Expert Systems With Applications 78 (2017) 386–395

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Hierarchically linked infinite hidden Markov model based trajectory analysis and semantic region retrieval in a trajectory dataset

Yongjin Kwon, Kyuchang Kang, Junho Jin, Jinyoung Moon, Jongyoul Park*

SW•Content Research Laboratory, Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, Republic of Korea

ARTICLE INFO

Article history:
Received 9 November 2016
Revised 14 February 2017
Accepted 15 February 2017
Available online 16 February 2017

Keywords:
Trajectory analysis
Semantic regions
Nonparametric Bayesian models
Infinite hidden Markov models
Sticky extensions

ABSTRACT

With an increasing attempt of finding latent semantics in a video dataset, trajectories have become key components since they intrinsically include concise characteristics of object movements. An approach to analyze a trajectory dataset has concentrated on semantic region retrieval, which extracts some regions in which have their own patterns of object movements. Semantic region retrieval has become an important topic since the semantic regions are useful for various applications, such as activity analysis. The previous literatures, however, have just revealed semantically relevant points, rather than actual regions, and have less consideration of temporal dependency of observations in a trajectory. In this paper, we propose a novel model for trajectory analysis and semantic region retrieval. We first extend the meaning of semantic regions that can cover actual regions. We build a model for the extended semantic regions based on a hierarchically linked infinite hidden Markov model, which can capture the temporal dependency between adjacent observations, and retrieve the semantic regions from a trajectory dataset. In addition, we propose a sticky extension to diminish redundant semantic regions that occur in a non-sticky model. The experimental results demonstrate that our models well extract semantic regions from a real trajectory dataset.

1. Introduction

Discovering latent semantics in a video is an essential topic in computer vision. Although human beings can readily understand the contents and semantics in a video, computers just regard the video as a sequence of frames. A large number of studies that help computers recognize the hidden semantics in a variety of topics have been conducted. For example, person recognition (Kim, Kim, Lee, & Jeong, 2015a), object recognition (He, Zhang, Ren, & Sun, 2016; Szegedy et al., 2015), object detection (Redmon, Divvala, Girshick, & Farhadi, 2016; Ren, He, Girshick, & Sun, 2015), object tracking (Nam & Han, 2016; Yoon, Lee, Yang, & Yoon, 2016), activity and event recognition (Ramanathan et al., 2016; Simonyan & Zisserman, 2014), and even prediction (Pintea, van Gemert, & Smeulders, 2014; Vondrick, Pirsiavash, & Torralba, 2016) are the most popular topics in computer vision. Even in the same topic, the approaches may differ according to applications, such as surveillance (Kim, Jang, Kim, & Kim, 2015b; Lei & Xu, 2006), entertainment (Wang & Mori, 2009), first-person camera (Ohnishi, Kanehira, Kanezaki, & Harada, 2016), etc.

Due to the intrinsic nature of videos, the semantics in videos are usually associated with the patterns of temporal variations and object movements. To reveal the semantics, it is necessary to find the implicit characteristics of object movements. After objects in a video are tracked by several methods (Bae, Kang, Liu, & Chung, 2016; Elbahri et al., 2016; Park, Lee, & Yoon, 2016), a *trajectory* of each object is described as a sequence of positions. Since Shao and Li (2015) indicated that trajectories contained concise characteristics of the movements, trajectory analysis has become an important part of video analysis. Among several types of latent semantics in trajectories, we focus on *semantic regions* (Kwon, Jin, Moon, Kang, & Park, 2016; Wang, Ma, Ng, & Grimson, 2011). In a given scene, a number of objects appear and move in certain directions. We can extract some regions that have similar patterns of object movements, and each object moves through a series of the regions. We refer to the region as a semantic region. The semantic region can be exploited for some applications, such as activity analysis (Zhou, Wang, & Tang, 2011).

The previous approaches to semantic region retrieval, however, have some drawbacks. Their approaches have no consideration for actual "regions" when modeling semantic regions. In other words, their approaches neither defined a semantic region as a region, nor provided a model to cover regions. For example, Wang et al. (2011), Zhou et al. (2011), and Zou, Chen, Wei, Han, & Jiao (2013) simply

* Corresponding author.
E-mail addresses: scocso@etri.re.kr (Y. Kwon), k2kang@etri.re.kr (K. Kang), junho@etri.re.kr (J. Jin), jymoon@etri.re.kr (J. Moon), jongyoul@etri.re.kr (J. Park).

http://dx.doi.org/10.1016/j.eswa.2017.02.026
0957-4174/

defined the semantic region as a subset of common paths, which did not include the notion of regions. In addition, these approaches suggested models for learning semantic regions based on the topic models (Blei, Ng., & Jordan, 2003; Teh, Jordan, Beal, & Blei, 2006). Their models assumed that observations and topics followed a discrete distribution, rather than a continuous distribution. Thus the expressive power of the models was limited to sets of points, not regions. In addition, their approaches did not capture the temporal dependency of observations in a trajectory. Since their base topic models had the "bag-of-words" assumption, the authors could not capture the order of observations.

In this paper, we propose a novel model that finds semantic regions by analyzing a set of trajectories, while it treats semantic regions as actual regions. We first revisit the semantic region and extend its meaning to cover actual regions. To model semantic regions, we adopt the concept of hierarchically linked infinite hidden Markov model (Sohn et al., 2015) and build a different model, called a *sest-hiHMM*, suitable for learning trajectories and semantic regions. We believe that our model has more expressive power than the previous models because it intrinsically includes the region-based concept of latent semantics and the temporal dependency of observations. For example, our model would compute more accurately the likelihood of a new trajectory that consists of new observations at the training dataset than the previous models. In addition, we also propose a sticky extension, called a *sticky sest-hiHMM*, to decrease redundant semantic regions that appear in a non-sticky model. We evaluate the two models on a real trajectory dataset. We observe whether semantic regions are well extracted, and investigate some interesting points of our models, such as how semantic regions are changed as the number of iterations increases and how they are shaped comparing to the trajectories.

The rest of this paper is organized as follows. Section 2 presents the previous approaches in video analysis and the importance of trajectory analysis. Section 3 revisits the semantic region and similar concept, and extends its meaning for our purpose. Section 4 describes essential nonparametric Bayesian models and our models based on hierarchically linked infinite hidden Markov models. Section 5 shows the experimental results of semantic region retrieval by our models and discuss the implications of the results. Finally Section 6 concludes our works and provides some future research directions.

2. Related work

In video analysis, a number of challenges have been addressed. One of the most challenging topics is action recognition. Robertson and Reid (2006) presented a system for activity recognition from videos. They computed positions, velocities, and local feature vectors of the target, and then classified activities. Donahue et al. (2015) developed a deep convolutional neural network for activity recognition. To handle temporal semantics, they appended Long-Short Term Memory (LSTM) networks. Piergiovanni, Fan, and Ryoo (2017) introduced the temporal attention filters that found important intervals in videos to improve the performance of activity recognition. Some literatures moved beyond action recognition. Scene understanding not only recognizes individual activities in the scene, but also figures out the characteristics of or the relationships between these activities. Brun, Saggese, and Vento (2014) tried to understand a given scene by capturing activity patterns. The authors proposed a kernel-based analysis of object trajectories to distinguish between normal and abnormal behaviors. Wu, Fu, Jiang, and Sigal (2016) proposed a neural network model to reveal relationships between objects/scene and actions, as well as to recognize actions (video classes). In addition to activity recognition and scene understanding, there have been challenging problems in video analysis.

As we mentioned, a large number of approaches in video analysis requires trajectory analysis. Prior to trajectory analysis in videos, the preprocessing step of single or multiple object tracking is required. A number of investigations of object tracking have been conducted. Ali and Dailey (2012) suggested multiple pedestrian tracking algorithm that works fully automatically, even for a very complicated scene, such as lots of crowds and occlusions. Their algorithm performed both human detection and tracking, and connected the detections to tracks by a confirmation-by-classification method. Cancela, Ortega, Fernández, and Penedo (2013) adopted a hierarchical approach that considered both low-level and high-level trackers. They leveraged a pool of histogram to preserve object identification during tracking. In Bae and Yoon (2014), the authors tackled the difficulties of multi-object tracking, such as occlusions, unreliable detections, and indistinctive appearances of different objects, by exploiting tracklet confidence and online appearance learning. Choi, Moon, and Yoo (2015) developed a robust multi-person tracker for real-time video surveillance. They concentrated on achieving real-time processing speed while ensuring a high accuracy and precision of tracking by reducing the computation and failure chance. Mithun, Howlader, and Rahman (2016) paid attention on a traffic management system and considered the vehicle tracking in the road network. The authors presented a tracking-by-detection approach by estimating vehicle positions accurately, while computing efficiently. Yildirim et al. (2016) revised a particle filter to take the direction of target vehicles into account. By weighting the target with similar direction, the performance of vehicle tracking under occlusion and camera vibration was enhanced.

Trajectory analysis itself involves many research topics according to applications. Trajectory clustering is a popular topic in trajectory analysis. It measures the similarity between each pair of trajectories, and then assigns similar trajectories to the same group. Piciarelli and Foresti (2006) introduced an online trajectory clustering that exploited a tree structure to contain the cluster information. For each trajectory, the clusters in the tree were updated, or were split to create a new cluster. Jung, Hennemann, and Musse (2008) explicitly dealt with outliers by postprocessing stages. They removed significantly broader clusters than the remaining ones because the clusters had isolated observations, and merged clusters that were significantly overlapped to decrease classification errors. Acevedo-Rodríguez, Maldonado-Bascón, López-Sastre, Gil-Jiménez, and Fernández-Caballero (2011) developed a novel clustering algorithm inspired by the notion of growing neural gas. The method could represent subtrajectories, and thus it could find prototype trajectories by constructing sequences of subtrajectories. Jeong, Yoo, Yi, and Choi (2014) suggested a latent Dirichlet allocation (LDA) based probabilistic model to perform trajectory clustering. The authors came up with a two-stage greedy inference method, rather than exact inference, Xu, Zhou, Lin, and Zha (2015) extended the mean shift clustering and manifold-based modeling to improve the quality of trajectory clustering. Their approach was robust to the noise, missing data, and changes of parameters. Kumar, Bezdek, Rajasegarar, Leckie, and Palaniswami (2015) introduced a two-stage clustering algorithms. The first stage performed a trajectory clustering, neglecting the direction of each trajectory. At the second stage, a direction-aware trajectory clustering was performed for each cluster.

Another topic in trajectory analysis is trajectory prediction. In this topic, the near future of object actions or movements are inferred before the observations are available in a given scene. Kitani, Ziebart, Bagnell, and Hebert (2012) considered the physical natures of objects and prior knowledge of scene configurations. Their work, however, paid attention on a single object movement, thereby being able to forecast the target trajectory. Walker, Gupta, and Herbert (2014) introduced a method of predicting the

changes of patches, such as transitions and appearances, rather than the changes of locations. The authors took into account not only patch movements and appearances, but also contextual information to see the relationship between the elements and surroundings. Akbarzadeh, Gagné, and Parizeau (2015) proposed a way of measuring similarities between two trajectories using kernel density estimation. They computed the similarities between the target trajectory and past trajectories, and then they used them as weight parameters for estimating the future positions of the target. Yi, Li, and Wang (2015) paid attention on stationary crowd groups. They believed that stationary crowd groups affected the movements of other people. Hence they built a model that captured the interactions between stationary crowd groups and pedestrians to forecast the trajectories. On the other hand, Alahi et al., (2016) modeled the interactions between pedestrians themselves. The authors exploited LSTMs to learn the interactions from the training dataset without specific settings. Yoo et al., (2016) also considered the interactions between co-occurring objects. They presented a hierarchical Bayesian model that learned both the movement patterns and relationships between the patterns. Ballan, Castaldo, Alahi, Palmieri, and Savarese (2016) introduced a navigation map that includes prior knowledge about interactions between the observed targets and the scene. They combined the navigation map with trajectory prediction to produce more probable paths.

There are other applications to exploit trajectory analysis. Li, Hu, Zhang, Zhang, and Luo (2008) exploited trajectories for video retrieval. The authors introduced a video retrieval framework that used nonparametric Bayesian models for trajectory index construction and video retrieval. Iwashita, Ryoo, Fuchs, and Padgett (2013) introduced human recognition using the results of object tracking from low resolution aerial videos. Their method interpreted a trajectory of each candidate to classify it as a human or just noise. Cui, Liu, and Xing (2014) provided a framework to train the hidden semantics in a scene from trajectories, and to detect whether an unseen trajectory is normal or abnormal activity. Lim, Tang, and Chan (2014) proposed a framework for multiple event detection under video surveillance. In their work, object tracking was adopted to estimate locations of the object or to identify the same object. In Arroyo, Yebes, Bergasa, Daza, and Almazán (2015), the situation in shopping malls was considered. The authors introduced an expert system for surveillance in shopping malls to track people trajectories and to find potentially suspicious behaviors.

Among the various topics in trajectory analysis, we have concentrated on semantic region retrievals. There have been previous studies for semantic region retrievals as well. Wang et al. (2011) introduced dual hierarchical Dirichlet processes (Dual-HDPs) to build the models of semantic regions. Zhou et al. (2011) integrated the LDA model (Blei et al., 2003) with Markov random fields (MRFs) as priors to capture both spatial and temporal relationships. On the other hand, Zou et al. (2013) integrated the correlated topic model (CTM) (Blei & Lafferty, 2007) with the forest of spanning trees, which is a kind of MRFs. The authors also concerned the topic clustering to improve the discriminative ability of semantic regions, which made it difficult to exploit the learnt models because it was not formulated as Bayesian approaches. These studies, however, still have limitations. Firstly these approaches converted trajectory data into another description to build the inputs of their models, which may cause the loss of semantics in the trajectory data. In addition, the observations of each trajectory were quantized by the conversions, and thus were treated as discrete values. It made their models concern sets of points, rather than actual "regions" while retrieving semantic regions. Furthermore, their approaches have less consideration of the temporal dependency of observations in each trajectory. Although the approaches in Zhou et al. (2011) and Zou et al. (2013) considered the sources and sinks, they were not sufficient to reflect the temporal dependency. In contrast, our work takes trajectory data themselves as the inputs without quantization. It is possible to assume that each semantic region has a continuous distribution. Therefore, our work can treat semantic regions as actual regions. Our model also concerns the temporal dependency of observations by exploiting the part of hidden Markov models (HMMs). Table 1 summarized a qualitative comparison between our work and previous works.

In video analysis, one of the well-known approach is to exploit HMMs. HMMs are useful for modeling invisible motions and latent structures in a given scene (Sun, Zhao, & Gao, 2015). Many literatures exploited HMMs on diverse applications. Chung and Liu (2008) introduced a hierarchical context HMM to recognize the elderly behaviors from video streams in a nursing home. Pruteanu-Malinici and Carin (2008) used an infinite HMM (iHMM) for modeling intrinsic characteristics of normal video. The authors detected infrequent events by computing the likelihood from the trained iHMM. In Kratz and Nishino (2009), a model for local motion patterns in very crowded scenes was proposed. The authors leveraged HMMs for describing both spatial and temporal relationships between motion patterns. Morris and Trivedi (2011) captured spatio-temporal dynamics of activities by HMMs. Hu et al. (2013) combined an improved hierarchical Dirichlet process with an HMM to discover motion patterns within trajectories and to cluster trajectories.

3. Semantic regions from videos

The concept of semantic regions originated with Wang et al. (2011). They regarded a semantic region as a subset of paths, but the concept was based on the topic models, such as LDA (Blei et al., 2003). They described semantic regions as follows. In a trajectory, each observation is drawn from a single semantic region. A trajectory itself is represented as a distribution over semantic regions. Note that the semantic regions are shared across trajectories. Hence a semantic region in Wang et al. (2011) corresponds to a topic in the topic models. As a similar concept, Emonet, Varadarajan, and Odobez (2011) proposed a motif. A motif was defined by the recurrent temporal pattern from a video. Similar to Wang et al. (2011), they associated a motif with a topic in the topic model as well. A different thing is that Emonet et al. (2011) used optical flow vectors in a static video, rather than trajectories, and thus a motif may not capture similar movements of two objects that have different velocities.

The approach in Wang et al. (2011), however, had some drawbacks. Their method required the conversion from each observation to a word. It was not ensured that the corresponding word contained the semantics that the observation originally had. In addition, their definition of semantic regions did not include the whole meaning of "semantic regions." Since trajectories were converted to documents of words, their model would be learnt from discrete data. Thus a point that was not observed could not be inferred. In other words, their models did not reveal the regions, but just the sets of semantically relevant points.

In this paper, we follow the definition of semantic regions in Kwon et al. (2016), which presented an extended definition that covered actual regions. It is difficult to define the term explicitly, but it can be described as follows. Suppose that an object appears and moves in a scene. It is believed that the object would had some purpose, and thus their movements would not be random. To complete the purpose, the object should move across some regions that are considered to be close to the purpose. Since a number of objects can appear in a video, one can aggregate all the regions, and then can map each region to some semantics. Therefore, the regions can be regarded as semantic regions. Fig. 1 shows an example of trajectories and semantic regions.

Table 1
Comparison of our work to of previous works.

	Our work	Wang et al. (2011)	Zhou et al. (2011)	Zou et al. (2013)
No transformation of trajectory data to another description is required.	√	×	×	×
Temporal dependency of observations in each trajectory is considered in the models.	√	×	Δ	Δ
Both spatial and temporal dependency of trajectories is considered in the models.	×	×	√	√
The number of semantic regions is automatically determined.	√	√	×	×
Semantic regions are modeled by fully Bayesian approaches.	√	√	√	×
Semantic regions cover actual regions, rather than sets of relevant points.	√	×	×	×

√ yes, × no, Δ partially.

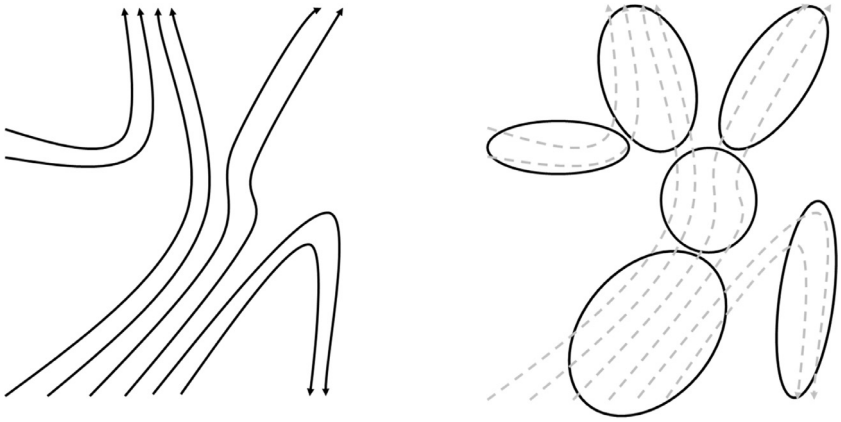


Fig. 1. An example of trajectories and semantic regions.

4. hiHMM-based trajectory and semantic region modeling

To compute semantic regions from a trajectory dataset, we first construct a model for trajectories and semantic regions. In this paper, a hierarchically infinite hidden Markov model (hiHMM) (Sohn et al., 2015) is considered as a base model to concern the temporal dependency between observations in each trajectory. Since the hiHMM itself is not suitable for semantic region retrieval, we modify it to model both trajectories and semantic regions. In the following subsections, we briefly describe essential models: HMMs, iHMMs, hiHMMs, and sticky HDP-HMMs, and introduce our models based on hiHMMs to model trajectories and semantic regions.

4.1. HMM, iHMM, and sticky HDP-HMM

HMMs are one of the most well-known model to assign a label to each observation in a sequence. In HMMs, given a sequence $y = (y_1, y_2, \cdots, y_T)$ and the number of labels or hidden states K, each observation y_t is associated with the hidden state $z_t \in \{1, 2, \cdots K\}$. The transition between hidden states in the HMM is denoted by the transition matrix $\pi \in \mathbb{R}^{K\times K}$. Each observation y_t has distribution $F(\cdot|\phi_{z_t})$, where ϕ_k is the emission parameter corresponding to the state k. As a Bayesian model (Beal, 2003), π and $\{\phi_k\}_{k=1}^K$ have their own priors as follows.

$$\beta|\gamma = \sim= Dir(\gamma/K, \cdots, \gamma/K)$$
$$\pi_k|\beta = \sim= Dir(\alpha_0\beta)$$
$$\phi_k|H = \sim= H$$

where π_k is the kth row of π, β is the shared prior parameter, and α_0 and γ are concentration parameters.

As a nonparametric Bayesian extension of HMMs, iHMMs (Beal, Ghahramani, & Rasmussen, 2001) were proposed. iHMMs have no limit to the number of hidden states. In other words, the number of hidden states K is regarded as the infinity. Then the transition and emission variables have an infinite number of parameters. In Beal et al. (2001), Dirichlet processes (DPs) are leveraged to integrate out the transitions and to express the model with a finite number of variables.

Beal et al. (2001) introduced the two-level state transition process. The state transition process shows how the sequence of hidden states is generated over time. Given the sequence of hidden states $z = (z_1, z_2, \cdots, z_T)$, the process determines z_t at each time step t, based on z_{t-1} that is already determined at the previous step. Let $n_{ij} = |\{z_{t'-1} = i \Lambda z_{t'} = j | 1 < t' < t\}|$. If $z_{t-1} = i$, then z_t is set to j with probability proportional to n_{ij}. Note that there is a probability proportional to α_0 of waiting for the decision of the *oracle*. The probabilities of transitions before inquiring of the oracle are determined as follows.

$$p(z_t = j|z_{t-1} = i, \alpha_0) = \begin{cases} \frac{n_{ij}}{\sum_{j=1}^K n_{ij} + \alpha_0} & j \text{ is chosen from } \{1, 2, \cdots, K\} \\ \frac{\alpha_0}{\sum_{j=1}^K n_{ij} + \alpha_0} & j \text{ is chosen by the oracle} \end{cases}$$

The decision of the oracle is made as follows. Let c_j be the number of hidden state j chosen by the oracle. The oracle chooses the state j with probability proportional to c_j. Note that there is a probability proportional to γ of assign to z_t a new state not in {1, 2, $\cdots$, K}. The probabilities of transitions determined by the oracle are determined as follows.

$$p(z_t = j|z_{t-1} = i, \gamma) = \begin{cases} \frac{c_j}{\sum_{j=1}^K c_j + \gamma} & j \text{ is chosen from } \{1, 2, \cdots, K\} \\ \frac{\gamma}{\sum_{j=1}^K c_j + \gamma} & j \text{ is a new state} \end{cases}$$

iHMMs were recalled in Teh et al. (2006) to associate the hierarchical Dirichlet process (HDP) with iHMM. Teh et al. (2006) defined a nonparametric HMM with the HDP, called the HDP–HMM, and showed that HDP–HMMs are actually equivalent to iHMMs. In the following subsection, we will explain the HDP and nonparametric Bayesian HMM with the HDP-based representation.

A drawback of iHMMs is that excessive transitions between redundant states are likely to be created. Fox, Sudderth, Jordan, and Willsky (2011) suggested sticky HDP–HMMs to give more weights on self-transitions while preserving nonparametric settings. For-

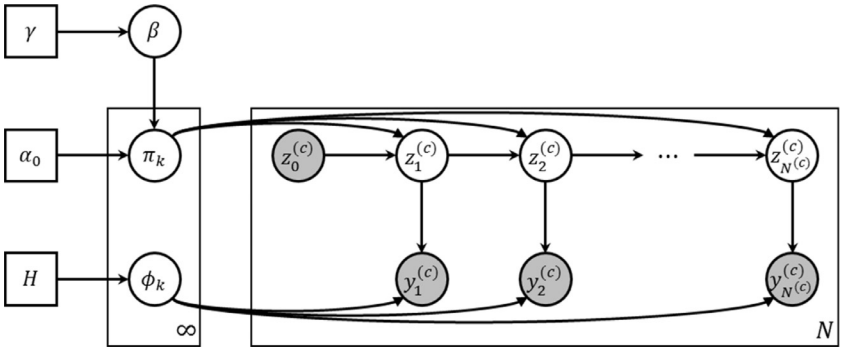


Fig. 2. A graphical representation of the sest-hiHMM.

mally, they introduced a sticky parameter κ to control the probability of self-transition from state j to itself to be proportional to $\alpha_0\beta_j + \kappa$, rather than $\alpha_0\beta_j$. The sticky model prevented excessive transitions because of less weights, and thereby reducing redundant states to a single state.

4.2. sest-hiHMMs and sticky sest-hiHMMs for modeling trajectories and semantic regions

We concern the hiHMM as a base model to concern the temporal dependency of observations in each trajectory. The concept of hiHMMs was originated from Sohn et al. (2015). While iHMMs inferred a single sequence of observations, hiHMMs jointly inferred a set of sequences by sharing priors or latent variables with multiple iHMMs. In Sohn et al. (2015), two hiHMMs were proposed; one assumed species-specific emissions, which loosely coupled iHMMs by sharing priors, and the other assumed shared emissions, which rather tightly coupled with a common emission matrix. The authors exploited hiHMMs to interpret chromatin state analysis.

hiHMMs, however, are not suitable for our problem. Semantic regions are considered as shared semantics of observations in trajectories, and semantic regions are associated with transitions. Thus we propose another version of the hiHMM that assumes *shared transitions*, as well as *shared emissions* (denoted by *sest-hiHMM*). Suppose we have N trajectories. Each trajectory $y^{(c)}$ has $N^{(c)}$ observations of positions $y_t^{(c)}$ $(1 \le t \le N^{(c)})$. Each observation is associated with the hidden state $z_t^{(c)}$, which in this paper is regarded as the indicator of semantic regions. The transition from $z_{t-1}^{(c)}$ to $z_t^{(c)}$ is associated with the shared transition matrix π, each row of which has the DP with a common base measure β. Note that β is the stick breaking construction for DPs (Teh et al. 2006). Each observation is thought of as a sample from the semantic region model $F(\cdot\,|\phi_k)$, where the shared emission parameter ϕ_k is generated from the base distribution H. A graphical representation of the sest-hiHMM is shown in Fig. 2. In summary, a sest-hiHMM can be described as follows.

$$
\begin{aligned}
\beta|\gamma &= \sim = GEM(\gamma)\\
\pi_k|\alpha_0,\beta &= \sim = DP(\alpha_0,\beta)\\
\phi_k|H &= \sim = H\\
z_t^{(c)}|z_{t-1}^{(c)},\{\pi_k\}_{k=1}^{\infty} &= \sim = \pi_{z_{t-1}^{(c)}}\\
y_t^{(c)}|z_t^{(c)},\{\phi_k\}_{k=1}^{\infty} &= \sim = F(y_t^{(c)}|\phi_{z_t^{(c)}})
\end{aligned}
$$

For each semantic region $z_t^{(c)}(=k)$, we assume to be a normal distribution $N(\mu_k,\ \Sigma_k)$, that is, $\phi_k = (\mu_k, \Sigma_k)$. The base distribution H is a normal-inverse-Wishart distribution $NIW(\mu_0,\ \lambda_0,\ \nu_0,\ \Psi_0)$ such that $\mu_0 \in \mathbb{R}^d$, $\lambda_0 > 0$, $\nu_0 > d-1$, and $\Psi_0 \in \mathbb{R}^{d\times d}$, where d is the dimension of observations, and thereby $d = 2$. Here we note that sest-hiHMMs assume that the semantics have continuous distributions, and thus it can cover actual regions that are not considered in previous studies.

Since sest-hiHMMs come from iHMMs and hiHMMs, they have the same drawback: redundant semantic regions and excessive transitions. We adopted the idea of Fox et al. (2011) to suggest a sticky version of the sest-hiHMM, called *sticky sest-hiHMM*. Similar to the sticky HDP–HMM, the transition matrix π is associated with both α_0 and κ. A different thing is that while the sticky HDP–HMM controls transitions in a single sequence, the sticky sest-hiHMM deals with transitions in multiple sequences as the sest-hiHMM assumes shared transitions. A graphical representation for the sticky sest-hiHMM is shown in Fig. 3. In summary, the sticky sest-hiHMM is described as follows.

$$
\begin{aligned}
\beta|\gamma &= \sim = GEM(\gamma)\\
\pi_k|\alpha_0,\kappa,\beta &= \sim = DP\left(\alpha_0+\kappa,\ \frac{\alpha_0\beta+\kappa\delta_k}{\alpha_0+\kappa}\right)\\
\phi_k|H &= \sim = H\\
z_t^{(c)}|z_{t-1}^{(c)},\{\pi_k\}_{k=1}^{\infty} &= \sim = \pi_{z_{t-1}^{(c)}}\\
y_t^{(c)}|z_t^{(c)},\{\phi_k\}_{k=1}^{\infty} &= \sim = F(y_t^{(c)}|\phi_{z_t^{(c)}})
\end{aligned}
$$

4.3. Posterior inference

We adopted beam sampling (Van Gael, Saatci, Teh, & Ghahramani, 2008) that is an efficient inference technique to deal with an infinite number of latent states by combining slice sampling and dynamic programming. It introduces auxiliary variables to limit the number of possible state sequences. Then, it computes the probabilities of all possible sequences, and samples the whole sequence by dynamic programing. For our models, we introduced auxiliary variables $u^{(c)} = (u_1^{(c)}, u_2^{(c)}, \cdots, u_{N^{(c)}}^{(c)})$ for each trajectory $y^{(c)}$. Based on the previous state sequences, the parameters of semantic regions ϕ_k, shared prior parameter β, and transition probabilities π are computed, and then the values of auxiliary variables are computed. Note that to sample β, we exploited the Chinese restaurant franchise (Teh et al., 2006) for sest-hiHMMs and the Chinese restaurant franchise for loyal customers (Fox et al., 2011) for sticky sest-hiHMMs, instead of Stirling numbers of the first kind, due to the heavy computation of Stirling numbers. The number of possible state sequences is determined by comparing the values of $u^{(c)}$ and transition probabilities. After the possible state sequences, transition probabilities, and other variables are determined, each state sequence is sampled independent of the other sequences.

5. Experiments

We explore the results of sest-hiHMMs and sticky sest-hiHMMs with the MIT trajectory dataset (Wang et al., 2011). This dataset

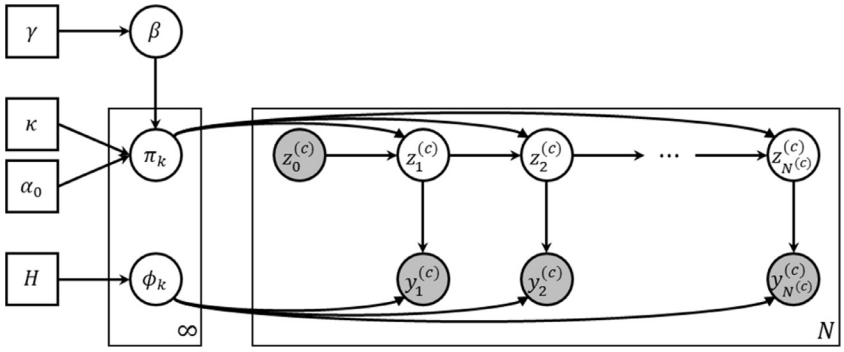
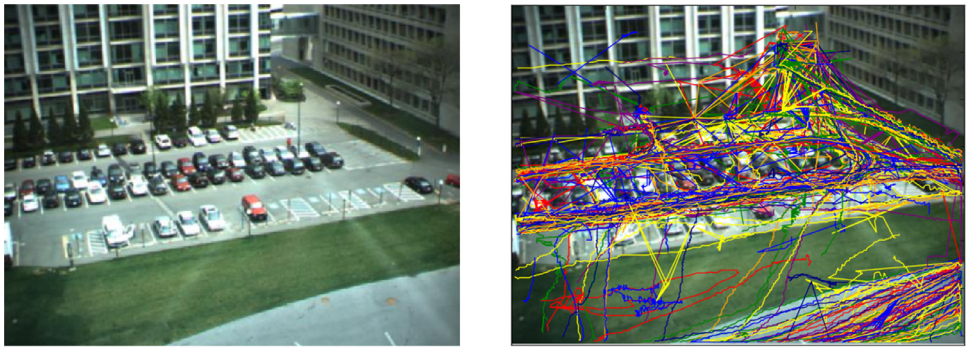


Fig. 3. A graphical representation of the sticky sest-hiHMM.



(a) Background image of the parking lot (b) 1% sample trajectories in the dataset

Fig. 4. Trajectories in the MIT trajectory dataset (Wang et al., 2011).

was composed of 40,453 trajectories of objects that appeared at a parking lot scene captured from a single static camera for five days. Fig. 4 shows the background and sample trajectories obtained from the parking lot scene. We first present the actual semantic regions obtained from sest-hiHMMs and sticky sest-hiHMMs. Although our models treat the semantic regions as continuous distributions, we represent each semantic region as an isoline, based on the values of the corresponding distribution at the pixels to see the location and range of the semantic region looks like and to check whether the semantic region is effectively extracted. Note that we set the initial number of semantic regions as 15, and we filter out small semantic regions in which the number of observations is smaller than the half of the total number of observations divided by the total number of semantic regions.

To avoid specifying the concentration parameters of sest-hiHMMs and sticky sest-hiHMMs, we provide hyperpriors on the concentration parameters as Escobar and West (1995), Teh et al. (2006), and Fox et al. (2011). Under the sest-hiHMM, the concentration parameters α_0 and γ have vague hyperprior *Gamma*(1, 1). Under the sticky sest-hiHMM, an auxiliary parameter ρ is introduced as follows:

$$\rho = \frac{\kappa}{\alpha_0 + \kappa}$$

Note that $\alpha_0 = (1-\rho)(\alpha_0+\kappa)$ and $\kappa = \rho(\alpha_0+\kappa)$. Thus instead of sampling α_0 and κ directly, we can sample ρ and $(\alpha_0+\kappa)$, and then compute α_0 and κ. The concentration parameters α_0 and γ have vague hyperprior *Gamma*(1, 1), and the parameter ρ have hyperprior *beta*(10, 1). For both models, the parameters of the base distribution H are set as follows. μ_0 is the vector pointing to the center of the scene, $\lambda_0 = 1$, $\nu_0 = 5$, and $\Psi_0 = \begin{bmatrix} w/2 & 0 \\ 0 & h/2 \end{bmatrix}$, where w and h are the width and height of the scene, respectively.

5.1. Semantic regions on trajectories

Fig. 5 shows the semantic regions obtained by sest-hiHMMs, with respect to the number of iterations. Based on the experimental results, we can notice some interesting points of the semantic regions. One about the shape of semantic regions. Unlike the previous studies, our models generate region-like results that include latent semantics. For sest-hiHMMs, the isoline of each semantic region is described as an ellipse since the semantic region has a normal distribution. Note that a point in the scene can be assigned to a semantic region, for example, by comparing the likelihoods for the semantic regions, even though the point is not covered by any semantic regions. For previous models, the point would be assigned to no semantic region.

Another point is about the shape of semantic regions. Each semantic region is likely to be stretched according to object movements. For example, at the parking area in the scene, objects usually appear at the middle-left side, move towards the middle-right side, turn around, move towards the middle-left side, and disappear. In other words, objects at the parking area usually moves horizontally. Similarly, the semantic regions at the parking area are horizontally stretched and vertically thin. It implies that the semantic regions retrieved by sest-hiHMMs can successfully include latent semantics from the trajectories, which are associated with the temporal dependency of observations in each trajectory.

The third one is about the number and granularity of semantic regions. As the number of iterations increases, the number of semantic regions gradually increases. On the other hand, the size of each semantic region is likely to decrease. This is because at first, sest-hiHMMs coarsely retrieve semantic regions, but as the number of iterations increases, sest-hiHMMs break the semantic regions into small pieces, to understand the scene more specifically. For example, the road at the bottom-right side is covered by

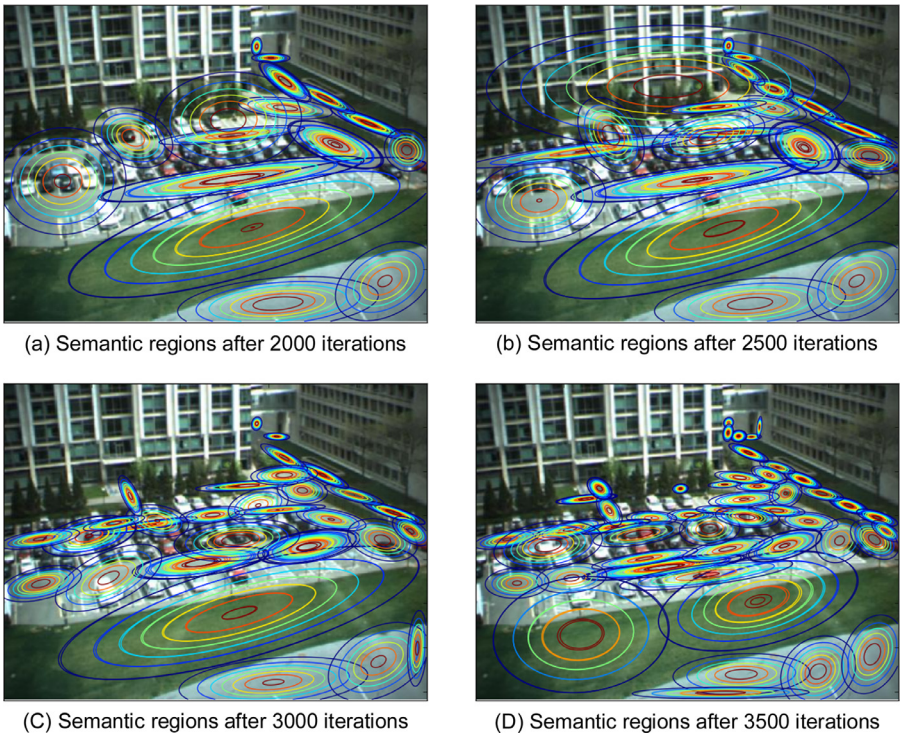


Fig. 5. Semantic regions retrieved by sest-hiHMMs with respect to the number of iterations.

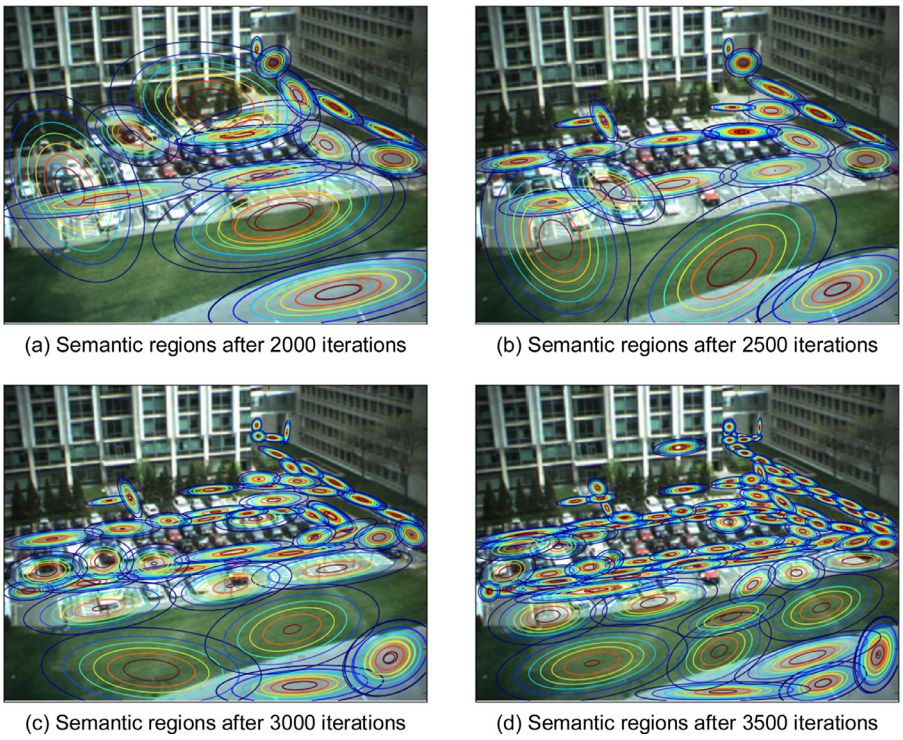


Fig. 6. Semantic regions retrieved by sticky sest-hiHMMs with respect to the number of iterations.

two semantic regions after 2000 iterations, but is covered by four semantic regions after 3500 iterations.

Sticky sest-hiHMMs show similar results to sest-hiHMMs as well. Fig. 6 shows the semantic regions obtained by sticky sest-hiHMMs. The semantic regions obtained by sticky sest-hiHMMs are described as actual regions and have similar shapes to the corresponding object movements. Sticky sest-hiHMMs have a tendency to generate a larger number of smaller semantic regions as the number of iterations increases.

The main difference between two models is about redundant states and excessive transitions. Fig. 7 shows the redundant semantic regions retieved by the two models. After 3000 iterations, 38.5% of valid semantic regions retrieved by sest-hiHMMs are determined as redundant (at 6 locations), while 18.9% of valid semantic regions retrieved by sticky sest-hiHMMs are determined as redun-

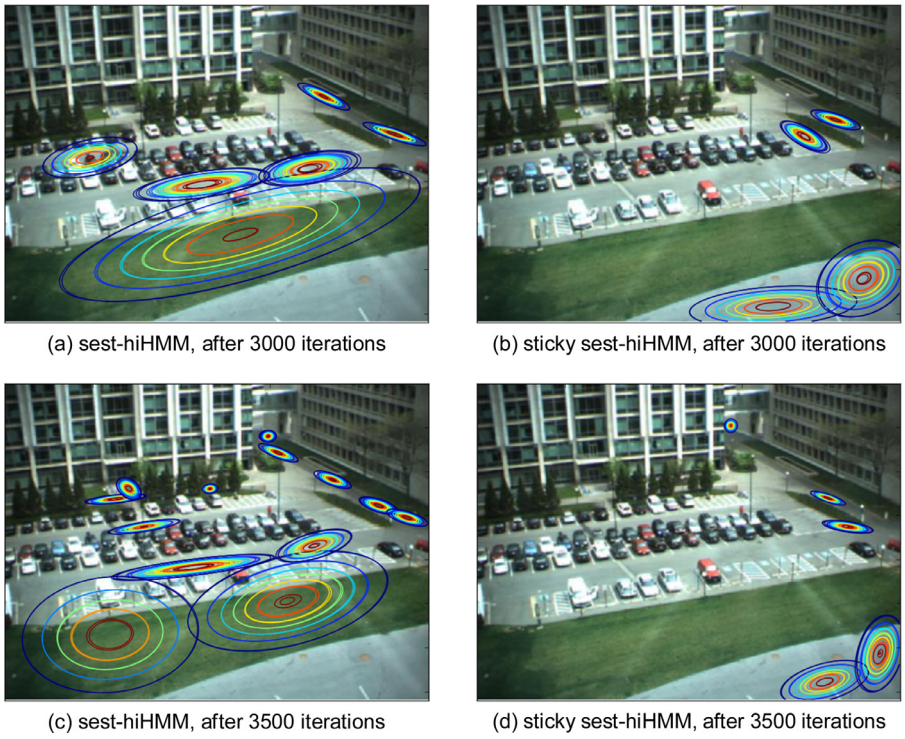


Fig. 7. Redundant semantic regions retrieved by both models.

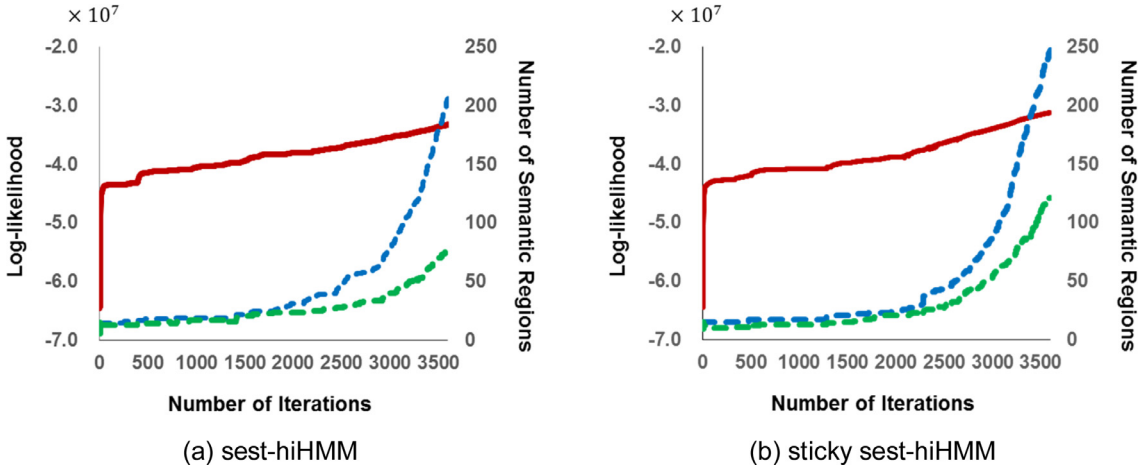


Fig. 8. Log-likelihood (red), number of non-filtered semantic regions (blue), and number of filtered semantic regions (green), with respect to the number of iterations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dant (at 4 locations). After 3500 iterations, 47.8% of valid semantic regions retrieved by sest-hiHMMs are determined as redundant (at 13 locations), while 11.0% of valid semantic regions retrieved by sticky sest-hiHMMs are determined as redundant (at 5 locations). Therefore, it is concluded that sest-hiHMMs are likely to extract redundant states, which leads to excessive transitions because sest-hiHMMs are likely to assign successive observations in a trajectory to different but redundant semantic regions, as we mentioned in Section 4.2. On the other hand, sticky sest-hiHMMs are less likely to retrieve redundant semantic regions because of the new parameter κ.

5.2. Quantitative evaluation of sest-hiHMMs and sticky sest-hiHMMs

We present how our models work as the number of iterations increases, using the number of semantic regions and log-likelihood. We first retrieve semantic regions and hidden state sequences at each iteration. Then we obtain the number of semantic regions and filtered semantic regions (which is computed as described in Section 5). For the log-likelihood, we sample the parameters for semantic regions, as well as hyperparameters. In the experiments, we perform 3600 iterations. The experimental results are shown in Fig. 8.

Both models show a similar tendency as the number of iterations increases. At initial, the log-likelihood dramatically decreases while the number of semantic regions does not show significant changes. The initial locations of semantic regions are given randomly, so that they are likely to be positioned at wrong locations. Thus at a few initial iterations, the semantic regions should move towards the right locations that contain latent semantics. After some iterations, the number of semantic regions begins to increase steeply, while the log-likelihood gradually increases. It implies that

after the semantic regions are positioned at the right locations, the semantic regions are split into two or more smaller semantic regions, to find more specific semantics. The problem is that a lot of tiny semantic regions, to which very few observations are assigned, are generated, and they are likely to be outliers. To deal with the problem, we use the aforementioned filtering method, and we can see that the number of filtered semantic regions are much smaller than that of non-filtered semantic regions.

5.3. Discussions

In the experiments, we investigated how the semantic regions were retrieved by our models from the real trajectory dataset. The experiments also verified that both sest-hiHMMs and sticky sest-hiHMMs included the temporal dependency of observations in each trajectory and captured actual regions as semantic regions. However, our models may encounter some problems. As shown in Fig. 8, our models generated a lot of tiny semantic regions as the number of iterations increased. Although each tiny semantic region may have a detailed meaning in the scene, we sometimes prefer to regulate the area of semantic regions. In the experiments, we removed tiny ones heuristically, but it is necessary to adjust the area of semantic regions by the models. Furthermore, the number of redundant semantic regions were considerably decreased by sticky sest-hiHMMs, but some of them were still existent. A stricter modification to the models, than sticky extensions, would be required to remove redundant semantic regions. We will explore how to overcome these problems as future works.

6. Conclusions

In this paper, we proposed a novel model for semantic region retrieval from a trajectory dataset, based on hiHMMs. We first revisited the semantic region to extend its meaning to cover actual regions, rather than sets of points. We built a novel model, called a sest-hiHMM, that was based on the notion of hiHMMs, but was modified for trajectories and semantic regions. We also built a sticky sest-hiHMM to reduce redundant semantic regions and excessive transitions between semantic regions. The experimental results showed that both models retrieved reasonable semantic regions from a real trajectory dataset. In particular, as the number of iterations increases, semantic regions retrieved by our models become smaller and include more fine-grained semantics to understand latent semantic given in a trajectory dataset. Moreover, semantic regions are likely to be shaped according to the corresponding object movements. The experimental results also verified that sticky hiHMMs suffered less from redundant semantic regions and excessive transitions than non-sticky hiHMMs.

Different from the existing semantic region retrieval approaches, we treated semantic regions as actual regions, rather than semantically relevant points. We assumed that each semantic region had a continuous distribution, and thus our models had more expressive power than the previous approaches. Furthermore, both the sest-hiHMM and the sticky sest-hiHMM encompassed HMMs to capture the temporal dependency of observations in a trajectory. Our models, however, did not concern the spatial and temporal dependency between trajectories. The dependency is considered to be important because two trajectories that are spatially and temporally close are likely to pass through similar semantic regions. We will explore how to extend our models to capture the spatial and temporal dependency between trajectories.

There are some future works to enhance the results of our work. Firstly the area of each semantic region should be regulated. As we discussed in Section 5.3, both our models have a tendency to generate smaller semantic regions, but too small ones may not be useful for further analysis. The constraint may be given by priors or convergence conditions. Another work is to eliminate more redundant semantic regions than a sticky sest-hiHMM, by concerning a stricter modification. Furthermore, we plan to improve the performance of posterior inference. We adopted beam sampling (Van Gael et al., 2008), but it suffered from the slow mixing rate because of the auxiliary variables (Fox et al., 2011). It can be resolved by adopting other algorithms, such as an infinite-state particle Gibbs algorithm (Tripuraneni, Gu, Ge, & Ghahramani, 2015). Lastly, trajectories and semantic regions may be modeled by other base models, rather than hiHMMs. We need to consider other nonparametric Bayesian models, such as beta process hidden Markov models (Sun et al., 2015) or infinite Beta-Liouville mixture models (Fan & Bouguila, 2015).

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).

References

Acevedo-Rodríguez, J., Maldonado-Bascón, S., López-Sastre, R., Gil-Jiménez, P., & Fernández-Caballero, A. (2011). Clustering of trajectories in video surveillance using growing neural gas. *Lecture Notes in Computer Science, 6686*, 461–470.

Akbarzadeh, V., Gagné, C., & Parizeau, M. (2015). Kernel density estimation for target trajectory prediction. In *Proceedings of the 2015 IEEE/RSJ international conference on intelligent robots and systems* (pp. 3449–3456).

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 961–971).

Ali, I., & Dailey, M. N. (2012). Multiple human tracking in high-density crowds. *Image and Vision Computing, 30*(12), 966–977.

Arroyo, R., Yebes, J. J., Bergasa, L. M., Daza, I. G., & Almazán, J. (2015). Expert video–surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Systems with Applications, 42*(21), 7991–8005.

Bae, C., Kang, K., Liu, G., & Chung, Y. Y. (2016). A novel real time video tracking framework using adaptive discrete swarm optimization. *Expert Systems with Applications, 64*, 385–399.

Bae, S., & Yoon, K. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the 2014 IEEE conference on computer vision and pattern recognition* (pp. 1218–1225).

Ballan, L., Castaldo, F., Alahi, A., Palmieri, F., & Savarese, S. (2016). Knowledge transfer for scene-specific motion prediction. In *Proceedings of the 14th European conference on computer vision* (pp. 697–713).

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2001). The infinite hidden markov model. *Advances in Neural Information Processing Systems, 14*, 577–584.

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference* PhD. Thesis. University College London, Gatsby Computational Neuroscience Unit.

Blei, D. M., Ng. , A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35.

Brun, L., Saggese, A., & Vento, M. (2014). Dynamic scene understanding for behavior analysis based on string kernels. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(10), 1669–1681.

Cancela, B., Ortega, M., Fernández, A., & Penedo, M. G. (2013). Hierarchical framework for robust and fast multiple-target tracking in surveillance scenarios. *Expert Systems with Applications, 40*(4), 1116–1131.

Choi, J., Moon, D., & Yoo, J. (2015). Robust multi-person tracking for real-time intelligent video surveillance. *ETRI Journal, 37*(3), 551–561.

Chung, P.-C., & Liu, C.-D. (2008). A daily behavior enabled hidden Markov model for human behavior understanding. *Pattern Recognition, 41*(5), 1572–1580.

Cui, J., Liu, W., & Xing, W. (2014). Crowd behaviors analysis and abnormal detection based on surveillance data. *Journal of Visual Languages and Computing, 25*(6), 628–636.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 2625–2634).

Elbahri, M., Taleb, N., Kpalma, K., & Ronsin, J. (2016). Parallel algorithm implementation for multi-object tracking and surveillance. *IET Computer Vision, 10*(3), 202–211.

Emonet, R., Varadarajan, J., & Odobez, J.-M. (2011). Extracting and locating temporal motifs in video scenes using a hierarchical non parametric Bayesian model. In *Proceedings of the 2011 IEEE conference on computer vision and pattern recognition* (pp. 3233–3240).

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90*(430), 577–588.

Fan, W., & Bouguila, N. (2015). Expectation propagation learning of a Dirichlet process mixture of Beta-Liouville distributions for proportional data clustering. *Engineering Applications of Artificial Intelligence, 43*, 1–14.

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics, 5*(2A), 1020–1056.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, W., Tian, G., Li, X., & Maybank, S. (2013). An improved hierarchical Dirichlet process-hidden markov model and its application to trajectory modeling and retrieval. *International Journal of Computer Vision, 105*(3), 246–268.

Iwashita, Y., Ryoo, M. S., Fuchs, T. J., & Padgett, C. (2013). Recognizing humans in motion: Trajectory-based aerial video analysis. In *Proceedings of the 24th British machine vision conference: 127* (pp. 1–127). 11.

Jeong, H., Yoo, Y., Yi, K., & Choi, J. (2014). Two-stage online inference model for traffic pattern analysis and anomaly detection. Machine Vision and Applications, 25(6), 1501-1517. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 4328–4336).

Jung, C. R., Hennemann, L., & Musse, S. R. (2008). Event detection using trajectory clustering and 4-D histograms. *IEEE Transactions on Circuits and Systems for Video Technology, 18*(11), 1565–1575.

Kim, H., Kim, D., Lee, J., & Jeong, I. (2015a). Uncooperative person recognition based on stochastic information updates and environment estimators. *ETRI Journal, 37*(2), 395–405.

Kim, J., Jang, G., Kim, G., & Kim, M. (2015b). Crowd activity recognition using optical flow orientation distribution. *KSII Transactions on Internet and Information Systems, 9*(8), 2948–2963.

Kitani, K. M., Ziebart, B. D., Bagnell, J. A., & Hebert, M. (2012). Activity forecasting. In *Proceedings of the 12th European conference on computer vision* (pp. 201–214).

Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proceedings of the 2011 IEEE conference on computer vision and pattern recognition* (pp. 1446–1453).

Kumar, D., Bezdek, J. C., Rajasegarar, S., Leckie, C., & Palaniswami, M. (2017). A visual-numeric approach to clustering and anomaly detection for trajectory data. *The Visual Computer, 33*(3), 265–281.

Kwon, Y., Jin, J., Moon, J., Kang, K., & Park, J. (2016). Trajectory segmentation based on spatio-temporal locality with multidimensional index structures. In *Proceedings of the 2016 international conference on collaboration technologies and systems* (pp. 212–217).

Lei, B., & Xu, L.-Q. (2006). Real-time outdoor video surveillance with robust foreground extraction and object tracking via multi-state transition management. *Pattern Recognition Letters, 27*(15), 1816–1825.

Li, X., Hu, W., Zhang, Z., Zhang, X., & Luo, G. (2008). Trajectory-based video retrieval using Dirichlet process mixture models. In *Proceedings of the 19th British machine vision conference: 106* 1-106.10.

Lim, M. K., Tang, S., & Chan, C. S. (2014). iSurveillance: Intelligent framework for multiple events detection in surveillance videos. *Expert Systems with Applications, 41*(10), 4704–4715.

Mithun, N. C., Howlader, T., & Rahman, S. M. M. (2016). Video-based tracking of vehicles using multiple time-spatial images. *Expert Systems with Applications, 62*, 17–31.

Morris, B. T., & Trivedi, M. M. (2011). Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(11), 2287–2301.

Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 4293–4302).

Ohnishi, K., Kanehira, A., Kanezaki, A., & Harada, T. (2016). Recognizing activities of daily living with a wrist-mounted camera. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 3103–3111).

Park, S., Lee, K., & Yoon, K. (2016). Robust online multiple object tracking based on the confidence-based relative motion network and correlation filter. In *Proceedings of the 2016 IEEE international conference on image processing* (pp. 3484–3488).

Piciarelli, C., & Foresti, G. L. (2006). On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters, 27*(15), 1835–1842.

Piergiovanni, A., Fan, C., & Ryoo, M. S. (2017). Learning latent sub-events in activity videos using temporal attention filters. In *Proceedings of the 31st AAAI conference on artificial intelligence*. in press.

Pintea, S. L., van Gemert, J. C., & Smeulders, A. W. M. (2014). Déj'a Vu: Motion prediction in static images. In *Proceedings of the 13th European conference on computer vision*. Part III (pp. 172–187).

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

Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, E., & Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 3043–3053).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*, 91–99.

Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding, 104*(2-3), 232–248.

Shao, Z., & Li, Y. (2015). Integral invariants for space motion trajectory matching and recognition. *Pattern Recognition, 48*(8), 2418–2432.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems, 27*, 568–576.

Sohn, K., Ho, J. W. K., Djordjevic, D., Jeong, H., Park, P. J., & Kim, J. (2015). hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics, 31*(13), 2066–2074.

Sun, S., Zhao, J., & Gao, Q. (2015). Modeling and recognizing human trajectories with beta process hidden Markov models. *Pattern Recognition, 48*(8), 2407–2417.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., & Anguelov, D. (2015). Going deeper with convolutions. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association, 101*(476), 1566–1581.

Tripuraneni, N., Gu, S., Ge, H., & Ghahramani, Z. (2015). Particle Gibbs for infinite hidden Markov models. *Advances in Neural Information Processing Systems, 28*, 2395–2403.

Van Gael, J., Saatci, Y., Teh, Y. W., & Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th international conference on machine learning* (pp. 1088–1095).

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Anticipating visual representations from unlabeled video. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 98–106).

Walker, J., Gupta, A., & Herbert, M. (2014). Patch to the future: Unsupervised visual prediction. In *Proceedings of the 2014 IEEE conference on computer vision and pattern recognition* (pp. 3302–3309).

Wang, X., Ma, K. T., Ng, G.-W., & Grimson, W. E. L. (2011). Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International Journal of Computer Vision, 95*(3), 287–312.

Wang, Y., & Mori, G. (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(10), 1762–1774.

Wu, Z., Fu, Y., Jiang, Y.-G., & Sigal, L. (2016). Harnessing object and scene semantics for large-scale video understanding. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 3112–3121).

Xu, H., Zhou, Y., Lin, W., & Zha, H. (2015). Unsupervised trajectory clustering via adaptive multi-kernel-based shrinkage.

Yi, S., Li, H., & Wang, X. (2015). Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the 2015 IEEE conference on computer vision and pattern recognition* (pp. 3488–3496).

Yildirim, M. E., Ince, I. F., Salman, Y. B., Song, J., Park, J., & Yoon, B. (2016). Direction-based modified particle filter for vehicle tracking. *ETRI Journal, 38*(2), 356–365.

Yoo, Y., Yun, K., Yun, S., Hong, J., Jeong, H., & Choi, J. (2016). Visual path prediction in complex scenes with crowded moving objects. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 2668–2677).

Yoon, J., Lee, C., Yang, M., & Yoon, K. (2016). Online multi-object tracking via structural constraint event aggregation. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 1392–1400).

Zhou, B., Wang, X., & Tang, X. (2011). Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Proceedings of the 2011 IEEE conference on computer vision and pattern recognition* (pp. 3441–3448).

Zou, J., Chen, X., Wei, P., Han, X., & Jiao, J. (2013). A belief based correlated topic model for semantic region analysis in far-field video surveillance systems. In *Proceedings of the 14th pacific-rim conference on multimedia* (pp. 779–790).