Joint Segmentation and Activity Discovery using Semantic and Temporal Priors

Julia Seiter*, Wei-Chen Chiu†, Mario Fritz†, Oliver Amft‡ and Gerhard Tröster*
*Wearable Computing Lab., ETH Zurich, Switzerland
{seiter,troester}@ife.ee.ethz.ch
†Max Planck Institute for Informatics, Saarbrücken, Germany
{walon,mfritz}@mpi-inf.mpg.de
‡ACTLab, University of Passau, Germany
amft@fim.uni-passau.de

***Abstract*—We introduce a hierarchical nonparametric topic modeling approach to infer activity routines from context sensor data streams based on a distance dependent Chinese restaurant process (ddCRP). Our approach does not require labeled data at any stage. Neither does our approach depend on time-invariant sliding windows to sample context word statistics. Our activity discovery approach builds on the idea that context words occurring within one activity are semantically similar, whereas context words of different activities are less similar. Context word streams are segmented into supersamples and then semantic and temporal features are obtained to construct a segmentation prior that relates supersamples via its context words. Our hierarchical model uses the segmentation prior and ddCRP to group supersamples and the Chinese restaurant process (CRP) to discover activities. We evaluate our approach using the Opportunity dataset that contains activities of daily living. Besides being nonparametric, our ddCRP based model outperforms both, classic parametric latent Dirichlet allocation (LDA) and the nonparametric Chinese restaurant franchise (CRF). We conclude that ddCRP+CRP is an adequate approach for fully unsupervised activity discovery from context sensor data.**

I. Introduction

Discovery of daily activities and routines from ubiquitous sensor data provides insights into individual behavior without prior model learning, which is relevant for assisted living, remote patient care, and related applications [20], [1]. A common approach to assess human behavior is to partition activity routines into abstract levels. For example, *office work* and *lunch* can be decomposed into context symbols, typically of shorter temporal duration. These may include activity primitives (*sit, walk*), locations (*home*), or object use (*computer, spoon*). Context symbols could be detected from the continuous acquired data of on-body and ambient sensors, where frequently supervised classification or data clustering were used [14]. Activity routine discovery requires methods for analyzing context symbol patterns, where often parametric topic models were applied, such as latent Dirichlet allocation (LDA) [14], [8]. Topic models originate from text mining and aim at discovering hidden themes from word statistics in documents. For parametric topic models it is assumed that one document contains a mixture of a finite number of topics and that each topic is described as probabilistic distribution over words from a predefined vocabulary.

In activity discovery, words correspond to context symbols and topics correspond to activities, which we call context words and activity topics respectively. Typically, documents are obtained using a temporal segmentation of the continuous context word stream with a predefined segment size that is large enough to capture context word statistics. Subsequently, discovery results per segment are retrieved. With frequently used segment sizes of 30 min [14], [22], activity transitions and activities with variations in duration may not be accurately identified. Moreover, parametric topic models, such as LDA, require to set the number of topics. Selecting topic model parameters, including segment size and number of topics, impacts activity discovery performance and highly depends on dataset properties that may be unknown [19]. Recently, Bayesian nonparametric topic models were proposed for activity discovery to overcome the dependency on a predefined topic count [17], [22]. However, existing nonparametric models also depend on a fixed segment size.

In this paper, we introduce a novel hierarchical topic model approach that does not depend on manually selecting parameters segment size and number of topics. Instead, segmentation and topic count estimation is performed based on the data and jointly with the activity topic discovery. We propose a framework that includes context word extraction and activity discovery. Context words are obtained from sensor data without statistical classifier training and thus do not require activity annotations. We introduce a segmentation prior considering semantic and temporal information and use the nonparametric distance dependent Chinese restaurant process (ddCRP) to group context words that belong to one activity. For example, segmentation of activity *lunch* would contain context words such as *spoon* and *plate*, whereas activity *office work* may contain *computer*. Thus, our semantic relationship representation of *spoon* is "closer" to *plate* than to *computer*.

The contributions in this paper are threefold: (1) We introduce a joint segmentation and activity discovery approach that is independent of the number of topics and the segment size. Here, we combine the nonparametric ddCRP and Chinese restaurant process (CRP) hierarchically and formulate a segmentation prior that considers semantic and temporal features of context words. Semantic representations were extracted from a corpus of Wikipedia articles. (2) We show that our approach outperforms the parametric LDA and the nonparametric Chinese restaurant franchise (CRF) on the Opportunity dataset that contains multi-modal sensor data [18]. (3) We demonstrate the increased robustness and performance of ddCRP+CRP

compared to other methods regarding activity discovery from context word annotations, actual context word detections from raw data, and synthetic noise.

II. Related Work

Several attempts towards activity discovery from sensor data were made. Gu et al. extracted characteristic object use fingerprints applying web-mining and discovered contrast patterns for each activity using emerging patterns [11]. Begole et al. applied data clustering to extract and visualize human's daily rhythms from computer activity [3]. As clustering-based methods cannot capture uncertainty in the structure of human activities, frequently probabilistic models have been applied. Barger et al. used probabilistic mixture models to infer daily life behavior patterns from clusters of sensor events in a smart home [2]. Probabilistic topic models have been applied to extract activity routines from mobile phone data [8], [24] and activity primitives [14]. However, all of these topic model approaches are parametric and assume a fixed model complexity. Thus, discovery performance critically depends on the number of topics specified. In contrast, our approach is nonparametric, thus estimates optimal topic count from the data structure.

Nonparametric models were recently applied for activity discovery. The hierarchical Dirichlet process HDP-HMM was used for abnormal activity detection [13] and activity discovery from smartphone sensor data [25]. Similarly, Nguyen et al. used HDP to discover latent activity topics from acceleration and proximity data [17]. Sun et al. used HDP to discover patterns of high-level activities [22] from data clusters. While nonparametric topic models estimate an optimal number of topics based on the data, their discovery performance remains sensitive to selecting proper segment sizes. The topic model-based discovery frequently used time-invariant segmentation, such as sliding windows [22], [14]. Yet, time-invariant segmentation fails to handle transitions, variations in activity duration, and short activities accurately. Our approach no longer requires selecting a segment size, but performs segmentation dynamically based on the data by introducing a segmentation prior to group context words of the same activity.

Nonparametric topic models were successfully applied to infer themes in text documents [10], [5]. In text mining, segmentation of text is not needed as text is naturally segmented in documents. However, using nonparametric topic models for object discovery in images and videos [15], [21] or activity discovery from sensor data [22] requires a proper segmentation. Recently, the distance dependent Chinese restaurant process (ddCRP) has been suggested to consider segmentation priors for nonparametric object discovery as well as joint video segmentation and inference of object appearance models in images and videos [9], [7]. Chiu and Fritz defined a video segmentation prior to group pixels of coherent motion using ddCRP and an infinite mixture model to extract global object classes based on CRP [7]. We introduce a similar approach for segmentation and nonparametric activity discovery from multi-modal sensor data. While in [7] the segmentation prior was based on spatio-temporal and motion similarities between pixel groups, we introduce semantic and temporal features to relate context words.

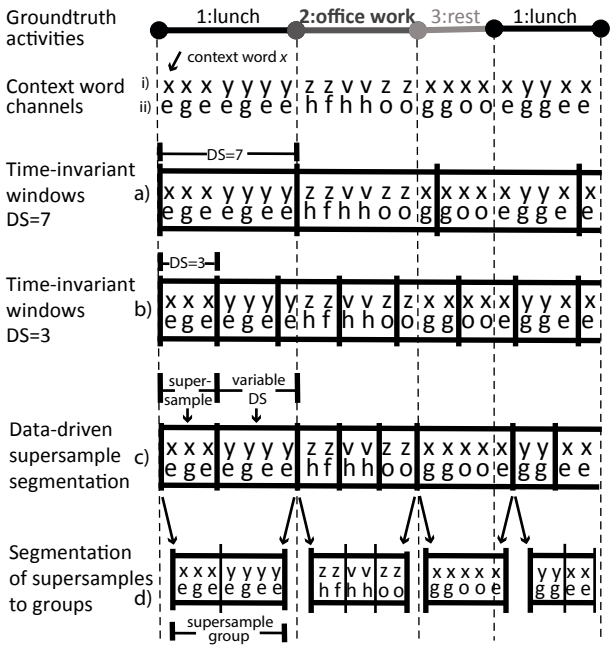

Fig. 1: Illustration of segmentation methods for activity discovery. Exemplary, 3 activities are shown and segmentations for 2 context word channels $\{i, ii\}$ with the context word vocabulary $\{e, f, g, h, v, x, y, z, o\}$. (a) Time-invariant windowing with segment size DS=7. (b) Time-invariant windowing with segment size DS=3. (c) Data-driven supersamples segmentation. A new supersample is formed each time a context change occurs in channel (i). (d) supersample groups are segmented by ddCRP with segmentation prior. Whereas in (a) and (b) windows intersect activities, (c,d) perform segmentation according to data.

III. Joint segmentation and discovery approach

Time-invariant sliding windows cannot adequately handle variations in activity duration. Figure 1 illustrates a segmentation problem of variable durations in activity discovery with examples: Large time-invariant windows, e.g. a segment size of DS=7, capture context word statistics of activity 1 exactly (see Fig. 1(a)). However, context word statistics for activity 3 would be incomplete, as the context word windows of activity 2 and 1 overlap. Contrary, a small segment size (e.g. DS=3) does not provide distinct context word statistics for activity 1*:lunch*, as illustrated in Fig. 1(b).

We introduce a joint segmentation and discovery approach as depicted in Fig. 2 to solve the segmentation problem. The first stage extracts data from multi-modal sensor sources into context words e.g., *sit, spoon moved*. The context word extraction relies on basic logic functions, thus avoiding supervised statistical learning and classification. Subsequently, we introduce a data-driven segmentation based on state changes in context words to obtain supersamples (see Fig. 1(c)). Supersamples represent short temporal segments of context words with variable size. As state changes in context words may occur within activities, supersamples may not comprehensively capture context word statistics that represent a particular activity, e.g., activity 1*:lunch* in Fig. 1(c). Therefore, supersamples will be grouped according to semantic and temporal context word relations.

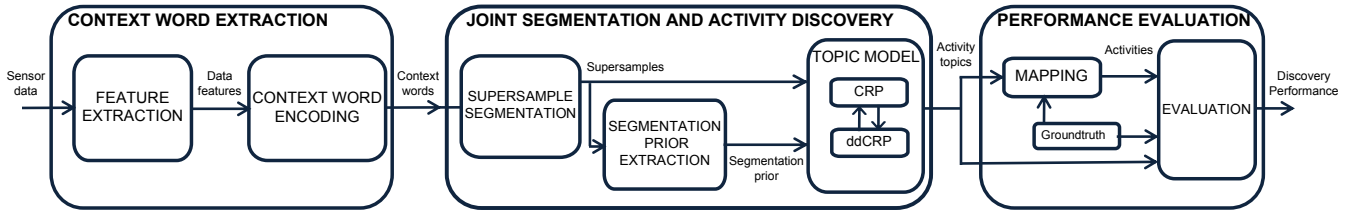

Fig. 2: Activity discovery framework: Sensor data is processed and encoded in context words using a predefined context vocabulary. Our approach jointly segments context words and performs activity discovery. Initially, a data-driven segmentation transforms context word streams into supersamples as input for a hierarchical, nonparametric topic model. We use semantic and temporal priors to group supersamples of the same activity with ddCRP. The CRP process is then used to infer activity topics from context word statistics of supersample groups. To evaluate performance, activity topics are mapped to a set of activities.

We assumed that an activity includes semantically similar context words, whereas the semantic relation of context words between different activities is lower. For example, context words *x* and *y* in Fig. 1 may correspond to *plate* and *spoon*. Then, *x:plate* and *y:spoon* are semantically more similar than *x:plate* and *z:computer*. To group supersamples that belong to the same activity (Fig. 1(d)), we introduce a segmentation prior that considers semantic and temporal relationships of supersamples based on the context words in each supersample. We deduced semantic distances between context words based on *word2vec* representations that were extracted from a corpus of Wikipedia articles [16]. For example, activity *1:lunch* contains a supersample i=1 with context words {*x,x,x,e,g,e*} and a supersample i=2 with {*y,y,y,y,e,g,e,e*} (Fig. 1(c)). The third supersample i=3 belongs to the activity *office work* and includes context words {*z,z,h,f*}. We expect higher prior probability to group supersamples 1 and 2 than supersamples 1 and 3 as the semantic and temporal distance of *x:plate* (supersample 1) and *y:spoon* (supersample 2) should be smaller than between *x:plate* (supersample 1) and *z:computer* (supersample 3). Contrary to the example here, distances for all pairs of context words were considered in the prior (see Sec. IV-D for details).

We then apply a hierarchical, nonparametric topic model for activity topic discovery using ddCRP and CRP as depicted in Fig. 3(c): ddCRP and CRP are clustering algorithms where the number of clusters is not given a priori but estimated from the data. In the local layer, supersamples are clustered into groups as illustrated in Fig. 1(d) and Fig. 3(c) using ddCRP. Grouping by ddCRP depends on the segmentation prior: In our example, supersamples i=1 and i=2 belong to activity *1:lunch* and have high prior probability to be grouped contrary to supersamples i=1 and i=3 that belong to different activities (see Fig. 1(c)). We expect supersample groups to provide comprehensive context word statistics describing activities (see Fig. 1(d)). Individual data recordings of a dataset likely contain the same activities. Thus, the global layer combines supersample groups that belong to the same activity by CRP to an activity topic group e.g. *q=1:lunch* (see Fig. 3(c)). For each activity topic group, the context word distribution is sampled from context word statistics of all assigned supersample groups such that the likelihood of the data is maximized. Retrieved activity topics were mapped to activities and discovery performance was analyzed (see Fig. 2).

IV. Discovery framework

The complete discovery framework is illustrated in Fig. 2.

A. Context Word Extraction

The context vocabulary covers X context words $\{e, f, g...\}$ that are extracted from body worn and ambient sensor data. First, features are extracted from raw sensor data (see Fig. 2). Each statistical feature from sensor data is transformed to a binary feature F using thresholds. Binary features are subsequently included in logic functions to obtain context words (see Sec. V-B, Tab. I for an example). Parallel operating context word detectors (e.g. *mode of locomotion, object usage*) result in several context word channels. Each context word channel provides either an active context word or a *null class* symbol, when no context word is active.

B. Segmenting Context Words into Supersamples

We use a data-driven segmentation for context word streams that result in variable sized segments, referred as supersamples similar to superpixels in vision [7]. New supersamples are formed each time a context state change occurs (see Fig. 1(c) for illustration). As there may be several parallel context channels from different sensors sources, we use the channel that includes the least sparse context word sequence. We consider that supersamples will typically have shorter temporal duration than activities and subsequently need to be grouped. We use a joint segmentation and activity discovery approach, as described below.

C. Chinese Restaurant Process (CRP) and Distance Dependent Chinese Restaurant Process (ddCRP)

The Chinese restaurant process (CRP) is based on a Dirichlet process $DP(\alpha, G_o)$ with base distribution G_o and concentration parameter α [5]. As metaphor it can be described by a Chinese restaurant with an infinite number of tables k and a menu of dishes ϕ as depicted in Fig. 3(a). Each table serves one dish ϕ_k. N customers enter the restaurant sequentially and randomly sit at a table. The probability that customer i is assigned to an existing table $k \in 1...K$ depends on the number of customers n_k already sitting at table k. The customer opens up a new table proportional to the parameter α:

$$p(z_i = k | z_{1:i}, \alpha) \propto \begin{cases} n_k & k \leq K \\ \alpha & k > K \end{cases}, \quad (1)$$

where z_i is the table assignment of customer i. In case a new table is opened up, the dish ϕ_k at the new table is sampled from G_o. In CRP, the table assignment is independent of previously entered customers. However, it might be more likely

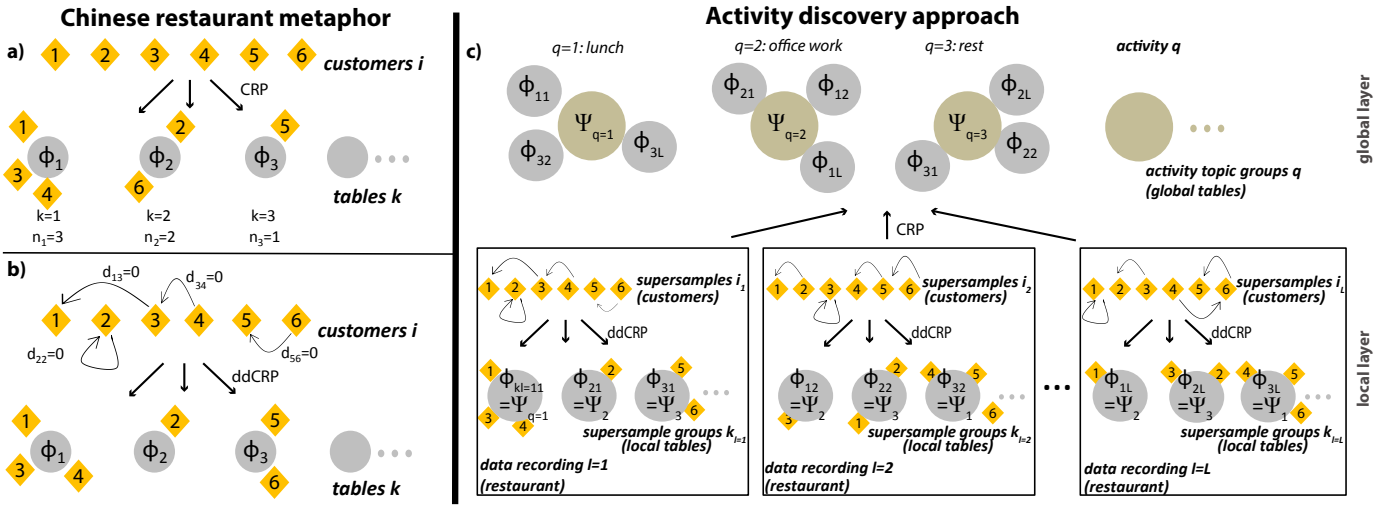

Fig. 3: Illustration of discovery processes. (a) Chinese restaurant process (CRP): customers are assigned to tables k according to Eq. (1). Customers sitting at the same table have the same dish ϕ_k. (b) Distance dependent Chinese restaurant process (ddCRP), introducing customer dependencies: customers i and j with small distances d_{ij} are likely to sit at the same table k as defined by Eq. (2). (c) hierarchical discovery framework ddCRP+CRP. Restaurants correspond to data recordings, customers to supersamples, tables are individual supersample groups, and dishes are activity topics. In the local layer, supersamples from recordings l are combined to supersample groups by ddCRP (see Section III). Each local supersample group k_l belongs to a local activity topic ϕ_{kl}. In the global layer, local supersample groups from all L recordings are assigned to global activity topic groups q using CRP and share global activity topic Ψ_q. Local activity topics ϕ_{kl} inherit the global activity topic Ψ_q. For example, supersamples that are assigned to supersample groups k_1=1, k_2=3 and k_L=3 belong to activity q=1:*lunch*.

that e.g. customers, who enter the restaurant in close temporal relation sit at the same table. Thus, ddCRP introduces customer dependencies [4], see Fig. 3(b). Customers i are linked to other customers j based on their dependency d_{ij}. Linked customers share the same table k. In ddCRP, the probability that customer i is linked to customer j is inverse proportional to their distance d_{ij}, whereas customer i sits alone proportional to α:

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f(d_{ij}) & j \neq i \\ \alpha & j = i \end{cases}, \qquad (2)$$

where c_i is the customers assignment, $f(d)$ denotes the decay function and D the set of all distances between customers. For activity discovery, restaurants correspond to data recordings, customers to supersamples, tables to supersample groups, and dishes to activity topics.

D. Segmentation Priors for Activity Discovery

In this work, the *word2vec* algorithm was used to extract vector representations of words, where the word vectors capture semantic relationships between words [16]. *Word2vec* is based on a continuous Skip-gram model that infers word vector representations unsupervised from a corpus of articles. Initially, the algorithm constructs a *word2vec* vocabulary of size W from the text corpus and then deduces vector representations based on neural networks. Finally, each word in the *word2vec* vocabulary is represented by the semantic relationship to W other words leading to a $1 \times W$ word vector for each word. We used a *word2vec* vocabulary of dimension $W = 1000$ to extract word vector representations from a corpus of Wikipedia articles (*available at https://code.google.com/p/word2vec/*). Context words represented a subset of the *word2vec* vocabulary ($X << W$) and were manually mapped to relevant word vectors v_x for X context words by searching the labels of X context words in the *word2vec* vocabulary.

We used the *word2vec*-based semantic as well as temporal distances between supersamples to form a segmentation prior over supersamples for ddCRP that likely groups supersamples belonging to the same activity (see Fig. 2). To semantically represent a context word x, we used word vector v_x. To semantically represent supersample i, we calculated the mean word vector v_i across all X_i unique context words x in supersample i: $v_i = \frac{1}{X_i} \sum_{x \in X_i}^{X_i} v_x$. The semantic distance d^s_{ij} of supersamples i and j is the Euclidean distance $d^s_{ij} = d(v_i, v_j)$ of their mean word vectors. The temporal distance d^t_{ij} counts the number of supersamples between supersample i and j. Considering our segmentation prior over supersamples, we modified Eq. (2) for supersample assignment c_i:

$$p(c_i = j|D, f, \alpha) \propto \begin{cases} f^t(d^t_{ij}) f^s(d^s_{ij}) & j \neq i \\ \alpha & j = i \end{cases}. \qquad (3)$$

The distance measure D and decay function f for ddCRP are composed of a temporal distance measure and decay function (D^t, f^t) and a semantic distance measure and decay function (D^s, f^s). The window decay function $f^t(d^t) = [d^t < A]$ assigns direct linkage probabilities for supersamples that are at most distance A apart. For the semantic distance d^s, we use an exponential decay function $f^s(d^s) = exp(-\frac{d^s}{B})$ that decreases linkage probability with increasing semantic distance. B is the width parameter.

E. Joint Segmentation and Activity Discovery (ddCRP+CRP)

Our activity discovery approach uses ddCRP in the local layer and CRP in the global layer as illustrated in Fig. 3(c).

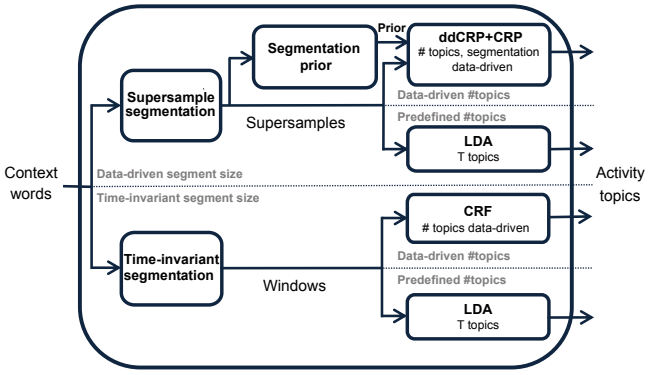

Fig. 4: Illustration of the evaluation strategy to assess and compare performance of our joint segmentation and activity discovery approach with other variants. We compare performance against LDA with time-invariant segmentation and predefined T, LDA with supersamples segmentation and predefined T, and CRF with time-invariant segmentation.

The ddCRP+CRP approach can be interpreted as follows: There is a set of L data recordings (restaurants) with a shared set of global activity topics Ψ (global dishes) across all recordings (restaurants). For each recording l, supersamples i_l and j_l with small semantic and temporal distances d_{ij_l} are likely grouped to the same supersample group k_l (local table). For example, linked supersamples in Fig. 3(c) (bottom) are assigned to the same supersample group. Each supersample group k_l of all data recordings l is assigned to one global activity topic Ψ_q (global dish) by CRP with the activity topic group q (global table) (see Fig. 3(c), top). The local activity topic ϕ_{kl} (local dish) in recording l inherits the global activity topic Ψ_q (global dish) from the activity topic group q where k_l is assigned to, e.g. Fig. 3(c) $\Psi_1 = \phi_{11} = \phi_{3L} = \phi_{32}$. Thus, multiple supersample groups k_l in multiple data recordings l can belong to the same activity topic Ψ_q.

The generative process ddCRP+CRP is described by:
(1) Each supersample i_l in recording l draws supersample assignment c_{i_l} with supersample group k_{l_i} from ddCRP(D,f,α).
(2) Each supersample group k_l in recording l draws a global activity topic group assignment q_{kl} from CRP(γ).
(3) Each global activity topic group q draws activity topic Ψ_q from G_0.
(4) For each supersample i_l in recording l, context word statistics u_{i_l} are drawn from η_q, where η_q is a multinomial distribution and $q_{k_{l_i}} = q$.
Given the observed context word statistics u_i for supersample i, the likelihood that u_i is sampled from the global activity topic q is $p(u_i|\Psi_q) = \eta_q(u_i)$. We used Gibbs sampling to infer the probabilities $p(u_i|\Psi_q)$ and thus the most likely activity topic assignment q for each supersample i as detailed in [7].

V. Evaluation Methodology

The evaluation strategy is illustrated in Figure 4. We compared performance of our nonparametric ddCRP+CRP approach with data-driven supersamples segmentation and joint segmentation and activity discovery to the parametric LDA-based topic model with time-invariant segmentation (segment size DS) and predefined activity topic count T. We further compared ddCRP+CRP to LDA with supersamples segmentation and predefined T and to the nonparametric model CRF with data-derived T, but time-invariant segmentation DS. CRF [23] is a hierarchical method as well. However, CRF uses CRP in the local layer instead of ddCRP and thus does not consider segmentation priors to segment supersample groups. All evaluations were performed per study participant and present average results across all participants and 10 topic model runs. For LDA, we varied DS within $[1, 5]$ min with empirically optimal $T = 10$ as well as varied T within $[5, 20]$ at $DS = 2.5$, as suggested in [19]. We evaluated discovery performance using context word annotations that can be seen as perfect context word detectors, as well as from encoded sensor data using the context vocabulary. We further investigated sensitivity to context word detector noise by adding equally distributed noise to context word annotations.

TABLE I: Summary of context vocabulary (25 context words). Features μ, σ^2 of acceleration signal $acc_{x,y,z}$ and binary signals b of switches were transformed to binary features F using thresholds. Logic equations were applied to obtain context words from binary features F of leg SL, back SB, and object sensors SO_i.

Context Vocabulary	Logic Equations
Mode of Locomotion	
(1)walk, (2)lie, (3)sit, (4)stand	(1): F^1_{SL}, (2): $F^2_{SB} \wedge \overline{F^1_{SB}} \wedge \overline{F^1_{SL}}$, (3): $F^3_{SL} \wedge \overline{F^1}_{SL} \wedge \overline{F^2}_{SB}$, (4): $\overline{F^3}_{SL} \wedge \overline{F^1}_{SL} \wedge \overline{F^2}_{SB}$; $F^1 = 1 : \sigma^2(\|acc_{xyz}\|) \geq 10000$ $F^2 = 1 : \mu(\|acc_{yz}\|) \geq \mu(\|acc_x\|)$ $F^3 = 1 : \mu(\|acc_z\|) \geq \mu(\|acc_{xy}\|)$
Object Usage	
(4+i)motion O_i, $i = 1...15$	(4+i): $F^1_{SO_i}$
(25)null class	(25): $\overline{F^1}_{SO_i}$
(15+i)motion O_i, $i = 16...20$	(15+i): $F^4_{SO_i}$; $F^4 = 1 : b = 1$
(25)null class	(25): $\overline{F^4}_{SO_i}$

A. Dataset

To evaluate our approach, we used the Opportunity dataset that consists of ~ 30 hours of activities of daily living (ADL) recorded at 30 Hz, including annotations for 5 recordings from 4 participants [18]. ADLs included *relaxing, coffee time, early morning, cleanup, sandwich time* plus a high-level *null class*, in total 120 instances. The dataset further provides annotations for mode of locomotion (4 labels) and object usage (20 labels), plus a low-level *null class*. We considered ADLs as activities, mode of locomotion and object usage corresponded to context words resulting in 25 individual words. To infer context words from sensor data, we used the 3-axis acceleration signals $acc_{x,y,z}$ of the right upper leg sensor SL and the back-worn sensor SB. We included 3-axis acceleration sensor data of sensors SO_i ($i = 1...15$) attached to 15 objects: *salami, bread, sugar, bottle, milk, spoon, knife cheese, glass, cheese, door1, door2, plate, cup, knife salami, lazychair*. We used binary signals b of reed switches SO_i attached to 5 objects ($i = 16...20$) including *fridge, top drawer, middle drawer, lower drawer, dishwasher*.

B. Framework Implementation

We extracted a context vocabulary with $X = 25$ context words from the sensor data as detailed in Table I. Context

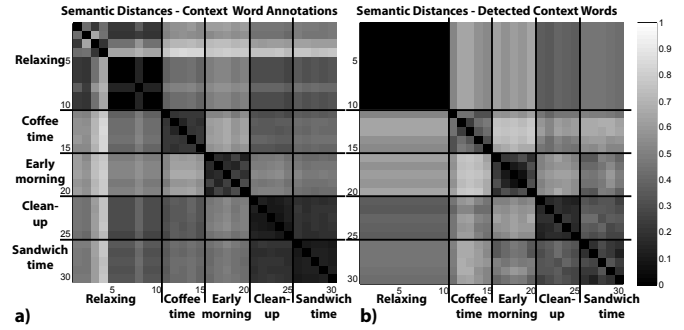

Fig. 5: Illustration of semantic distances D^s between activity instances of 5 activities in the Opportunity dataset for (a) context word labels and (b) context word detections from sensor data. Semantic distances were calculated from context word vector representations and averaged across all 4 subjects. The graph indicates that context words used within the same activity have close semantic relation.

words included mode of locomotion and object use (O_i) resulting in 21 parallel context word channels (20 object channels, 1 channel for mode of locomotion). Supersamples segmentation from the context word stream was performed using mode of locomotion as context state information, which is the least sparse context word channel of the Opportunity dataset. We used all 20 context word channels with object information to calculate the semantic distance d^s_{ij} between supersamples i and j. For ddCRP+CRP, we used the implementation of [7] with width parameters $B = 0.1$ for f^s and $A = 3$ for f^t, and hyperparameters $\alpha = 50$, $\gamma = 1$ and $\eta = 1$. For CRF, we used hyperparameters $\alpha = 1$, $\gamma = 1$ and $\eta = 1$. For LDA we used the implementation of [6] with $\alpha = 1$ and T topics. For time-invariant segmentation, we used sliding windows of size DS and segment step $0.1 * DS$ and applied Borda Count ranking to overlapping segments [12].

C. Performance Estimation

To assess activity discovery performance we mapped discovered activity topics to activities by assigning the most frequent activity per predicted activity topic using the groundtruth. *Null class* data was included for topic discovery, but removed in the performance analysis. As performance measure we used class-normalized accuracy across all 5 activities and the Rand index RI.

VI. Results

A. Semantic Relationships within and between Activities

Figure 5 shows that semantic distances D^s of context word vectors were small among instances of the same activity, e.g. *early morning*. Confirming our approach to use semantic priors, independent activities showed high distances, e.g. *early morning* and *coffee time*. Detector errors may have decreased within-activity similarity of detected context words compared to context labels, e.g. *coffee time*. Nevertheless, we also observed a reverse trend, where context word annotations appeared to be imperfect and incomplete compared to detections, e.g. for *relaxing*, *clean up* and *sandwich time*.

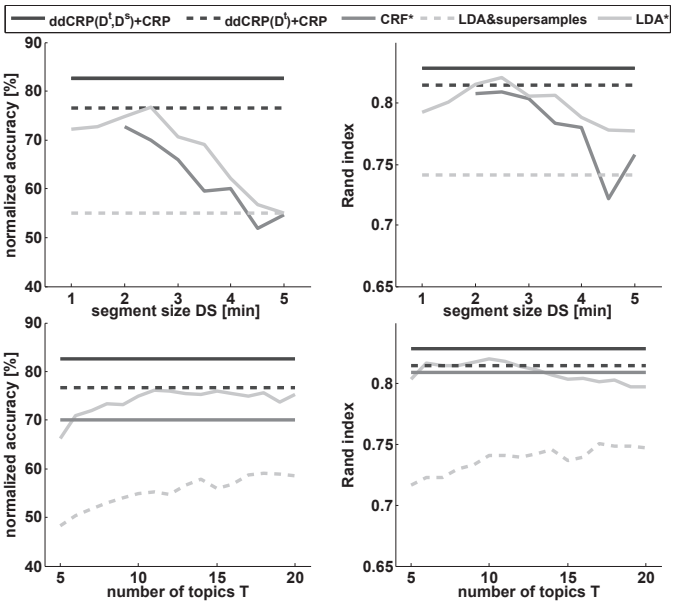

Fig. 6: Averaged normalized accuracy and Rand index for discovering activities from context word labels. Results are shown for ddCRP+CRP with temporal (D^t) and semantic (D^s) segmentation priors, CRF, and LDA. ddCRP+CRP outperformed nonparametric CRF and parametric LDA. (*) We varied segmentation window and number of topics for CRF and LDA-based methods when parameter dependency was present.

B. Activity Discovery from Context Word Labels

1) ddCRP+CRP versus LDA: Our ddCRP+CRP approach yielded 83% accuracy and Rand index $RI = 0.83$, clearly outperforming LDA as depicted in Fig. 6. LDA using time-invariant segmentation showed a peak in accuracy and Rand index for $DS = 2.5$ min and $T = 10$ topics.

2) ddCRP+CRP versus CRF: ddCRP+CRP outperformed nonparametric CRF with optimal segment size by 10% in accuracy and by $RI = 0.02$. With decreasing segment sizes accuracy of CRF increased up to 73%. The Rand index showed a peak at $DS = 2.5$ min with $RI = 0.81$.

3) Temporal and Semantic Priors: We assessed the benefit of temporal and semantic priors. ddCRP+CRP with semantic prior increased accuracy by 6% compared to ddCRP+CRP with only temporal prior. The performance of ddCRP+CRP with temporal prior was close to the performance of LDA and CRF with optimal parameters.

C. Sensitivity to Context Word Noise

Figure 7 shows that ddCRP+CRP was robust against deletion noise with up to 60% deletions and 20% insertion noise, outperforming LDA at optimal parameter settings. In practice, uniformly distributed noise across context word detectors is unlikely to occur. Besides deletions and insertions also timing and substitution errors may hamper discovery. The noise analysis may thus rather illustrate boundaries of our ddCRP+CRP approach: ddCRP+CRP performance depends on the segmentation prior. For uniformly distributed insertion noise, ddCRP likely grouped supersamples of different activities in the local layer leading to less distinct context word statistics of

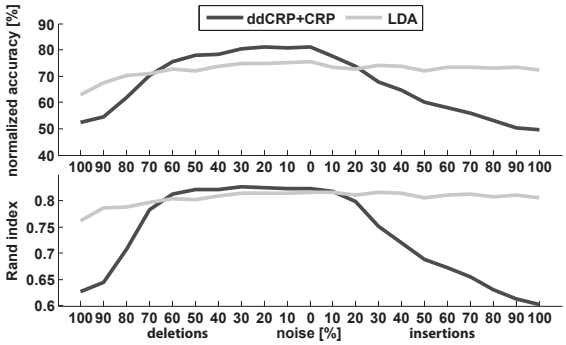

Fig. 7: Influence of evenly distributed noise over context word detectors on discovery performance. ddCRP+CRP was robust against context word deletions up to 60%, but showed sensitivity to insertions. Our approach outperformed LDA between 60% deletion and 20% insertion noise.

supersample groups at the global CRP layer. Contrary, ddCRP+CRP was less affected by deletions, as they only reduced priors grouping supersamples of the same activity. In contrast, LDA estimates activity topics exclusively from context word statistics in time-invariant segments. Evenly distributed noise offsets all context word statistics and therefore barely changes the context word structure in a segment. The sensitivity of ddCRP+CRP for insertions and robustness against deletions suggests tuning context detectors for high precision.

D. Activity Discovery from Sensor Data

For activity discovery from detected context words using our context word extraction approach, all methods showed decreased performance compared to discovery from annotations. Figure 8 shows that our ddCRP+CRP model outperformed LDA with time-invariant segmentation and optimal parameters ($T = 7$, $DS = 2.5$) by 4.5% accuracy and $\Delta RI = 0.1$. For LDA, optimal activity topic count decreased for detected context words, compared to the discovery from annotations ($T = 7$ vs. $T = 10$). Moreover, optimal segment size changed ($DS = 3.5$ vs. $DS = 2.5$). Our ddCRP+CRP model automatically selected a smaller number of activity topics for activity discovery from detected context words compared to context word labels ($T = 7$ vs. $T = 15$). ddCRP+CRP with just temporal segmentation prior performed with 60% accuracy worse than ddCRP(D^t, D^s)+CRP (78%). CRF with optimal segment size $DS = 2$ min yielded the same accuracy and slightly smaller Rand index $\Delta RI = -0.1$ as ddCRP+CRP, but more activity topics $\Delta T = 5$.

VII. Discussion

By introducing a framework for joint segmentation and activity discovery in this work, the time-invariant segmentation and parameters used in previous works towards unsupervised activity discovery were removed. Our ddCRP+CRP approach performed supersamples segmentation and activity discovery simultaneously and outperformed the parametric LDA as well as nonparametric CRF, both using time-invariant segmentation. In future work, the performance of motif search or other data-driven segmentation approaches for discretized data separately or in combination with LDA, ddCRP+CRP or other discovery methods could be compared to our ddCRP+CRP approach.

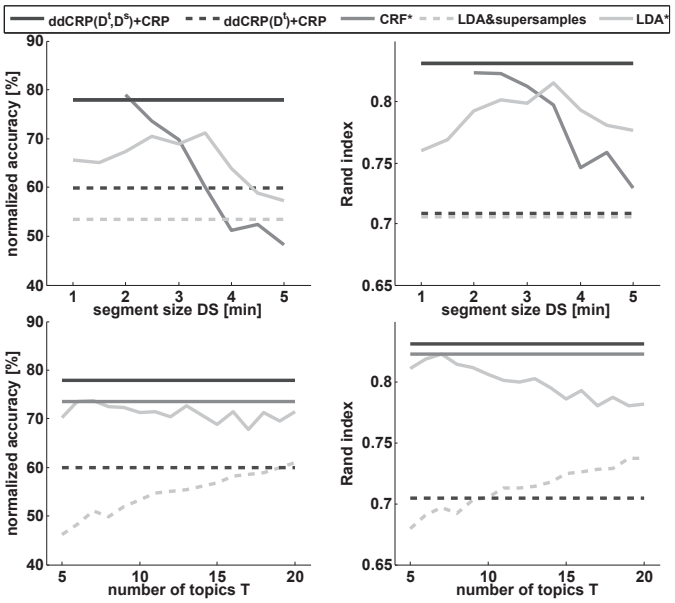

Fig. 8: Performance of activity discovery from context word detections for ddCRP+CRP including temporal D^t and semantic D^s segmentation priors, CRF and LDA. Our nonparametric ddCRP+CRP approach outperformed parametric LDA and nonparametric CRF at their optimal parameter settings. (*) We varied segmentation window and number of topics for CRF and LDA-based methods when parameter dependency was present.

We used basic logic functions to extract context words, thus avoiding statistical classifier learning. Our data-driven context word segmentation generated supersamples that were typically shorter than activities. Hence, individual supersamples did not capture distinct context word statistics to describe activities and may explain the poor performance of LDA using supersamples segmentation. Using a temporal segmentation prior for ddCRP+CRP increased accuracy over the LDA-based approach, but performed less accurate compared to ddCRP+CRP using a temporal and semantic segmentation prior. For discovery from context labels, ddCRP+CRP outperformed CRF by 10% in accuracy. For discovery from detected context words, both methods showed similar peak accuracy. However, CRF yielded 5 additional activity topics compared to ddCRP+CRP (CRF: $T = 12$ at $DS = 2$, ddCRP: $T = 7$). For an intuitive mapping, T in the range of M was desirable. Thus, results were shown for $T < 20$ activity topics to describe the $M = 5$ activities. CRF obtained $T > 20$ for $DS < 2$ min.

While the nonparametric model ddCRP+CRP and CRF infer optimal activity topic count T, hyperparameters α, γ determine the expectation over T. There are no established strategies to select α, γ. However, if a range for $\hat{T} \approx T \pm 5$ is estimated, ddCRP+CRP and CRF can automatically choose an optimal T. In our work, we used the same hyperparameter setting for discovery using labels and detected context words. In our tests, ddCRP(D^t, D^s)+CRP showed similar discovery performance even when hyperparameters were varied, indicating robustness of the method. Omitting semantic priors, i.e. ddCRP(D^t)+CRP showed lower robustness to hyperparameter variation that may explain the performance difference between labels and detected context words.

We segmented context words into supersamples and used context state changes to determine supersample bounds. In this work, we only considered context changes in one selected context word channel. State changes could similarly be estimated by combining several context word channels to create a virtual context state. Selecting and constructing a segmentation source still requires expert knowledge about the targeted discovery objective and context word processing. Similarly, constructing logic functions to derive context words requires knowledge about the sensor modalities and discovery goals. Nevertheless, we consider that such logic functions could be cataloged according to sensor type and scenario, thus become reusable for similar discovery applications without parametric adjustments.

Due to the limited availability of datasets exhibiting a hierarchical annotation structure we evaluated our method with the Opportunity dataset. In future work, other datasets could be analyzed to verify scalability of the method. Our approach required context words with semantic meaning, as we used *word2vec* to formulate a segmentation prior. Nevertheless, *word2vec* is flexible and could be applied to a different corpus, e.g. containing data clusters or other symbols extracted from sensor data. In our approach context words corresponded to a small subset of the *word2vec* word vocabulary and we manually extracted context word vectors from the word vocabulary. Instead, string matching could be applied in future to automate the mapping. In our evaluation, all context words could be mapped to a word vocabulary of $W = 1000$. It is nevertheless simple to increase the vocabulary, if corresponding words could not be found or to search synonyms using language processing algorithms. We used a generic text corpus from Wikipedia to extract the *word2vec* vocabulary. Domain specific text corpora might yield even more relevant word coverage and context word vectors.

VIII. Conclusion and future work

We introduced a novel non-parametric topic model approach for joint segmentation and activity discovery from sensor data that is independent from topic model parameters, such as segment size and number of topics. We segmented context words into supersamples using context state and formulated a segmentation prior with semantic and temporal information to group supersamples that belong to individual activities using ddCRP and CRP. With this method, segmentation is adjusted to the underlying data. Evaluation results show that our approach can outperform classical non-parametric LDA and non-parametric CRF even at optimal parameter settings. We concluded that combining segmentation and non-parametric activity discovery by using a segmentation prior and ddCRP+CRP is an adequate technique for activity discovery. In future work, we like to adapt the segmentation prior to datasets with different sensor modalities and discovery objectives.

Acknowledgment

We thank Zeynep Akata for providing the *word2vec* vector representations extracted from Wikipedia articles. This work was supported by the EU Marie Curie Network iCareNet under grant number 264738.

References

[1] A. Aztiria, J. C. Augusto, R. Basagoiti, A. Izaguirre, and D. J. Cook, "Discovering frequent user–environment interactions in intelligent environments," *PUC*, vol. 16, no. 1, pp. 91–103, 2012.

[2] T. S. Barger, D. E. Brown, and M. Alwan, "Health-status monitoring through analysis of behavioral patterns," *IEEE SMC*, vol. 35, no. 1, pp. 22–27, 2005.

[3] J. Begole, J. Tang, and R. Hill, "Rhythm modeling, visualizations and applications," in *UIST*. ACM, 2003, pp. 11–20.

[4] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes," *JMLR*, vol. 12, pp. 2461–2488, 2011.

[5] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *JACM*, vol. 57, no. 2, p. 7, 2010.

[6] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[7] W.-C. Chiu and M. Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *CVPR*. IEEE, 2013, pp. 321–328.

[8] K. Farrahi and D. Gatica-Perez, "Discovering routines from large-scale human locations using probabilistic topic models," *TIST*, vol. 2, p. 3, 2011.

[9] S. Ghosh, A. B. Ungureanu, E. B. Sudderth, and D. M. Blei, "Spatial distance dependent chinese restaurant processes for image segmentation," in *Adv Neural Inf Process Syst*, 2011, pp. 1476–1484.

[10] D. Griffiths and M. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," *Adv Neural Inf Process Syst*, vol. 16, p. 17, 2004.

[11] T. Gu, S. Chen, X. Tao, and J. Lu, "An unsupervised approach to activity recognition and segmentation based on object-use fingerprints," *DKE*, vol. 69, no. 6, pp. 533–544, 2010.

[12] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *TPAMI*, vol. 16, no. 1, pp. 66–75, 1994.

[13] D. H. Hu, X.-X. Zhang, J. Yin, V. W. Zheng, and Q. Yang, "Abnormal activity recognition based on hdp-hmm models." in *IJCAI*, 2009, pp. 1715–1720.

[14] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," in *UBICOMP*. ACM, 2008, pp. 10–19.

[15] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes," in *CVPR*. IEEE, 2010, pp. 1951–1958.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Adv Neural Inf Process Syst*, 2013, pp. 3111–3119.

[17] T. Nguyen, "Bayesian nonparametric extraction of hidden contexts from pervasive honest signals," in *PERCOM Workshops*, 2014, pp. 168–170.

[18] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *INSS*, 2010, pp. 233–240.

[19] J. Seiter, O. Amft, M. Rossi, and G. Tröster, "Discovery of activity composites using topic models: An analysis of unsupervised methods," *PMC*, vol. 15, no. 0, pp. 215 – 227, 2014.

[20] J. Seiter, A. Derungs, C. Schuster-Amft, O. Amft, and G. Tröster, "Activity routine discovery in stroke rehabilitation patients without data annotation," in *REHAB*, 2014.

[21] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *INT J COMPUT VISION*, vol. 77, pp. 291–330, 2008.

[22] F.-T. Sun, Y.-T. Yeh, H.-T. Cheng, C. Kuo, and M. Griss, "Nonparametric discovery of human routines from sensor data," in *PERCOM*. IEEE, 2014, pp. 11–19.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *JASA*, vol. 101, no. 476, 2006.

[24] J. Zheng and L. M. Ni, "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data," in *Ubicomp*, 2012, pp. 153–162.

[25] Y. Zhu, Y. Arase, X. Xie, and Q. Yang, "Bayesian nonparametric modeling of user activities," in *TDMA*. ACM, 2011, pp. 1–4.