Assessing Topic Models: How to Obtain Robustness?

Julia Seiter [1], Oliver Amft [1,2], and Gerhard Tröster [1]

[1] Wearable Computing Lab., ETH Zurich, Switzerland
[2] ACTLab, Signal Processing Systems, TU Eindhoven
{seiter,amft,troester}@ife.ee.ethz.ch
http://www.ife.ee.ethz.ch

Abstract. In this work we investigate the influence of varying daily activity dataset characteristics on topic model performance stability for daily routine discovery. For this purpose, we denote a set of key dataset properties that influence the experimental design regarding recording, as well as data pre-processing steps.
Using generated daily activity datasets, we identified optimal topic model stability for particular dataset properties. Results indicated that topic model routine duration should exceed document size by a factor of more than two. Recording durations of more than 9 days were required for a set of four routines and activity primitive overlap may not exceed 5%.

Keywords: activity recognition, topic models, behaviour inference

1 Introduction

The discovery of complex daily routines from sensor data is relevant for a variety of applications stretching from medical diagnosis to independent living. Wearable sensors can provide information on the structure and routines in daily life, including complex routines such as *hygiene*, *lunch* and *dinner*. As daily life activities are very subject dependent and vary regarding duration and individual activities involved, discovering structures in daily activities is a challenging research problem. A commonly considered concept is to partition daily activity into abstraction levels, where regular *daily routine* structures can be composed of *activity primitive* sets. The latter typically has finer temporal granularity and - at the lowest level - must be suitable for recognition from sensors. Figure 1 illustrates this concept, where daily routine structures form a composition of different activity primitives.

Several approaches exist towards complex activity recognition and discovery that could describe daily routines. For instance, Huynh et al. [5] used probabilistic topic models to reveal specific activity patterns from a number of primitives, which were mapped to complex daily routines, such as *office work* or *commuting*. As primitives, Huynh et al. used activities, including *queuing in a line* and *sitting/desk activities*. Using topic models seems a very promising approach to discover structures in daily activities. Depending on the application, however,

routines vary highly in e.g. duration and primitive composition. It is yet not clear, under which dataset conditions topic models can perform robustly. To design future experimental evaluations and topic model system designs, it is thus critical to identify key properties, including training dataset duration, primitive specification, etc. that could influence performance.

In this work, we investigated the topic model stability by varying selected dataset properties. As an exhaustive evaluation of potentially influential properties is beyond feasibility, we focused on a set of key elements that influence the dataset recordings, including duration of routines, amount of training data, and specificity of routines. We considered that these properties profoundly influence data needs and number and granularity of primitives for obtaining robust discovery results. In order to evaluate dataset properties, we implemented a daily activity simulation model. The simulation model allowed us to generate datasets with different characteristics. We based our investigation on the UbiComp'08 dataset presented in Huynh et al. [5].

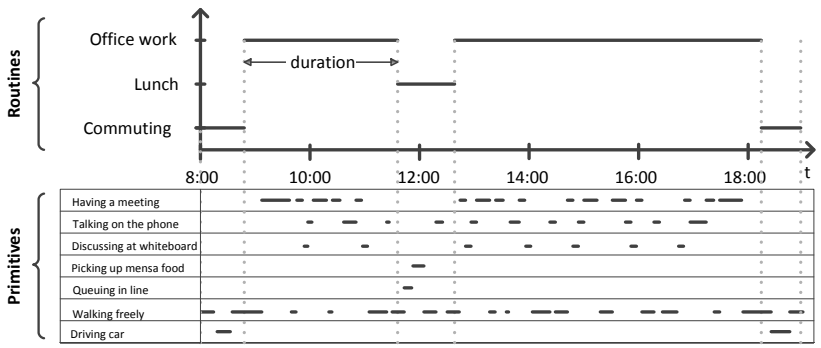

Fig. 1. Illustration of occurrence and duration of three daily routines and their composition out of primitives derived from the UbiComp'08 dataset [5]. The example visualises the common assumption to abstract daily activities. Our approach incorporates this concept for generating datasets with different properties and evaluating topic model performance stability.

In order to validate our daily activity simulation model and the topic model implementation, we used the UbiComp'08 dataset to (1) compare the performance reported by Huynh et al. in [5] to ours, and (2) confirm that the dataset creation can replicate the results in [5]. Subsequently, we investigated the topic model performance stability using a framework of simulation-based dataset generation and stability measurement. As common stability criterion we used the standard deviation of the routine prediction accuracy across multiple generated datasets. From the simulation results, requirements for a system regarding sensor modalities and data pre-processing could be derived.

The paper is structured as follows: first, we review related initiatives in complex activity recognition and topic models in Section 2. We then describe the daily activity simulation model and its formalities in Section 3, followed by fun-

damentals of the topic model framework for daily routine discovery in Section 4. We formally introduce the dataset properties considered for daily activity simulations in Section 5 and describe the analysis implementation. Sections 6 and 7 present results and the conclusion of this investigation.

2 Related Work

Hierarchical activity recognition. For complex activity recognition hierarchical models have been frequently used. Olivier at al. [8] recognized office activities using multiple HMM layers. In their work, video, audio and computer work was processed, and activities at different granularity levels were recognized. Complex activities, such as *giving a presentation* were inferred at the top layer. Lee at al. [6] presented a framework to infer activities from a variety of contextual data in the mobile setting using hierarchical Bayesian networks. Amft et al. [1] inferred composite activities from wearable and environmental sensors in a two layered model. Huynh et al. [5] classified low-level activity data such as *walking freely* or *standing* using Naive Bayes. In a second layer, they derived activity patterns from a probabilistic topic model. These activity patterns were matched to *daily routines*, such as *office work* and *lunch*. They achieved an averaged recall of 86.1% and precision of 67.2% on a set of four specific routines.

Topic models. Besides the approach of Huynh et al. [5] who focused on activity recognition to discover routines, e.g. Farrahi et al. [4] applied topic models in an unsupervised manner. They introduced a framework to discover daily routines from location and proximity data using topic models.

Topic models have been successfully applied for activity recognition in video frames. For example, in [7] and [9] human action categories were recognized from complex video streams using topic models. While topic models are common in many application fields besides text processing, we have not found investigations on topic models and its input demands, focusing on model stability. For daily routine discovery, robustly performing topic models could be applied in various applications to reveal the daily activity structure.

3 Simulation of Daily Routines

In this section, we introduce the daily activity simulation model, used to generate daily activity datasets. With the simulated daily activities we subsequently investigated the topic model stability.

For the daily activity simulation in this work, we assumed routines to be composed of several activity primitives of finer temporal granularity. For instance, the routine *office work* would consists of several primitives, such as *making a phone call*, *walking freely* or *having a meeting* (see Fig. 1). Due to this hierarchical structure in daily activities, we defined a three layered simulation model for sampling daily routines and its primitives. The top layer defines the sequence of routines, the intermediate layer describes their duration. The primitives of each

routine are derived from a lower layer model. An illustration of our simulation model is provided in Figure 2.

Top layer. The top layer consists of an HMM, which describes the set of consecutive routines during a considered number of days. Routines are chosen from k routines $r^i; i \in 1, 2, ..., k$. An HMM with n states x_j and k observations $e_j = r^i$ is used to model the routine sequence. In daily life, the same routine r^i may occur several times during a day with varying durations. Thus, a routine may be represented by several HMM states ($n \geq k$) in this model layer. The HMM is described by an $n \times n$ state transition matrix T and an $n \times n$ emission matrix E. When sampling data, the outputs of the top layer are a z-dimensional state vector $\boldsymbol{x}$ and an emission vector $\boldsymbol{e}$ containing the sequence of z states and z assigned routines.

The sequence of routines is very specific during the day. High fluctuations between different routines may thus not represent a realistic sequence. As an HMM is based on a probabilistic process, we observed that the HMM typically shows more frequent alternations between routines even when trained with a realistic set of routine sequences. To obtain realistic routine representations, we only used the HMM to sample sequences containing state transitions in different states ($x_j \rightarrow x_h; j \neq h$).

Intermediate layer. The duration d_j of each HMM state x_j is estimated from its corresponding normal distribution $N_j(\mu_j, \sigma_j^2)$, $j \in 1, 2, ..., n$. The sampling output of the intermediate layer contains a duration vector $\boldsymbol{d}$ describing the duration of each routine in a sequence denoted by the emission vector $\boldsymbol{e}$.

Lower layer. In this work, the occurrence of primitives for a particular routine is of interest. The lower layer consists of k independent Markov chains MC_{r^i} describing the sequence of primitives for each of the k routines r^i individually. Each state in the MCs represents one of the m activity primitives. When sampling primitives, for each of the routines r^i in $\boldsymbol{e}$, the corresponding Markov chain MC_{r^i} is selected. The number of primitives sampled from the Markov chains is denoted by the duration vector $\boldsymbol{d}$. As output, the primitive label vector $\boldsymbol{p}$ is formed by the ordered output of all the z Markov chain calls belonging to the sequence of z routines in $\boldsymbol{e}$. Additionally, a routine label vector $\boldsymbol{l}$ is emitted. We use a $m \times m$ primitive transition matrix T^i, $i \in 1, 2, ..., k$ for each Markov chain MC_{r^i}.

4 Topic Modeling Approach of Daily Routines

For the daily routine discovery, we used a similar topic model framework as reported by Huynh et al. [5]. Topic models find their origin in the text processing community and are used to discover k_T hidden topics in a corpus of documents filled with words from an alphabet. In this work we applied the Latent Dirichlet Allocation (LDA) algorithm, which assumes distributions of topics over documents to be derived from a Dirichlet distribution. When applying LDA on a

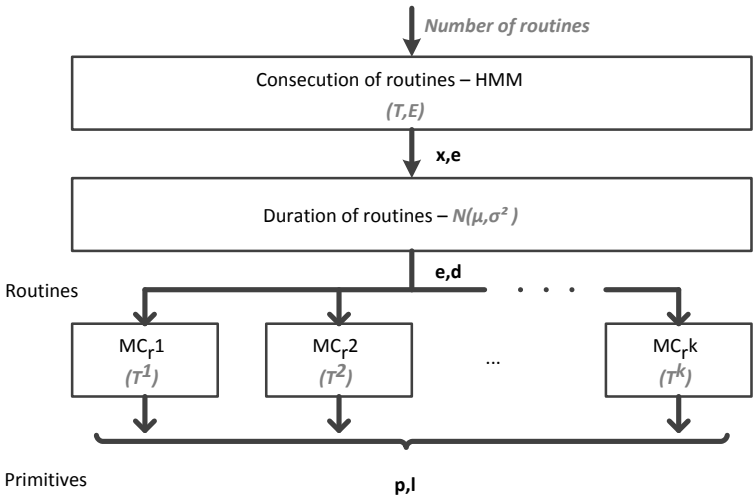

Fig. 2. Daily activity simulation model for sampling of daily routines and its activity primitives. The top layer estimates a sequence vector of routines $\boldsymbol{e}$ and states $\boldsymbol{x}$. In the intermediate layer, the duration vector $\boldsymbol{d}$ is derived from normal distributions $N_j(\mu_j, \sigma_j^2)$ assigned to the HMM states in $\boldsymbol{x}$. The lower layer applies a Markov chain MC_{r^i} for each routine r^i to sample primitives. The number of primitives is defined by $\boldsymbol{d}$. The outputs of all MCs form the primitive vector $\boldsymbol{p}$ and the routine vector $\boldsymbol{l}$. Model parameters are denoted in grey.

corpus of documents, the algorithm infers for each document d a k_T-dimensional topic activation vector gamma $\boldsymbol{\gamma_d}$ from the bag-of-words in d. The normalized vector $\boldsymbol{\gamma_d}$ describes the estimated occurrence ratio of each topic in d. More detailed information on LDA can be found in [3].

Here, we used the topic model to infer routine patterns from primitives. Routines correspond to topics, words are formed by primitives. Documents cover a time slice of a day, containing all the activity primitives in that time slice. Structuring a day into subsequent time slices is equal to structuring it in subsequent documents. The inputs for the topic model are the primitive histograms of the documents. The topic model then reveals patterns in the primitives and infers a topic activation vector $\boldsymbol{\gamma_d}$ for every document. The number of topics does not necessarily match the number of routines. Therefore, we use a superordinated kNN classifier for mapping topics to routines. The topic activation vector $\boldsymbol{\gamma_d}$ is used as feature vector for the kNN. Both the topic model and the kNN are trained by a subset of the considered daily activity data.

5 Analysis Methodology

The topic model performance stability could be affected by various dataset properties. We describe in this section the properties considered in this work and our

overall evaluation strategy. In the evaluation, we firstly validated the daily activity simulation model against the UbiComp'08 dataset presented in Huynh et al. [5]. Subsequently, we used the simulation model to generate datasets with explicit properties and analysed the topic model stability. The UbiComp'08 dataset was considered as basis of the evaluation. Our simulation approach could be generalised to other datasets that include a two-layer data hierarchy of primitives and routines by inferring the simulation model parameters as shown below.

5.1 Dataset Properties Considered in the Daily Activity Simulation

In order to determine requirements for a stable topic model performance, we considered the following dataset properties in our daily activity simulation: duration of routines, amount of data and specificity of routines. The dataset properties are detailed in this section. Each property was individually investigated to avoid co-occurring effects.

Duration of Routines and Amount of Data. The duration of a specific routine r^i was changed by varying the means μ_j of the normals $N_j(\mu_j, \sigma_j^2)$ in the simulation model. This was done for all states x_j showing an emission $e_j = r^i$. Given the varied mean μ_j^*, a corresponding standard deviation σ_j^* was adapted according to $\sigma_j^* = \sigma_j \mu_j^* / \mu_j$. We varied the number of simulated recording days to investigate the amount of data.

Specificity of Routines. We investigated the similarity of different routines to gain insight into how specific routines need to be for stable topic model performance. As measure for the similarity of two routines r^i, r^j the overlap o^{ij} of their primitive histograms h^i and h^j was derived according to:

$$o^{ij} = 1 - \sum_{s=1}^{m} |h_s^i - h_s^j|/2 \ . \tag{1}$$

The parameter h_s^i is the occurrence ratio of primitive s in routine i. The specificity of all routines in a dataset is described by the overlap o_{total}, which is the mean over all pairwise routine overlaps $o^{ij} | i \neq j, j > i$. The transition matrices T_i from the lower layer of our simulation model (see Fig. 2) were used as tuning parameters when sampling data. In order to derive less specific routines, the new transition matrix T_*^i is a combination of the original T^i and the transition matrices T^j of the other routines r^j: $j \in 1, .., k$; $i \neq j$. We derive the new transition matrices T_*^i by:

$$T_*^i = (1 - (k-1)p)T^i + \sum_{j=1; j \neq i}^{k} pT^j \ . \tag{2}$$

The tuning parameter $p \in [0, \frac{1}{m}]$ was used in our analysis. For $p = \frac{1}{m}$, all routines share an identical transition matrix and therefore would show an equal primitive histogram (see Fig. 3). For $p = 0$, T_*^i and T^i are identical.

Two routines are highly specific when they do not share the same activated primitives in their histograms. Therefore, we define the transition matrices T^i_{spec} for each routine r^i. T^i_{spec} was derived from T^i by copying a subset of state transition matrix components $t^i_{hj} \in T^i$ for selected primitives $h, j \in 1, 2, ..., m$. The other components $t^i_{rs}; rs \neq hj$ in T^i_{spec} were set to zero. The matrix T^i_{spec} was then normalized to a row sum of 1. Primitives h, j were chosen such that different routines do not share any primitives. Figure 4 provides an illustration of this setting. We derived the transition matrices T^i_* according to ($p \in [0, 1]$):

$$T^i_* = (1 - p)T_i + pT^i_{spec} \ . \tag{3}$$

For Eq. 3 and $p = 1$, different routines do not show overlap in their primitive histograms. In our evaluation, more specific routines were obtained by adapting T^i according to Eq. 3, whereas applying of Eq. 2 resulted in less specific routines compared to the basis (UbiComp'08 dataset).

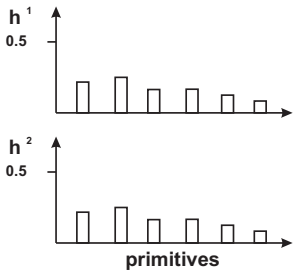

Fig. 3. Histograms h^1 and h^2 of routines 1 and 2 showing an equal mixture of all T^i's (100% primitive overlap).

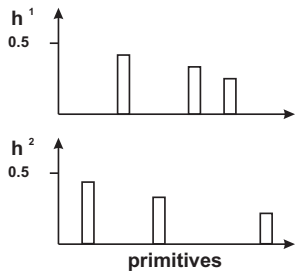

Fig. 4. Histograms h^1 and h^2 of routines 1 and 2 showing T^1_{spec} and T^2_{spec}, with no overlap in primitives.

5.2 Implementation

Simulation of Data. All dataset properties used in our simulation model were sampled from the UbiComp'08 dataset (Huynh et al. [5]). This setup formed the basis (*number of days*, k, T, E, μ_k,σ_k, m, T^i) for all simulations in this work.

The UbiComp'08 dataset covers seven days without sleeping phases. It contains 4 routine labels *dinner*, *commuting*, *lunch* and *office work*, and a *null* class. In total, 24 primitive labels are available as user annotations at a frequency of $f = 2.5Hz$, including activities such as *using the toilet*, *preparing food*, and *sitting/desk activities*.

To analyse routine duration r^i, the means μ_j with a corresponding emission $e_j = r^i$ and $\mu_j > 5\,\text{min}$ were swept in the interval $[10\,\text{min}, 150\,\text{min}]$. For the amount of data, we varied the number of sampling days in [3 *days*, 25 *days*]. To analyse routine specificity, we varied the tuning parameter p as described above.

Daily Routine Discovery. To evaluate the topic model we used the LDA implementation according to [2]. All topic model parameters were set corresponding to [5]. Documents were formed over a duration of 30 min, shifted by 2.5 min, the number of topics k_T was set to 10.

To obtain stable topic model estimation results, three training runs were performed, choosing the one with the highest likelihood. Primitives that occurred in a single day only were not considered, as well as the primitive *unlabeled*. We applied the *Borda Count* ranking method to the topic activations $\boldsymbol{\gamma_d}$ of all documents d covering the same 2.5 min time slot. The resulting total topic activation vectors showed a resolution of 2.5 min. After upsampling to the ground truth frequency, topic activations were used as input for the kNN classifier.

Evaluation. We used a leave-one-day-out scheme in our analysis. For data samples exceeding 11 days, we applied a 10-fold-cross validation. The *null* class was used for training the topic model and the kNN, but left out for evaluation. To compensate for the probabilistic data acquisition via sampling, we repeated for each dataset property analysis both, the dataset simulation and topic model calculations 20 times. The averaged accuracy and standard deviation of the 20 runs were analysed. In the evaluation we considered a standard deviation of less than 5% as stable performance. Performance variations below this standard deviation could be considered random, e.g. caused by the initialisation of models.

6 Results

Firstly, we validated our daily activity simulation model against the performance reported by Huynh et al. [5]. In the subsequent sections, we investigated selected dataset properties regarding the topic model stability for daily routine discovery.

6.1 Validation of the Daily Activity Simulation Model

Table 1. Comparison of performance regarding daily routine discovery. The performances for our topic model implementation and a simulated dataset sampled from the actual UbiComp'08 in [5] were compared against the results of [5].

Routine	**Huynh [5]**		**Our topic model**		**Simulated data**	
	recall	precision	recall/accuracy	precision	recall/accuracy	precision
dinner	40.2	75.5	73.6	71.8	74,8	71,8
commuting	51.8	85.5	82.9	90.6	86,3	85,1
lunch	83.3	87.0	86.7	91.6	75,4	76,2
office work	93.7	96.4	94.5	96.2	91,0	94,7
mean	67.2	86.1	84.4	87.6	81.9	82.0

When using the UbiComp'08 dataset [5], our topic model achieved a recognition performance as reported in Table 1. For the evaluation approach in this work, it is sufficient to consider the class-specific accuracies. In order to compare against Huynh et al., we show the precision here as well. Compared to Huynh et al., our topic model results showed higher precision and recall since we used labels and not classified primitives as topic model input.

In order to validate our simulation model we compared the recognition performance of our topic model implementation using the UbiComp'08 dataset [5] against a simulated dataset sampled from the UbiComp'08 dataset. On the simulated data, we achieved similar recognition performance compared to using the dataset directly, except for the *lunch* routine. For *lunch*, performance results of the simulated data were lower. We assume that the difference for this routine was due to an inadequate sampling of the primitive transition matrix T^i in the lower layer of our simulation model. We attributed the low performance variations of other routines to the random topic model initialisation. Comparing the routine consecution in a day and the histogram of primitives per routine we found that the UbiComp'08 dataset and the simulated dataset showed high similarity. Overall, structure and performances of simulated and actual dataset correspond well. Consequently, we considered the simulation model as capable of generating realistic datasets and suitable to analyse specific dataset properties.

6.2 Influence of Routine Duration

Figure 5 shows that with increasing document length performance and stability of the topic model for the best and the lowest performing routines of the dataset (*office work*, *dinner*) increase. This effect depends marginally on the total amount of data available for each routine: *dinner* occurred once a day, while *office work* occurred three times a day, and therefore comprises three times the data amount of *dinner*. Nevertheless, both routines show the same trend in accuracy and standard deviation for sufficiently long routines.

However for short durations below 50 min, larger data amounts lead to higher stability. Too short routines relative to the document length are highly instable, such as seen for *dinner*. Following our assumption of topic model stability (std<5%), both routines require 80 min duration. Hence, routine duration must considerably exceed the document length (30 min).

6.3 Influence of the Amount of Training Data

Amount of data particularly influences model stability, if few days of data (below 5 days) are available. Figure 6 shows the effects on performance. *Office work* and *commuting* are already stable (std<5%) at 9 days of data, while the total occurrence times of *office work* was 61.0 hours and of *commuting* 6.4 hours.

Although *lunch* (8.4 hours) had more data compared to *commuting* during 9 days, stable results were obtained for $\sim$ 14 days only (totalling to 13 hours of data). This result indicates that the total amount of data does not have an unique impact on stability. There are other parameters, such as specificity of

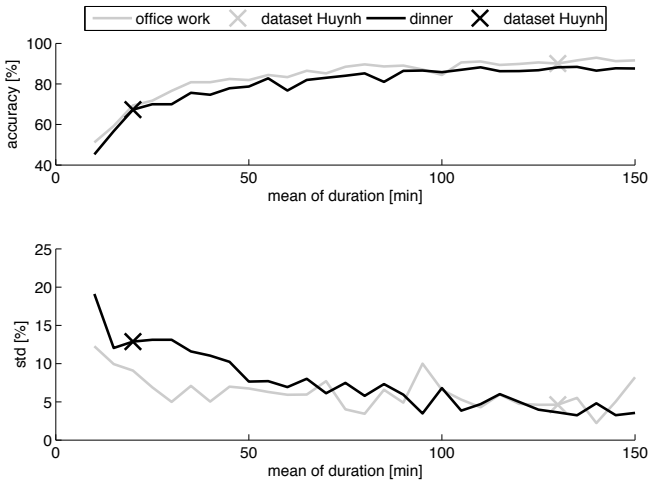

Fig. 5. Recognition accuracy and stability over 20 runs for *dinner* and *office work* using simulated data. Both routines become more stable for longer routine durations.

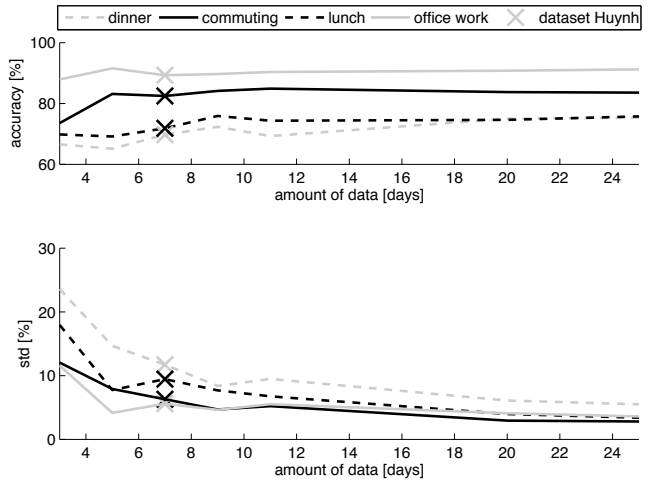

Fig. 6. Recognition accuracy and stability over 20 runs for simulated data conditioned on the number of recorded days for four routines.

the routine influencing it. Nevertheless, the number of recording days turns out to be one crucial tuning parameter for stable topic models. *Dinner* does not become stable in the considered interval at all. The analysis shows that, in order to ensure stability, data acquisition should be performed for ~ 14 days. More data can yield performance benefits for low frequent routines such as *dinner*.

6.4 Influence of the Specificity of Routines

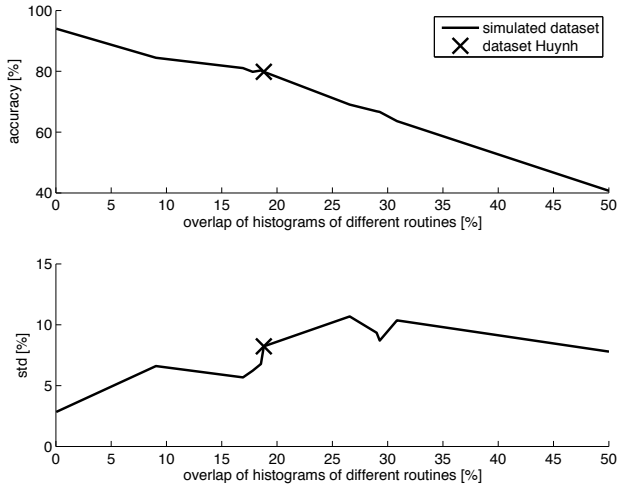

Fig. 7. Recognition accuracy and stability over 20 runs for simulated data conditioned on the specificity of routines. With increasing histogram overlap of different routines, the topic model becomes less stable.

The specificity of routines highly influences stability and performance of the topic model, as Figure 7 illustrates. With increasing overlap, routines become more similar and the topic model yields less stable results. Overlap highly depends on how specific routines are. For a 20% overall overlap in the dataset, *commuting* and *dinner* show a very low routine-to-routine overlap (4.4%), whereas *dinner* and *office work* appear similar in terms of their primitives (routine-to-routine overlap was 37%). This implies that the choice of primitives is highly connected to performance and stability. Thus, when targeting stability under the given topic modeling approach, the overlap may not exceed 5%.

7 Conclusion and Outlook

In order to analyse key dataset properties that could provide stable topic model performance, we investigated the duration of routines, amount of data and specificity of routines in this work. The choice of these dataset properties appears

essential for the experiment and data recording design, when targeting daily routine discovery.

The validation of our daily activity simulation model confirmed that performances closely resembling those in previously published work can be achieved. Subsequently, we created datasets of different characteristics by varying the selected dataset properties. Our investigations showed that specific requirements exist that would ensure stable topic model performance. In particular, routine durations need to be considerably longer than the document size and 14 recording days appeared essential in the considered conditions. Furthermore, we found that the primitive histogram overlap of different routines highly corresponds to topic model stability. Thus, a bound on the primitive overlap can be given to support the design of lower recognition layers.

Using our results, a suitable primitive set regarding number and granularity could be defined. The potential performance of the selected primitive set could be predicted by deriving the histogram overlap of routines. The preliminaries found in this paper can be used towards a stable topic model application in activity recognition.

Acknowledgments. This work was supported by the EU Marie Curie Network iCareNet under grant number 264738.

References

1. O. Amft, C. Lombriser, T. Stiefmeier, and G. Tröster. Recognition of user activity sequences using distributed event detection. *Smart Sensing and Context*, pages 126–141, 2007.
2. D. Blei. Implementation of lda at http://www.cs.princeton.edu/ blei/lda-c.
3. D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
4. K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2:3:1–3:27, January 2011.
5. T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 10–19, New York, NY, USA, 2008. ACM.
6. Y.-S. Lee and S.-B. Cho. Human activity inference using hierarchical bayesian network in mobile contexts. In B.-L. Lu, L. Zhang, and J. Kwok, editors, *Neural Information Processing*, volume 7062 of *Lecture Notes in Computer Science*, pages 38–45. Springer Berlin / Heidelberg, 2011.
7. J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299–318, 2008. 10.1007/s11263-007-0122-4.
8. N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.
9. Y. Wang and G. Mori. Human action recognition by semilatent topic models. 31(10):1762–1774, 2009.