Nonparametric Discovery of Human Routines from Sensor Data

Feng-Tso Sun*, Yi-Ting Yeh†, Heng-Tze Cheng*, Cynthia Kuo*, Martin Griss*
*Electrical and Computer Engineering Department, Carnegie Mellon University
{fengtso, hengtze}@cmu.edu, cynthia@vibradotech.com, martin.griss@sv.cmu.edu
†Computer Science Department, Stanford University
yitingy@stanford.edu

***Abstract*—People engage in routine behaviors. Automatic routine discovery goes beyond low-level activity recognition such as sitting or standing and analyzes human behaviors at a higher level (e.g., commuting to work). With recent developments in ubiquitous sensor technologies, it becomes easier to acquire a massive amount of sensor data. One main line of research is to mine human routines from sensor data using parametric topic models such as latent Dirichlet allocation. The main shortcoming of parametric models is that it assumes a fixed, pre-specified parameter regardless of the data. Choosing an appropriate parameter usually requires an inefficient trial-and-error model selection process. Furthermore, it is even more difficult to find optimal parameter values in advance for personalized applications.**

In this paper, we present a novel nonparametric framework for human routine discovery that can infer high-level routines without knowing the number of latent topics beforehand. Our approach is evaluated on public datasets in two routine domains: a 34-daily-activity dataset and a transportation mode dataset. Experimental results show that our nonparametric framework can automatically learn the appropriate model parameters from sensor data without any form of model selection procedure and can outperform traditional parametric approaches for human routine discovery tasks.

I. Introduction

Recent advances in sensor technologies and the growing interest in context-aware applications, such as detecting changes in activties for elderly care, targeted advertising, and location-based services have led to a demand for understanding human behavior patterns from sensor data.

Many techniques have been proposed to perform low-level activity recognition such as walking, sitting, or opening a door. This performs using various kinds of sensor data such as motion data [1], GPS/Bluetooth/WiFi signals [2], ambient sound [3], and RFID-tagged objects [4]. While low-level activity recognition provides us building blocks to understand human behaviors, they are not enough to convery more high-level semantic meanings. For example, people usually describe what they did during the day using high-level routines (e.g., "commuting to work" or "having lunch") rather than a sequence of low-level activities such as "walk-walk-stand-sit-sit-run-walk-stand". This leads to a main line of research focusing on modeling a higher-level characterization of human behaviors from sensor data, namely at the routine level.

More specifically, routine discovery is about extracting temporal regularities in people's daily lives [5]. A routine can be seen as a composition of multiple low-level activities. Multiple low-level activities can occur within the same routine. Different routines may contain the same kinds of low-level activities, but with different proportions. For example, the "`Grocery Shopping`" routine may invovle more "standing" and "walking" activities compared to the "`Office Work`" routine.

Most existing approaches for automatic routine discovery are built on parametric topic models such as latent Dirichlet allocation (LDA) [6]. In a topic model for text mining, a document is a mixture of a number of hidden topics which can be represented by a multinomial topic proportion. Topic models were initially designed for text mining, but they are also effective in extracting understandable human behavior patterns [7][5].

We can make an analogy between text and sensor data. In the context of mining a sequence of sensor data, sensor data features are first mapped into a set of discrete *labels* (vocabulary). Each mapped data feature becomes a word. Then, the bag of words in each temporal window (document) is used to train the topic model. Documents belong to the same routine if they have similar topic proportions. In the parametric setting, the above procedure requires two types of parameters to be predefined: the size of vocabulary and the number of latent topics. Typically, they are chosen in a trial-and-error fashion [7][5].

However, for routine discovery, such parameter specification poses several challenges. First of all, the best parameter values for personalized models may be different for different users. For example, due to the fact that different people usually have very distinct behavior patterns based on their lifestyles, job types, or ages, their routine patterns may require different number of latent topics to model appropriately. Moreover, even for a single user, it is possible that her behavior patterns change over time. The best parameter values must also be adjusted accordingly. Hence, we need the model to automatically select parameter values based on individual users' behavior patterns.

In this paper, we propose a novel human routine discovery framework using nonparametric Bayesian methods. The goal is to avoid declaring the number of activities and routines in a person's daily life beforehand in parametric settings (see Figure 1). Our framework consists of two phases: vocabulary extraction and routine extraction. During the first

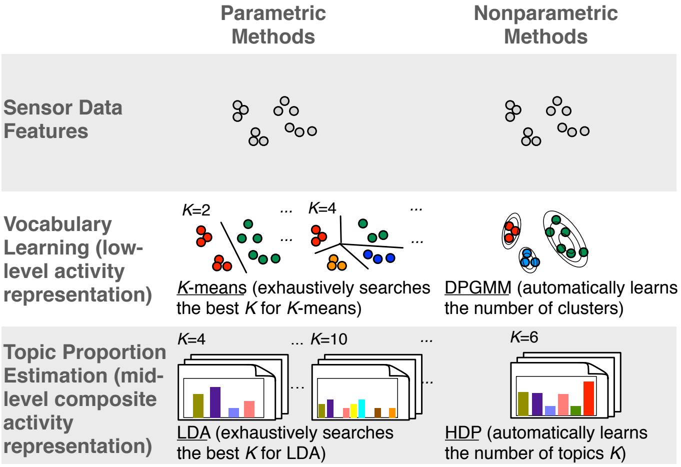

Parametric Methods
Nonparametric Methods
Sensor Data Features
Vocabulary Learning (low-level activity representation)
K=2
...
K=4
...
K-means (exhaustively searches the best K for K-means)
DPGMM (automatically learns the number of clusters)
Topic Proportion Estimation (mid-level composite activity representation)
K=4
...
K=10
...
K=6
LDA (exhaustively searches the best K for LDA)
HDP (automatically learns the number of topics K)

Fig. 1. Comparison of parametric and nonparametric methods for routine discovery. For parametric methods, the size of vocabulary and the number of topics must be specified in advance. Since the appropriate parameters are difficult to specify initially, searching through the parameter space to achieve the best performance is necessary. Our proposed framework uses nonparametric methods to automatically select the proper size of vocabulary and the number of topics from the data.

phase (vocabulary extraction), we build up the vocabulary and automatically determine the size of vocabulary (low-level activity representation) from raw sensor data using the Dirichlet process Gaussian mixture model (DPGMM). In the second phase (routine extraction), we infer topic proportions (high-level routine representation) for each document with the automatically determined number of latent topics using hierarchical Dirichlet process (HDP) and extract latent routines. We demonstrate the effectiveness of our framework with two public real-world datasets for the routine discovery tasks. Experimental results indicate that HDP automatically selects the appropriate number of latent topics and achieves better clustering performance compared to LDA, with the best number of routines setting. Moreover, DPGMM outperforms parametric clustering method (i.e., K-means) for the vocabulary extraction in terms of routine clustering performance.

The main contributions of this paper are summarized as follows: (1) We design and implement a nonparametric framework that can learn a low-level activity vocabulary with automatically chosen size and discover high-level routines from sensor data without declaring the number of latent topics in advance. (2) We conduct experiments on public datasets in two routine domains: a 34-daily-activity dataset and a transportation mode dataset to demonstrate the effectiveness of our approach in discovering human routines from sensor data.

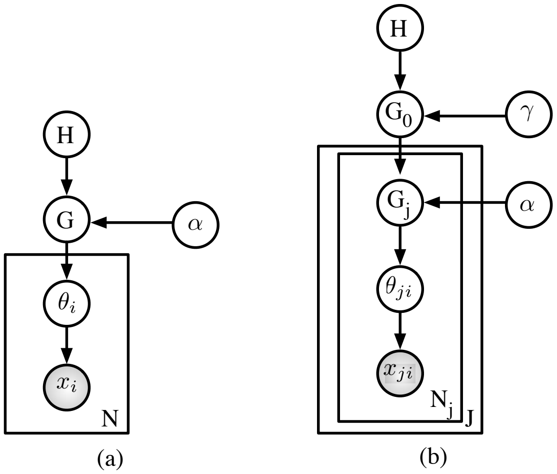
In Section II, we describe related work in the area of activity recognition and human routine discovery. In Section III, we review the background of nonparametric Bayesian methods. Section IV introduces and analyzes the characteristics of two public real-world datasets: a 34-daily-activity dataset and a GPS trajectory dataset. In Section V, we show how our human routine discovery framework can be applied on these two datasets. Experimental results are given in Section VI. Finally, conclusion and future research direction are presented in Section VII.

II. Related Work

In the activity recognition field, researchers have developed and applied several machine learning paradigms to recognize low-level human activities (e.g., sitting, standing, or walking) from various types of sensor data. First, supervised learning approaches [1][8][9] require that the training data are completely labeled with their ground truth activity labels. Second, zero-shot learning [10] and semi-supervised learning [11][12] aim to learn models based on partially labeled data. That is, the training data consist of both labeled and unlabeled data. However, ground truth labels sometimes are simply unavailable. Finally, unsupervised learning focuses on performing pattern discovery and clustering based on the similarity of observed samples without providing ground truth labels [7][13]. Our method is a type of unsupervised learning.

Beyond low-level activity recognition, extracting routines (e.g., dining, office work, or taking bus) has received attention because routine information provides high-level semantics for understanding human behaviors. Eagle et al. [14] use principal component analysis (PCA), a general-purpose dimensionality reduction method, to obtain main components that construct human daily routines from the MIT Reality Mining dataset. However, PCA's eigen-vectors have no explicit semantic meanings and they do not encode uncertainty in human daily routines at a specific time point. Alternatively, researchers have been looking at the problem from a probabilistic view point. Huynh et al. [7] discover daily routines from wearable sensor data using K-means clustering to build activity vocabulary and using LDA to learn topic proportions for each time window. Farrahi et al. [5] also apply LDA on labeled cell tower data to automatically discover routines, including "being at work" or "going home from work". Zheng et al. [15] propose a probabilistic generative model for learning users' latent behavior patterns based on unlabeled cell tower data. However, these methods all require parameter selection such as the size of vocabulary, the number of topics, and the number of typical states of a user. Parameter selection is usually done via trial and error, cross validation, or perplexity measurement with a wide range of settings for the parameters on the training data. They assume that the model complexity is fixed. In contrast, our routine discovery framework is nonparametric such that the parameters are automatically selected and it allows the model complexity to change as more data are available.

Recently, the concept of nonparametric methods has shown promise in the field of mobile computing. For example, Hu et al. [16] solve low-level abnormal activity recognition problem by using hierarchical Dirichlet process hidden Markov model (HDP-HMM) to automatically decide the right number of states for HMM. Similarly, Zhu et al. [17] segment a small number of activities using HDP-HMM. Nguyeh et al. [18] apply HDP model to extract users' proximity patterns from sociometric badge data. Our work extends the previous work



(a) (b)

Fig. 2. Graphical models of (a) the Dirichlet process mixture model and (b) the hierarchical Dirichlet process.

by proposing a new nonparametric framework for human routine discovery that can construct low-level activity primitives and extract high-level routines from raw sensor data without the need of model selection procedures.

III. Background of Nonparametric Bayesian Modeling

In this section, we review two nonparametric Bayesian methods: the Dirichlet process mixture model (DPMM) and the hierarchical Dirichlet process (HDP).

A. Dirichlet Process

The basic building block of DPMM and HDP is the Dirichlet process which is an infinite-dimensional generalization of the Dirichlet distribution [19].

A Dirichlet distribution is a distribution over multinomial distributions. It is parameterized by a vector of $\{\alpha_1, \ldots, \alpha_m\}$. A random vector $(\pi_1, \ldots \pi_m)$ $(\sum_{k=1}^{m} \pi_k = 1)$ is Dirichlet distributed if

$$P(\pi_1, \ldots \pi_m) = \frac{\Gamma(\Sigma_k \alpha_k)}{\prod_k (\Gamma(\alpha_k))} \prod_{k=1}^{m} \pi_k^{(\alpha_k - 1)} \tag{1}$$

where $\alpha_1, \ldots, \alpha_m > 0$. The support of an m-dimensional Dirichlet distribution is the $m-1$-dimensional probability simplex.

The random vector $(\pi_1, \ldots \pi_m) \sim Dirichlet(\alpha_1, \ldots, \alpha_m)$ defines the possible parameters for a multinomial distribution on the discrete space $\Theta = \theta_1, ..., \theta_m$ such that $P(\theta = \theta_i) = \pi_i$. From Eq (1), we see that, if $(\alpha_1, \ldots, \alpha_m) = (1, \ldots, 1)$, the distribution of $(\pi_1, \ldots \pi_m)$ is uniform.

The Dirichlet process generalizes the Dirichlet distribution, and it is denoted as $DP(\alpha, H)$, where α is the concentration parameter and H is the base distribution. Like the Dirichlet distribution, the Dirichlet process is also a distribution over distributions. A distribution $G(\theta)$ is Dirichlet process distributed if $G \sim DP(\alpha, H)$. Mathematically, it can be written as

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta = \theta_k) \tag{2}$$

where $\theta_k \sim H$ and $\delta(\theta = \theta_k)$ is a Dirac delta function at θ_k. The infinite sequence of mixture weights π_k can be constructed by the stick-breaking scheme [20]:

$$\begin{aligned} &\beta_k \sim Beta(1, \alpha) \\ &\pi_1 = \beta_1 \\ &\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \text{ for } k = 2, 3, \ldots. \end{aligned} \tag{3}$$

From the above equations, we see that α defines how concentrated the distribution of Θ is. For a small value of α, only the first few π_k will have significant weights. Higher α value results in mixture weights that are distributed more uniformly. Also, as k increases, the weights tend to decrease.

B. Dirichlet Process Mixture Model

DPMM generalizes traditional finite mixture models by allowing the number of mixture components to be infinite. Let's start by considering the finite case. In traditional finite mixture models, it is assumed that the number of mixture components is given. For example, let $x_1, \ldots, x_N$ be N observation data points, and they are exchangeable. In a mixture model, a data point x_i is assumed to be drawn from the distribution $p(x) = \Sigma_{k=1}^{K} \pi_k f(x|\theta_k)$ where K is the number of mixture components, π_k is the mixture weight of component k, and $f(x|\theta_k)$ is the mixture component parameterized by θ_k. For example, one common choice for the distribution of the mixture components is the Gaussian distribution, which is parameterized by mean and variance. The mixture weights sum to one.

In the mixture model problem, given the observation data points, it is convenient to introduce a latent discrete random variable, c_i, associated with each data point. It is often referred to as the indicator variable, whose domain is $\{1, \ldots, K\}$. It specifies which component the corresponding data point belongs to.

Therefore, the generative process of the finite mixture model can alternatively be described by:

$$\begin{aligned} &p(c_i = k) = \pi_k \\ &x_i \sim f(x|\theta_k). \end{aligned} \tag{4}$$

This describes how each data point x_i has been generated by first sampling the component id c_i, and then sampling from the distribution of that mixture component.

Since the mixture weight $\pi = \pi_1, ..., \pi_k$ is multinomial distribution, it is convenient to use the Dirichlet distribution as the prior. We can also use the Dirichlet distribution to construct the finite mixture model by the following steps:

$$\begin{aligned} \theta_{c_i} &\sim H \text{ for } c_i = \{1, \ldots, K\} \\ (\pi_1, \ldots, \pi_K) &\sim Dirichlet(\alpha/K, \ldots \alpha/K) \\ c_i &\sim Multinomial(\pi_1, \ldots \pi_K) \\ x_i &\sim f(x|\theta_{c_i}). \end{aligned} \tag{5}$$

where H is the base distribution encoding the prior beliefs about the parameters of the mixture components.

Next, consider the limiting case where $k \to \infty$, so that the mixture model becomes

$$p(x) = \sum_{k=1}^{\infty} \pi_k f(x|\theta_k). \tag{6}$$

The model described in Eq (5) then becomes

$$\begin{aligned} G &\sim DP(\alpha, H) \\ \theta_i &\sim G \\ x_i &\sim f(x|\theta_i). \end{aligned} \tag{7}$$

Here, the Dirichlet distribution becomes the Dirichlet process. Therefore, the infinite mixture model is also called the Dirichlet process mixture model. In DPMM, each component is still described by some set of parameters. These parameters come from the Dirichlet process.

Figure 2(a) illustrates the graphical model of DPMM. First, the prior distribution function G is drawn from a Dirichlet process $G \sim DP(\alpha, H)$ where α is the concentration parameter and H is the base prior. Second, given G, we sample θ_i, the parameters for the component that x_i belongs to. Finally, given the parameters θ_i, we generate each data point x_i.

In this paper, we perform inference of DPMM by using the Gibbs sampling algorithm [21]. It starts by randomly initializing c_i's for all x_i's and then iterates the following steps:

1) Pick a data point x_i.
2) Sample its corresponding indicator variable c_i conditioned on fixing all other indicator variables $\{c_{-i}\}$ using the Chinese restaurant process [22]:

$$\begin{aligned} P(c_i = k, k \leq K | c_{-i}, \alpha) &= \frac{n_k}{\alpha + N - 1} f(x_i|\theta_k) \\ P(c_i = K+1 | c_{-i}, \alpha) &= \frac{\alpha}{\alpha + N - 1} f_{K+1}(x_i) \end{aligned} \tag{8}$$

   where N is the total number of data points, n_k is the number of data points being assigned to component k, and $f_{K+1}(x_i) = \int f(x_i|\theta)H(\theta)d\theta$. Note that with probability $\frac{\alpha}{(\alpha+N-1)}$, x_i is assigned to a new component $K+1$.
3) If we get a new component, we can draw its corresponding parameter values θ_{K+1} by

$$P(\theta_{K+1}|x_i) \propto f(x_i|\theta_{K+1})H(\theta_{K+1}). \tag{9}$$

C. Hierarchical Dirichlet Process

Recall that in finite mixture models, we assume that data points are drawn from a distribution consisting of K components. A topic model is one type of mixture model. A topic model is used to model a document using a mixture of a number of topics where a topic is a multinomial distribution over the vocabulary. The generative process of the topic model is achieved by first selecting a topic from the topic distribution. Then, choose a word from the word distribution defined by the topic. One particular topic model is the latent Dirichlet allocation (LDA) where the topic distributions and the word distributions both have the Dirichlet prior. In particular, each document has its own topic distribution over a finite number of topics.

In HDP, we relaxed the assumption that the number of topics is known. Thus, HDP can also be thought of as a nonparametric generalization of the LDA. Figure 2(b) illustrates the graphical model of HDP, which consists of two levels of Dirichlet processes.

Assume that each document is indexed by $j = 1, \ldots, J$ and that x_{ji} denotes the ith word in document j. This generative model of HDP can be described as:

$$\begin{aligned} &G_0 \sim DP(\gamma, H) \\ &G_j \sim DP(\alpha, G_0) \text{ , for } j = 1, \ldots, J \\ &\theta_{ji} \sim G_j \text{ , for } i = 1, \ldots, N_j \\ &x_{ji} \sim f(x|\theta_{ji}). \end{aligned} \tag{10}$$

In the upper level, G_0 is the distribution of an infinite mixture of topics for all documents. It is drawn from a Dirichlet process with the base distribution H and the concentration parameter γ. In the lower level, G_j defines the mixture of topics of document j. The words are generated by repeatedly drawing from the corresponding topic distribution.

From the graphical model, we can see that the base distribution of G_j is G_0 which is also a random draw from another Dirichlet process. The reason that we need two levels of Dirichlet processes is because we want G_0 to be the common base distribution that is shared across all G_j. In this way, we achieve the goal that different documents share the same mixture components but with different mixture weights.

Recall that in DPMM, the Dirichlet process allows us to avoid pre-specifying the number of mixture components. In HDP, the number of topics activated in G_j also need not to be specified.

This concept of topic modeling can be applied in a general grouped data setting. For example, when mining a sensor data stream, we can consider the sensor data within a time window as a group. Each group consists of a number of sensor readings.

IV. Dataset Description

To evaluate our nonparametric Bayesian framework for routine discovery, we experimented with two realistic and published datasets including daily life routines and transportation modes [7][23].

A. Daily Life Routine Dataset

The daily life routine dataset was released by Technische Universitat Darmstadt (TU Darmstadt). It contains 34 daily low-level activity classes (including the unlabeled class). In addition to the low-level activity class annotations, this dataset also provides 4 high-level routine class annotations (i.e., commuting, lunch, office work, and dinner). The sensor data were collected from two wearable 3-axis accelerometers worn on the right hip pocket and the dominant wrist. The accelerometer data were sampled at a rate of 100Hz, and the features (i.e., mean and variance of acceleration of each axis) were computed over a window of 0.4 seconds (i.e., 2.5Hz).

B. Transportation Mode Dataset

The transportation mode dataset was collected by Microsoft Research Asia (MSRA). This dataset contains GPS trajectory data from 182 users in a period of five years. A GPS trajectory is represented by a sequence of timestamped coordinates including longitude, latitude, and altitude. The GPS data were sampled around every two seconds (i.e., 0.5Hz). Each GPS trajectory was provided with a specific transportation mode annotation such as bike, walk, bus, and subway. In this paper, we extracted a week of GPS trajectory data from one user for the routine discovery task.

V. Approach of Routine Discovery

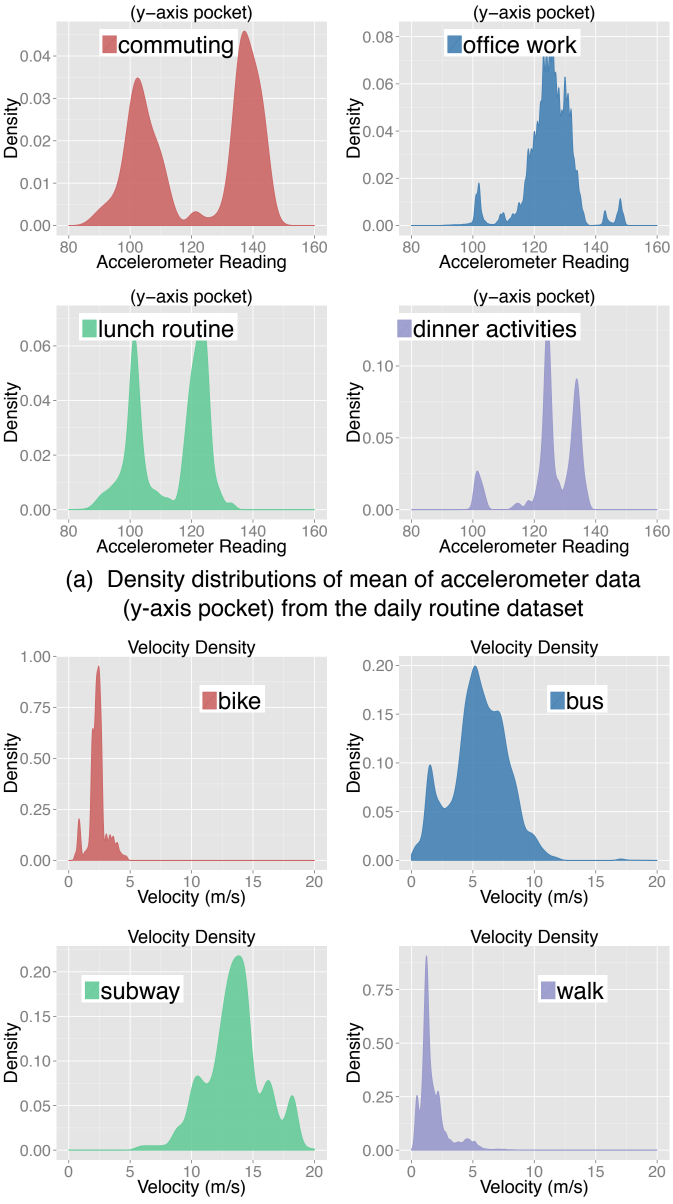
In this section, we describe our nonparametric Bayesian framework for routine discovery. First, we extract low-level signal features from raw sensor data such as accelerometer and GPS data (Section V-A). Second, we describe how to map feature vectors into artificial words using DPGMM (Section V-B). Finally, a set of artificial words over a period of time window are grouped as a document for routine discovery using HDP (Section V-C). The whole pipeline is illustrated in Figure 4.

A. Feature extraction

Our framework is agnostic to the type of input sensor data. Once the low-level features are extracted from raw sensor data, we input them to the activity vocabulary extraction module.

For the daily life routine dataset, we use the *mean* and *standard deviation* of accelerometer data in dimension x, y, and z from wrist and pocket sensors. These features have been proven effective in previous work [7]. Figure 3 (a) shows an example of feature distributions across four different daily routine classes. We see that the histograms of the same feature type behave quite differently across different routine classes.

The features for the transportation mode dataset include *velocity*, *heading direction change rate*, and *stop rate* derived from the GPS trajectory data. Previous work identified that this set of GPS features are robust to traffic conditions [23]. Figure 3 (b) illustrates the distributions of GPS feature (i.e., velocity) across four different transportation modes.



(a) Density distributions of mean of accelerometer data (y-axis pocket) from the daily routine dataset

(b) Density distributions of velocity from the transportation mode dataset

Fig. 3. Examples of feature density distributions across different ground truth routines. (a) Density distributions of the mean of the pocket accelerometer data in dimension y from the daily routine dataset. (b) Density distributions of the velocity feature from the GPS trajectory dataset. This suggests that feature density distributions of different routines can be modeled by Gaussian mixture models with different numbers of components.

B. Learning vocabulary with Dirichlet process Gaussian mixture model

Next, we describe how we use DPGMM to infer the set of discrete artificial words from the feature vectors in the context of the daily life routine dataset.

Recall that in the daily life routine dataset, each data point is represented by a 12-dimensional feature vector ($\mu_{x-pocket}$, $\sigma_{x-pocket}$, ... , $\sigma_{z-wrist}$). Assume that there are N number of data points. One way to model these N data points with

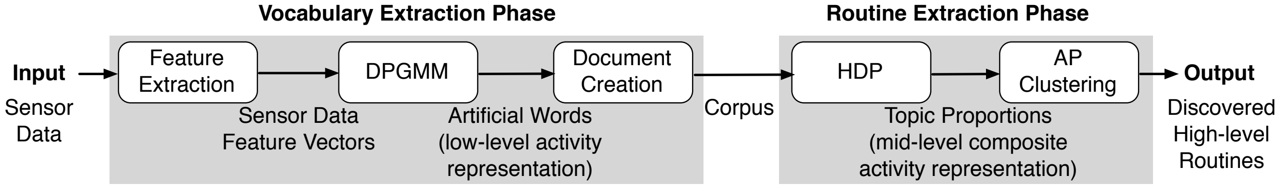

Vocabulary Extraction Phase
Routine Extraction Phase
Input
Sensor Data
Feature Extraction
DPGMM
Document Creation
Sensor Data Feature Vectors
Artificial Words (low-level activity representation)
Corpus
HDP
AP Clustering
Topic Proportions (mid-level composite activity representation)
Output
Discovered High-level Routines

Fig. 4. Overview of our two-phase nonparametric routine discovery framework. In the first phase, starting from the input sensor data, we extract feature vectors, build up vocabulary using DPGMM, and then transform sensor data stream to a collection of documents. In the second phase, we discover routines by clustering the topic proportion vectors of the documents learned from HDP. Note that the size of the vocabulary, the number of topics, and the number of routines are learned automatically.

DPGMM is to use 12-dimensional Gaussian distributions as the component distributions. Each component is parameterized by a 12-dimensional mean vector and a 12×12 covariance matrix. However, a large number of parameters in the covariance matrix would create the problem of overfitting [24].

Therefore, we use the idea of dividing the feature space into lower-dimensional feature subspaces, fitting a different DPGMM for each subspace, and then combine the results together. More specifically, we organize the feature vectors into 6 subspaces by their sensor types and axis types. For example, one subspace corresponds to data points collected by the pocket sensor in the x-axis, $(\mu_{x-pocket}, \sigma_{x-pocket})$.

To simplify the notation, we use $\mathbf{t}_{ij}$ to denote the ith data point in subspace j where $i = 1, \dots N$ and $j = 1, \dots, 6$. The use of DPGMM to fit the data in subspace j can be formulated as

$$\begin{aligned} G_\mu &\sim DP(\alpha, \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{\gamma}^{-1})) \\ G_S &\sim DP(\alpha, Gamma(\boldsymbol{\beta}, \boldsymbol{\omega}^{-1})) \\ \boldsymbol{\mu}_{c_{ij}} &\sim G_\mu \\ \boldsymbol{S}_{c_{ij}} &\sim G_S \\ \boldsymbol{t}_{ij}|\theta_{c_{ij}} &\sim \mathcal{N}(\boldsymbol{\mu}_{c_{ij}}, S_{c_{ij}}) \end{aligned} \tag{11}$$

where c_{ij} is an indicator variable specifying the cluster associated with $\boldsymbol{t}_{ij}$, and $\{\boldsymbol{\mu}_{c_i}, \boldsymbol{S}_{c_i}\}$ is the set of parameters of the Gaussian component for cluster c_{ij}. $\boldsymbol{\mu}_{c_{ij}}$ and $\boldsymbol{S}_{ij}$ are generated by Dirichelet processes with base distributions Gaussian and Gammas, respectively. Four hyperparameters $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ specify the base distributions expressing the strength of the prior belief on the distribution of the parameter space.

During inference, we run Gibbs sampling for 500 iterations with a burn-in period of 100 iterations to infer c_{ij} for each t_{ij} [21]. Finally, for each data point, we concatenate the corresponding cluster assignments from the 6 subspaces to form a discrete artificial word $w_i = (c_{i1}, \dots, c_{i6})$. Note that we do not need to specify the number of unique artificial words (vocabulary size) beforehand.

Similarly, for the transportation mode dataset, we applied DPGMM for each of the three feature vectors (i.e., *velocity*, *heading direction change rate*, and *stop rate*) to construct artificial words representing the data points.

C. *Discovering routines with hierarchical Dirichlet process*

Based on the artificial word representation for the sensor data, we now describe how to construct documents and extract routines using HDP.

To construct documents from a stream of artificial words, we represent each document as a histogram of artificial word occurrences in a sliding window with overlapping. For the daily routine dataset, the sliding windows are set 30 minute duration with 2.5 minute overlapping, similar to previous work by Huynh et al. [7]. For the transportation mode dataset, the sliding window is 10 minutes duration with 1 minute overlap.

Using a Gibbs sampling scheme similar in the inference stage of DPGMM, we obtain the mixture proportion of latent topics for each document.

Finally, we cluster document topic proportions using the affinity propagation (AP) algorithm [25]. AP algorithm takes the similarity of pairs of data points (topic proportion vectors) as input and forms clusters by finding data points that are representative of clusters. We use Jensen-Shannon divergence as the distance function. Let π_i and π_j denote the topic proportions of document i and j. Their Jensen-Shannon divergence is defined by

$$JSD(\pi_i, \pi_j) = \frac{1}{N} D_{KL}(\pi_i||M) + \frac{1}{2} D_{KL}(\pi_j||M) \tag{12}$$

where $M = \frac{1}{2}(\pi_i + \pi_j)$ and $D_{KL}(P||Q)$ is the Kullback-Leibler divergence which is formulated as $D_{KL}(P||Q) = \sum_i ln(\frac{P(i)}{Q(i)})P(i)$.

The similarity between two topic proportions are then computed by

$$D(\pi_i, \pi_j) = e^{-JSD(\pi_i, \pi_j)} \tag{13}$$

Each cluster corresponds to a discovered routine. Intuitively, two sliding windows (documents) are assigned to the same routine label if their topic proportions are similar. Using the AP algorithm to perform routine clustering, the number of routines needs not to be specified.

VI. EXPERIMENTAL RESULTS

In this section, we present qualitative and quantitative evaluation results of our nonparametric routine discovery framework. First, we show the discovered routines from the daily life routine dataset and the transportation mode dataset. Second,

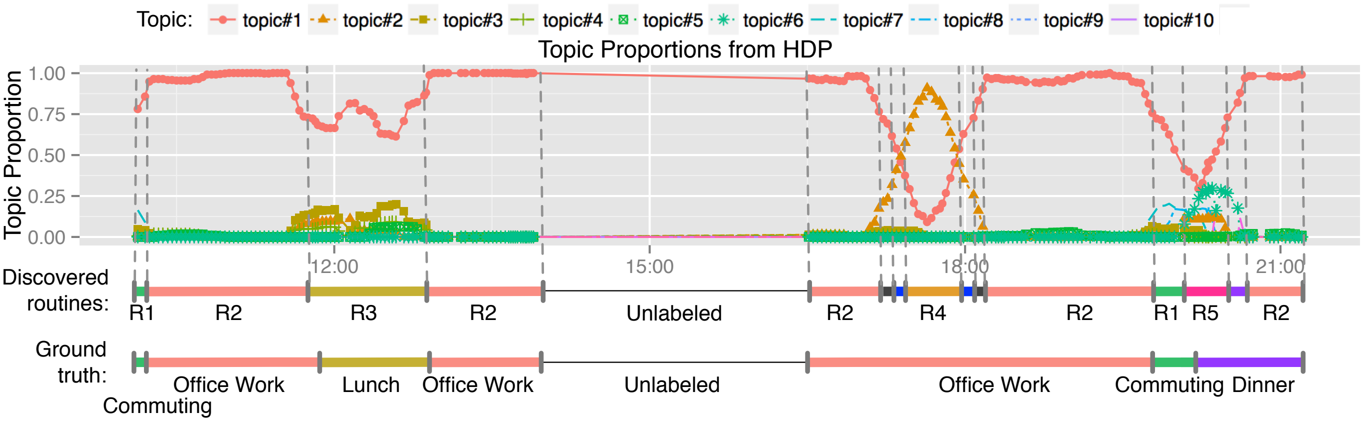


Fig. 5. The top plot visualizes topic proportions inferred from HDP during the course of a day in the daily routine data. The bottom plot shows the comparison of the ground truth routine labels and the discovered routines from our framework. Note that the inferred topics reveal high correlation with annotated routine labels. For example, "`Office Work`" and "`Lunch`" correspond to higher proportion of Topic#1 and Topic#3 respectively. Hence, topic proportions allow us represent and discover high-level daily routines.

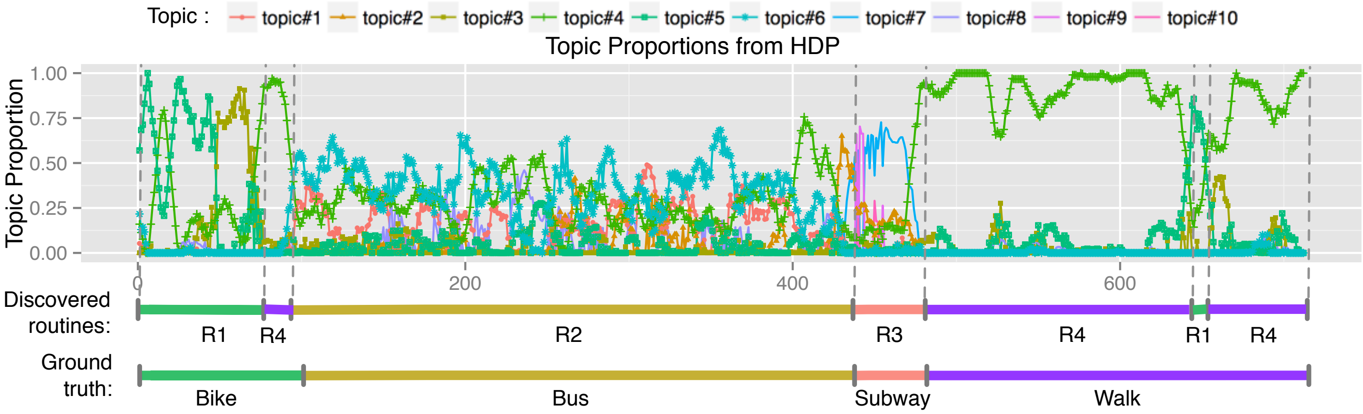


Fig. 6. The top plot illustrates topic proportions inferred from HDP for a week of the transportation mode data. The bottom plot shows the comparison of the ground truth transportation mode labels and the discovered routines from our framework. Note that the inferred topics reveal high correlation with annotated transportation mode labels. For example, "`Bike`" and "`Walk`" correspond to higher proportion of Topic#5 and Topic#4 respectively. Hence, it suggests that we can use topic proportions to represent and discover transportation mode routines.

we compare DPGMM and K-means, the baseline clustering method used in previous work [7], for vocabulary construction. Finally, the performance comparison of nonparametric (HDP) and parametric (LDA) topic models is presented.

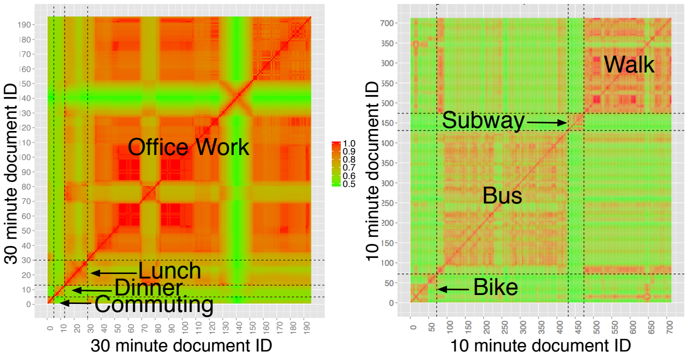
A. Qualitative analysis

We first show the output of our routine discovery method on the two datasets. Figure 5 illustrates the extracted routines for one day from the daily life routine dataset. The top plot shows how the learned topic proportion vectors change over time. The documents are constructed using a 30-minute sliding window with 2.5 minutes overlap. (Data between 14:00pm-16:30pm are unlabeled.) The middle plot shows the extracted routine classes by clustering these topic proportion vectors using AP. The bottom plot shows the ground truth. The ground truth label of a specific sliding window is assigned with the most frequent routine class label in that window. Different colors indicate different routines. We see that the discovered routines "R1" (green), "R2" (pink), and "R3" (yellow) match ground truth labels "`Commuting`", "`Office Work`", and "`Lunch`".

Our method extracted more routines than the ground truth labels. For example, 17:00-18:00 pm is labeled as "R4" while the ground truth label is still "`Office Work`". We examined the ground truth activity labels in the dataset and found that "walking freely" occurs more frequently during this period of time compared to other parts of the "`Office Work`". Thus, it is expected that our method would identify it as an additional routine.

Moreover, our method labels "R2" in the last part of the "`Dinner`" period. This is because the labeled "`Dinner`" period consists of "*cooking in the kitchen*" and "*sitting at the table to dine*". From the accelerometer sensors' point of view, "*sitting at the table to dine*" and "*sitting at the office table*" have similar word distributions. Therefore, based on the sensor data, our method is not able to distinguish them. This scenario suggests that using other types of sensor data, such as location and time, might be useful.

Similarly, Figure 6 shows the inferred transportaion modes



(a) Daily routine data similarity matrix (b) Transportation data similarity matrix

Fig. 7. Similarity matrix for (a) the daily routine dataset and (b) the transportation mode dataset of the topic proportion vectors learned from HDP. Red color refers to higher similarity and green color refers to lower similarity. Note that learned topic proportions of documents with the same ground truth labels have higher similarity.

on a week of transportation mode data. The documents are constructed using a 10 minute sliding window with 1 minute overlap. We reorder and group documents with the same ground truth labels for the better visualization. We see that the discovered routines are highly correlated to the ground truth transportation modes.

Figure 7 shows two similarity matrices of the extracted topic proportion vectors on the daily routine and the transportation mode datasets using Eq (13). The Red color refers to higher similarity, and the green color refers to lower similarity. For better visualization, we group documents based on their ground truth routine labels. In both similarity matrices, we see that topic proportion vectors of documents with the same ground truth label are similar, forming red sub-blocks. Moreover, the green bands in the "Office Work" block in Figure 7(a) correspond to "Office Work" with more "walking freely" activity occurrences.

B. Quantitative evaluation

1) Evaluation metrics: To measure the alignment between the discovered routines and the ground truth routines, we compute three widely used clustering evaluation metrics: the *cluster purity*, the *rand index* and the *pair-counting F-measure* [26]. Given clustered documents, *cluster purity* assigns the cluster to the ground truth routine label which is most frequent in the cluster, and then computes the percentage of documents whose ground-truth label is the same as its cluster label. The *Rand index* looks at all pairs of documents and caculates the percentage of document pairs that are correctly classified, $\frac{TP+TN}{TP+FP+FN+TN}$, where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives respectively. More specifically, TP is the number of similar document pairs that are assigned to the same cluster and TN is the number of dissimilar document pairs that are assigned to different clusters. *F-measure* is defined as $F_\beta = \frac{(\beta^2+1)P\times R}{\beta^2 P+R}$ where P is precision and R is recall. In this paper, we set $\beta = 1$ and *F-measure* becomes $\frac{2\cdot TP}{2\cdot TP+FP+FN}$ in which true positive are double counted. Intuitively, *F-measure* penalizes false negatives more than false positives.

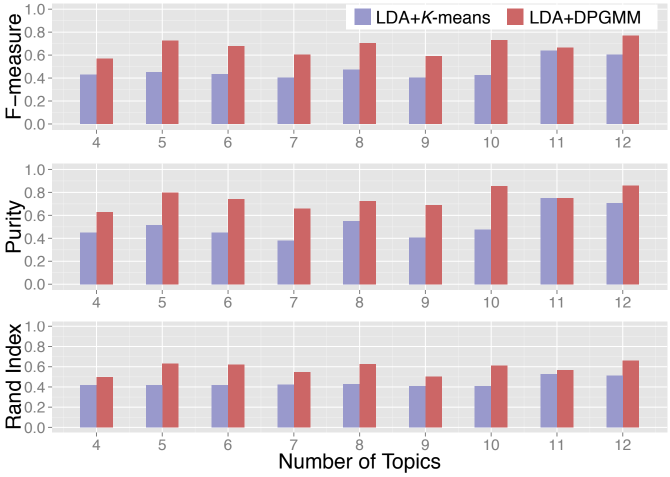


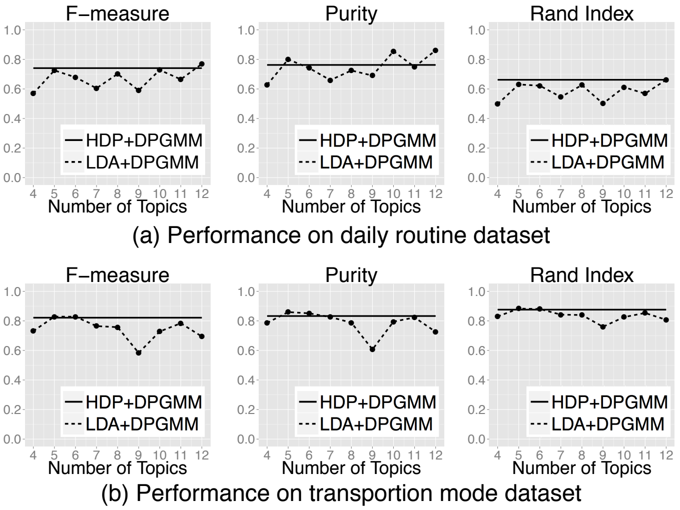
Fig. 8. Performance comparison of DPGMM (nonparametric) against K-means (parametric) for vocabulary extraction on the daily routine dataset, in terms of the F-measure (top), the purity (middle), and Rand index (bottom) over different numbers of topics in LDA.

2) DPGMM vs. K-means: We first study the effect of using nonparametric and parametric methods in the vocabulary construction phase. We consider K-means as the baseline parametric method for constructing vocabulary. Figure 8 shows results on the daily routine dataset using K-means and DPGMM. The high level topic model is LDA. For the K-means baseline, we use $K = 60$, which was used in previous routine discovery work on the same dataset [7]. We see that DPGMM performs consistently better than K-means in all three evaluation metrics as we change the number of topics. This demonstrates the advantage of nonparametric methods.

3) HDP vs. LDA: Next, we fix the vocabulary construction scheme (DPGMM) and compare the performance of HDP and LDA with various number of topics shown in Figure 9. The number of topics used in LDA varies from 4 to 12. Since the number of topics is not part of the problem formulation in HDP, they are horizontal lines in the plots. As we would expect, LDA is sensitive to the selection of the number of topics. Also, HDP and the optimal setting in LDA have comparable performance. Note that the number of topics does not correspond to the number of routines. Therefore, even if we know that we have 4 routines in advance, the optimal number of topics is not 4. The number of topics inferred in HDP on the daily routine and the transportation datasets is 12 and 10, respectively.

VII. Conclusion and Future Work

In this paper, we have presented a novel routine discovery framework which adopts nonparametric Bayesian methods. Most previous work in topic model based routine discovery used parametric methods such as K-means for low-level activity clustering and LDA for routine discovery. However, model selection (e.g., via a trial-and-error process) is the main challenge for the adoption of parametric models. Our two-phase



(a) Performance on daily routine dataset

(b) Performance on transportion mode dataset

Fig. 9. Performance comparison of HDP+DPGMM (nonparametric) against LDA+DPGMM (parametric) on the daily routine and the transportation mode dataset in terms of the F-measure (left), the purity (middle), and Rand index (right). We can see that HDP+DPGMM performs as well as the best LDA model without a model selection procedure.

nonparametric routine discovery framework uncovers latent routines from sensor data in a fully unsupervised fashion. More specifically, the framework automatically finds the size of the low-level activity vocabulary from multi-dimensional feature vectors using DPGMM at the vocabulary extraction phase. At the routine discovery phase, the framework further applies HDP to automatically select the appropriate number of latent topics and discover latent routines. The framework has been validated on two public datasets. Experimental results show that our nonparametric framework can achieve comparable performance against parametric models without the need of specifying parameters in advance.

In the future, we would like to further investigate how to incorporate multi-modal sensor data streams for the routine discovery task [27] using nonparametric Bayesian methods. We would also like to explore a more systematic and automatic mechanism to select the appropriate length of the document [28]. Furthermore, it would be interesting to explore how to transfer the knowledge learned from a single user's routine discovery model to a group of users [15].

References

[1] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data." Springer, 2004, pp. 1–17.

[2] Q. Yang, "Activity recognition: linking low-level sensors to high-level intelligence," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, 2009, pp. 20–25.

[3] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*. ACM, 2009, pp. 165–178.

[4] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing daily activities with RFID-based sensors," in *Proceedings of the 11th International Conference on Ubiquitous Computing*. New York, NY, USA: ACM, 2009, pp. 51–60.

[5] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 3:1–3:27, Jan. 2011.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[7] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," in *Proceedings of the 10th international conference on Ubiquitous computing*, New York, NY, USA, 2008, pp. 10–19.

[8] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille, "A long-term evaluation of sensing modalities for activity recognition," in *Proceedings of the 9th International Conference on Ubiquitous Computing*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 483–500.

[9] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *Comm. Mag.*, vol. 48, no. 9, pp. 140–150, Sep. 2010.

[10] H.-T. Cheng, F.-T. Sun, M. Griss, P. Davis, J. Li, and D. You, "Nuactiv: recognizing unseen new activities using semantic attribute-based learning," in *Proceeding of the 11th annual International Conference on Mobile Systems, Applications, and Services*. New York, NY, USA: ACM, 2013, pp. 361–374.

[11] B. Longstaff, S. Reddy, and D. Estrin, "Improving activity classification for health applications on mobile devices using active and semi-supervised learning," in *PervasiveHealth*, 2010, pp. 1–7.

[12] M. Mahdaviani and T. Choudhury, "Semi-supervised and active training of conditional random fields for activity recognition," in *Advances in Neural Information Proceeding Systems*, 2007, pp. 977–984.

[13] D. Minnen, T. Starner, I. Essa, and C. Isbell, "Discovering characteristic actions from on-body sensor data," *2012 16th International Symposium on Wearable Computers*, pp. 11–18, 2006.

[14] N. Eagle and A. Pentland, "Eigenbehaviors: Identifying structure in routine," in *Behavioral Ecology and Sociobiology*, 2009, pp. 1057–1066.

[15] J. Zheng and L. M. Ni, "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. New York, NY, USA: ACM, 2012, pp. 153–162.

[16] D. H. Hu, X.-X. Zhang, J. Yin, V. W. Zheng, and Q. Yang, "Abnormal activity recognition based on HDP-HMM models," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009, pp. 1715–1720.

[17] Y. Zhu, Y. Arase, X. Xie, and Q. Yang, "Bayesian nonparametric modeling of user activities," in *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*. New York, NY, USA: ACM, 2011, pp. 1–4.

[18] T. Nguyen, D. Phung, S. Gupta, and S. Venkatesh, "Extraction of latent patterns and contexts from social honest signals using hierarchical Dirichlet processes." Los Alamitos, CA, USA: IEEE Computer Society, 2013, pp. 47–55.

[19] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2009.

[20] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[21] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," vol. 9, no. 2, 2000, pp. 249–265.

[22] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in Neural Information Processing Systems*. MIT Press, 2004, p. 2003.

[23] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," in *Proceedings of the 10th International Conference on Ubiquitous Computing*. New York, NY, USA: ACM, 2008, pp. 312–321.

[24] A. Krishnamurthy, "High-dimensional clustering with sparse Gaussian mixture models," 2011.

[25] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, p. 2007, 2007.

[26] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[27] O. Yakhnenko and V. Honavar, "Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence," in *In SIAM SDM*, 2009.

[28] J. Seiter, O. Amft, and G. Tröster, "Assessing topic models: How to obtain robustness?" in *AwareCast 2012: Workshop on Recent Advances in Behavior Prediction and Pro-active Pervasive Computing*, 2012.