \bullet)$ instead of $p(w, t|\bullet)$ where $\bullet$ represents all model parameters and the training data. For MOTIF, the learned results are topics as well as their

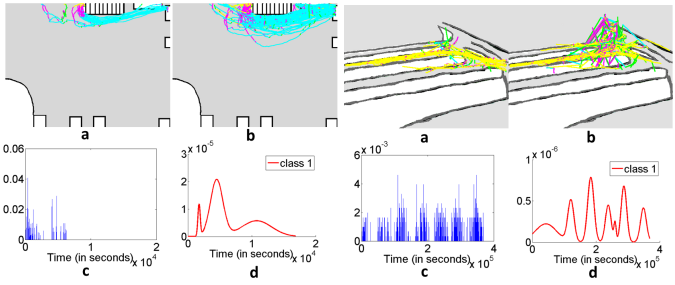

Fig. 8. Left: Forum dataset. (a) A pattern learned by DHDP. (b) A pattern learned by MOTIF. (c) The topic starting time distribution over time from MOTIF. (d) The time activities of Figure 4 (P3) from STHDP. Right: Carpark dataset. (a) A pattern learned by DHDP. (b) A pattern learned by MOTIF. (c) The topic starting time distribution over time from MOTIF. (d) The time activities of Figure 5 (1) from STHDP.

	STHDP	DHDP	MOTIF
Forum	-3.84	-9.8	-54.38
Carpark	-2.8	-7.75	-62.13
TrainStation	-3.5	-4.9	-62.2

$r_{correct}/r_{complete}$	STHDP	DHDP	MOTIF
Forum	0.92/0.88	0.95/0.63	0.87/0.78
Carpark	0.83/0.9	0.89/0.31	0.85/0.42
TrainStation	0.84/0.75	0.72/0.55	0.69/0.58

Table 2. Left: Best per-word log likelihoods. Right: $r_{correct}$ and $r_{complete}$ accuracies from 0-1, 1 is the best.

durations in time. We compute the likelihood of a testing sample by averaging the likelihoods across all topics whose durations contains t, i.e., $p(w, t|\beta_k, t \in rt_k, \bullet)$ where β_k is the topic with duration rt_k. For STHDP, the likelihood is computed across all word topics and all time topics, $p(w, t|\bullet)$, which is much more strict. We found that STHDP outperforms both DHDP and MOTIF, with MOTIF performing more poorly than the other two. We initially found the results surprising given the fact that MOTIF learns similar spatial activities to the other two. Further investigations showed that, since the testing data is randomly selected, this causes gaps in the training data in time. As a consequence of the discrete nature of the time representation in MOTIF, all MOTIF topics have low probabilities in those gaps, thus causing the low likelihoods. Removing the time and only considering spatial activities in this case may help but would not be fair to the other two methods.

Next, we compute the correctness/completeness of the three methods as in [16]. Correctness is the accuracy of trajectories of different activities not clustered together while completeness is the accuracy of trajectories of the same activity clustered together. To get the ground truth data for each dataset, the trajectories were first roughly clustered into activities. Then 2000 pairs of trajectories were randomly selected where each pair comes from the same activity and another 2000 pairs were randomly selected where each pair comes from two different activities. Finally these 4000 pairs of trajectories for each dataset were labeled and compared with the results of our method. We denote the correctness as $r_{correct}$ and the completeness as $r_{complete}$. Because estimating the number of clusters is hard, it was only needed to judge whether a pair

of trajectories was from the same activity or not. The correctness/completeness metric indicates that grouping all the trajectories in the same cluster results in 100% completeness and 0% correctness while putting every trajectory into a singleton cluster results in 100% correctness and 0% completeness. So only an accurate clustering can give good overall accuracies. Table 2 Right shows the accuracies. STHDP outperforms the other two on $r_{complete}$ across all datasets. Its $r_{correct}$ is higher in TrainStation and slightly worse in Forum and Carpark but the difference is small (within 6%).

Finally, we discuss how different temporal modeling could lead to different temporal anomaly detection. Most of the outliers in Figure 7 are *spatial* outliers and also detected by DHDP. However, some are not (Figure 7 (d) (k) (l) (o)). Figure 7 (d) is a good example. Spatially its probability is high because it is on one major activity shown in Figure 4 P7. However, if its temporal information is considered, our method gives a low probability because its timing is very different from the observations in Figure 4 P7. In contrast, DHDP gives a high probability because it first identifies the segment in which this trajectory is, then computes the probability based on the activities computed within the segment and the segments before. Since Figure 4 P7 and Figure 7 (d) are in the same segment, a high probability is given. The result is caused by the fact that DHDP models progressions between segments but the temporal information within a segment is not modeled. Meanwhile, MOTIF reports a higher probability on Figure 7 (d). However, it suffers from the situation explained by the low likelihoods in Table 2 Left. When a continuous chunk of data are missing, there is a void spanning a short period in the training data, which causes low probabilities on any observations in the time span. This kind of temporal information loss leads to false alarms for anomaly detections (all our testing data report low probabilities). In our method, if an activity is seen before and after the void, it will be inferred that there is a probability that the activity also exists in the middle by putting a Gaussian over it. Even if the activity only appears before or after the missing data, the Gaussian there prevents the probability from decreasing as quickly as it does in MOTIF.

4 Limitation and Conclusions

For performance comparison, we tried to run three models and stopped them once satisfactory activities were computed and compared the time. Our method is approximately the same as DHDP and can be slightly slower than MOTIF depending on the dataset. But we did not use larger datasets because although being able to report good likelihoods, sampling is in general slow for largest datasets and it applies to all three models. So we focused on experiments that show the differences between our model and the other two. Also, we find the data is abundant in terms of activities where a random sampling suffices to reveal all activities. Quicker methods for training such as variational inference [39] or parallel sampling can be employed in future.

In summary, we propose a new non-parametric hierarchical Bayesian model with a new hybrid sampling strategy for the posterior estimation for activity analysis. Its unique feature in time modeling provides better likelihoods, correctness/completeness and anomaly detection, which makes it a good alternative to existing models. We have shown its effectiveness on multiple datasets.

References

1. Zhou, S., Chen, D., Cai, W., Luo, L., Low, M.Y.H., Tian, F., Tay, V.S.H., Ong, D.W.S., Hamilton, B.D.: Crowd Modeling and Simulation Technologies. ACM Trans. Model. Comput. Simul. **20**(4) (November 2010) 20:1–20:35
2. Ali, S., Shah, M.: Floor Fields for Tracking in High Density Crowd Scenes. In Forsyth, D., Torr, P., Zisserman, A., eds.: Computer Vision ECCV 2008. Number 5303 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (October 2008) 1–14 DOI: 10.1007/978-3-540-88688-4_1.
3. Antonini, G., Martinez, S.V., Bierlaire, M., Thiran, J.P.: Behavioral Priors for Detection and Tracking of Pedestrians in Video Sequences. Int J Comput Vision **69**(2) (May 2006) 159–180
4. Emonet, R., Varadarajan, J., Odobez, J.: Extracting and locating temporal motifs in video scenes using a hierarchical non parametric Bayesian model. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2011) 3233–3240
5. Zhou, B., Tang, X., Wang, X.: Learning Collective Crowd Behaviors with Dynamic Pedestrian-Agents. Int J Comput Vis **111**(1) (June 2014) 50–68
6. Wang, X., Ma, K.T., Ng, G.W., Grimson, W.: Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. (June 2008) 1–8
7. Wang, X., Ma, X., Grimson, W.: Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. IEEE Trans. Patt. Anal. Machine Intel. **31**(3) (2009) 539–555
8. Stauffer, C., Grimson, W.E.L.: Learning Patterns of Activity Using Real-Time Tracking. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INLIGENCE **22** (2000) 747–757
9. Oliver, N., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8) (August 2000) 831–843
10. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001. Volume 2. (2001) II–123–II–130 vol.2
11. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Volume 2. (June 2004) II–819–II–826 Vol.2
12. Lin, D., Grimson, E., Fisher, J.: Learning visual flows: A Lie algebraic approach. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. (June 2009) 747–754
13. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 3488–3496
14. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity Forecasting. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: Computer Vision ECCV 2012. Number 7575 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (October 2012) 201–214 DOI: 10.1007/978-3-642-33765-9_15.
15. Xie, D., Todorovic, S., Zhu, S.C.: Inferring "Dark Matter" and "Dark Energy" from Videos. In: 2013 IEEE International Conference on Computer Vision (ICCV). (December 2013) 2224–2231
16. Wang, X., Ma, K.T., Ng, G.W., Grimson, W.E.L.: Trajectory Analysis and Semantic Region Modeling Using Nonparametric Hierarchical Bayesian Models. Int J Comput Vis **95**(3) (May 2011) 287–312

17. Varadarajan, J., Emonet, R., Odobez, J.M.: A Sequential Topic Model for Mining Recurrent Activities from Long Term Video Logs. Int J Comput Vis **103**(1) (December 2012) 100–126
18. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005. Volume 2. (June 2005) 524–531 vol. 2
19. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Describing Visual Scenes Using Transformed Objects and Parts. Int J Comput Vis **77**(1-3) (2007) 291–330
20. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. ICCV 2005 (2005)
21. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. Int. J. Comp. Vision **79**(3) (2008) 299–318
22. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience (2005)
23. Emonet, R., Varadarajan, J., Odobez, J.M.: Temporal Analysis of Motif Mixtures Using Dirichlet Processes. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(1) (January 2014) 140–156
24. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. J. Am. Stat. Assoc. **101**(476) (2006) 1566–1581
25. Dubey, A., Hefny, A., Williamson, S., Xing, E.P.: A non-parametric mixture model for topic modeling over time. arXiv:1208.4411 [stat] (August 2012) arXiv: 1208.4411.
26. Wang, C., Blei, D.M.: A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process. arXiv:1201.1657 [cs, stat] (January 2012) arXiv: 1201.1657.
27. Lin, D., Grimson, E., Fisher, J.W.: Construction of Dependent Dirichlet Processes based on Poisson Processes. In Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A., eds.: Advances in Neural Information Processing Systems 23. Curran Associates, Inc. (2010) 1396–1404
28. Blei, D.M., Frazier, P.I.: Distance Dependent Chinese Restaurant Processes. J. Mach. Learn. Res. **12** (November 2011) 2461–2488
29. Wang, X., McCallum, A.: Topics over Time: A non-Markov Continuous-time Model of Topical Trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06, New York, NY, USA, ACM (2006) 424–433
30. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica **4** (1994) 639–650
31. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. Journal of Machine Learning Research **14**(1) (2013) 1303–1347
32. Chang, J., Fisher, J.W.: Parallel Sampling of HDPs using Sub-Cluster Splits. (2014)
33. Hughes, M.C., Fox, E.B., Sudderth, E.B.: Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data. (2012)
34. Jain, S., Neal, R.: A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. Journal of Computational and Graphical Statistics **13** (2000) 158–182
35. Dahl, D.B.: Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models. (2005)
36. Majecka, B.: Statistical models of pedestrian behaviour in the Forum. MSc Dissertation, School of Informatics, University of Edinburgh, Edinburgh (2009)
37. Luber, M., Spinello, L., Silva, J., Arras, K.O.: Socially-aware robot navigation: A learning approach. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). (October 2012) 902–907
38. Almingol, J., Montesano, L., Lopes, M.: Learning Multiple Behaviors from Unlabeled Demonstrations in a Latent Controller Space. (2013) 136–144

39. Wang, H., Ondej, J., O'Sullivan, C.: Path Patterns: Analyzing and Comparing Real and Simulated Crowds. In: Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. I3D '16, New York, NY, USA, ACM (2016) 49–57