SAMMPLE: Detecting Semantic Indoor Activities in Practical Settings using Locomotive Signatures

Zhixian Yan[1], Dipanjan Chakraborty[2], Archan Misra[3], Hoyoung Jeung[4], Karl Aberer[1]

[1]EPFL
Switzerland
{zhixian.yan, karl.aberer}@epfl.ch

[2]IBM Research Lab
India
cdipanjan@in.ibm.com

[3]Singapore Management University
Singapore
archanm@smu.edu.sg

[4]SAP Research
Australia
hoyoung.jeung@sap.com

Abstract

We analyze the ability of mobile phone-generated accelerometer data to detect high-level (i.e., at the semantic level) indoor lifestyle activities, such as cooking at home and working at the workplace, in practical settings. We design a 2-$Tier$ activity extraction framework (called SAMMPLE[1]) for our purpose. Using this, we evaluate discriminatory power of activity structures along the dimension of statistical features and after a transformation to a sequence of individual locomotive micro-activities (e.g. sitting or standing). Our findings from 152 days of real-life behavioral traces reveal that locomotive signatures achieve an average accuracy of 77.14%, an improvement of 16.37% over directly using statistical features.

1 Introduction

Semantic-level activity mining in wearable computing literature has traditionally been investigated in *smart home* environments, using object-embedded & multiple wearable sensors. Recently, activity recognition in the wild (i.e., in real-life practical settings with mobile phone sensors), is receiving a lot of attention [5]. We investigate the performance of detecting an individual's *lifestyle-related indoor semantic activities*, solely based on observations from a single phone-based accelerometer, without constraints on orientation and usage. It is natural that multiple sensors, while improving activity recognition accuracy, burdens the power-constrained mobile device and impacts user comfort. We investigate discriminatory power of the accelerometer sensor as it constitutes the most commonly-used, low-energy smartphone sensor.

In contrast to laboratory studies, we focus on *real-life naturalistic environments*. We utilize two user-generated data traces from 5 users. The first data set (*MICRO-SHORT*) is used to determine the best features for classifying locomotive activities in controlled but naturalistic conditions, where the smartphone's usage and on-body position varied dynamically. The second data set (*SEMANTIC-LONG*) captured accelerometer readings from the phones of these 5 users as they went about performing their daily lifestyle-based semantic activities, for a period of 6–8 weeks, giving us a rich observation data for 152 days. The key question we investigate is: In real life, does an individual's semantic activities possess enough regularity and discriminatory power, as observed by the phone accelerometer, and what level of accuracies can be achieved for various lifestyle activities?

Traditional approaches (e.g., [4]) towards recognition tasks from accelerometer data use various statistical features computed from the raw sensor streams. These work well when the activity exhibits recurring behavior over short spans of time, typically of the order of seconds. Semantic activities are often long running (minutes) and non-homogeneous (i.e. complex), resulting in temporal variations in the statistical features over the observation period. Moreover, in real-life settings, the placement & orientation of the accelerometers/phones vary unpredictably, affecting the statistical features computed over a longer period.

We analyze the discriminatory power of statistical features (1-$Tier$) towards classification of these semantic activity structures for real-life settings. Further, using the intuition that that locomotive and postural states are likely to be recorded relatively well from short observation windows, we study the discriminatory power of locomotive features (2-$Tier$) by first converting the signal to a sequence of an individual's micro-activities (MAs), as a specific set of these locomotion or postural states.

Prior investigations on building hierarchical representations of complex activities from low-level sensors [6, 3, 2] have mostly operated using multiple wearable accelerometers (e.g. on wristwatch and right pocket) along with object interaction data and showed that good recognition accuracy is achievable. In contrast to prior work, our data set is on the most practical usage scenario – a phone with unconstrained orientation and usage, with a very large data set (152 days of readings from 5 users). Accordingly, our re-

[1] SAMMPLE: Semantic Activity Mining via Mobile Phone-based Locomotive Estimation

sults on achievable accuracy serves a valuable baseline to the "smartphone-based sensing" research community.

2 The SAMMPLE Inference Framework

SAMMPLE is a 2-Tier classification framework (Fig. 1) that infers an individual's semantic activity (HA), using *micro-activities* (MA) as an intermediate step. In Layer I, the raw accelerometer data corresponding to a HA is first partitioned into a sequence of non-overlapping "frames" (T_f) of small duration (e.g., 5 secs, 10 secs). We extract statistical features for each frame, and classify each frame to a corresponding MA. The Layer II accepts this sequence of inferred MAs and extracts activity structures using features from this MA sequence.

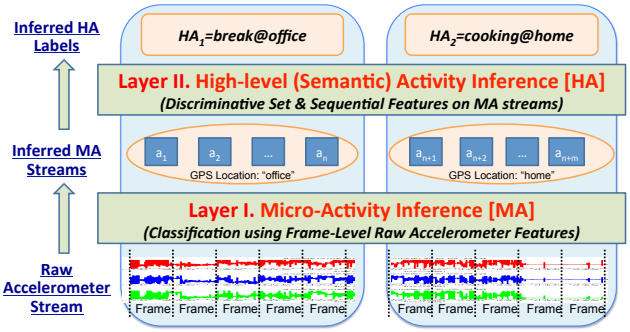


Figure 1: Our 2-Tier Semantic Activity Inferencing Process

2.1 Layer I: Micro-Activity Inference

For extracting MAs, we consider a feature vector, consisting of both *time* and *frequency* domain features from: (1) the 3D axis of the accelerometer, referred to as 3D-features; (2) a projection of the readings on the gravity direction ($\vec{p}$) and the plane perpendicular to gravity($\vec{h}$), which makes it orientation-independent (referred to as 2D-features). We use the fact that the mean of accelerometer readings, computed over a long time period gives an estimate of g [5], to project the raw signal to this "2D" reference frame. For the frequency domain, the features are computed by first transforming the (x_i, y_i, z_i) segment into a 250-point FFT vector [7]. Finally, a total of ~70 features (summarized in Table 1) are used per frame (F_i). A state-of-the-art calibration technique on the Nokia N95[9] is used to calibrate the sensor readings, before feature extraction.

2.2 Layer II: Semantic Activity Inference

We investigate two feature extraction approaches: *a)* the *Order-Oblivious* (*OO*) and *b)* the *Sequence-Aware* (*SA*) approach. For explanation, we utilize the illustrative example in Table 2, showing two HA instances (i.e., HA_1 - *Office_Break* and HA_2 - *Office_Lunch*), with a simple set of MAs (viz. 'walk (w)', 'sit (s)', 'stand (t)').

Order-Oblivious (OO) – Given the MA sequence associated with an HA instance, this approach creates an M-dimensional feature vector (M = No. of MAs), where the i^{th} element of the vector denotes the number of MAs of type MA_i. The feature vector thus captures the duration (as T_f is a constant) of each specific MA in the given HA instance.

Table 1: Features Used for Micro-Activity Classification

	Name	*Definition*
3D	calibrated (3D)	(x_i, y_i, z_i)
Projected 2D	Vertical [$\vec{p}$]	$\vec{p} = \frac{\vec{d}\cdot\vec{v}}{\vec{v}\cdot\vec{v}} \cdot \vec{v}$, where $v = \langle \bar{x}, \bar{y}, \bar{z} \rangle$ (the mean of x,y,z) and $\vec{d} = \langle x - \bar{x}, y - \bar{y}, z - \bar{z} \rangle$
	Horizontal [$\vec{h}$]	$\vec{d} - \vec{p}$
	Magnitude [mag]	$\lvert\vec{h}\rvert, \lvert\vec{p}\rvert, corr(\lvert\vec{h}\rvert, \lvert\vec{p}\rvert)$
Time Feature	Mean	AVG($\sum x_i$); AVG($\sum y_i$); AVG($\sum z_i$)
	Variance	VAR($\sum x_i$); VAR($\sum y_i$); VAR($\sum z_i$)
	Mean-Magnitude	AVG($\sqrt{x_i^2 + y_i^2 + z_i^2}$)
	Magnitude-Mean	$\sqrt{\bar{x}^2 + \bar{y}^2 + \bar{z}^2}$
	2-Axis Correlation	$corr(xy) = \frac{cov(xy)}{\sigma_x, \sigma_y}$; $corr(yz)$; $corr(xz)$
	Signal-Magnitude-Area	$\frac{1}{n}\sum_{i=1}^{n}(\lvert x_i\rvert + \lvert y_i\rvert + \lvert z_i\rvert)$
Frequency	FFT Magnitude	$m_j^{(x)} = \lvert a_j + b_j i\rvert$; similarly, $m_j^{(y)}$, $m_j^{(z)}$
	FFT Energy	$\frac{\sum_{j=1}^{N}(m_j^2)}{N}$, for x,y,z respectively
	FFT Entropy	$-\frac{\sum_{j=1}^{n}(p*log(p))}{n}$, for each axis, where p is normalized histogram count of FFT components

Sequence-Aware (SA) – This approach extracts features that also capture the *order* in which the various MAs occur within a HA instance. This approach should improve discriminatory capability of the resulting features, compared to the *OO* approach which does not utilize such knowledge. However, it comes at the expense of higher dimensionality of the feature vector. We consider two pattern mining-based techniques to learn key discriminatory features from the underlying traces: SA-TD (a *duration-preserving* strategy) and SA-TP (a *transition-preserving* strategy). To explain them, we first define a few terms.

Let M_i=[$MS_1^{(i)}, \ldots MS_l^{(i)}$] be l MA sequences as the different instances of HA_i, e.g., $MS_1^{(1)}$ =[t t t t t t w w w w t] in Table 2. the set of MA sequences associated with the Let $\mathcal{S}_j^{(i)}$ be the set of all *subsequences* of $MS_j^{(i)}$. Let sub_c be a MA subsequence, i.e. $sub_c \subseteq \mathcal{S}_j^{(i)}$.

Definition 1 (Cover of sub_c)**.** *Denoted as* ***cov****(sub_c, M_i), equals the number of instances in M_i where sub_c is present.*

Definition 2 (Support of sub_c)**.** *Denoted as* ***supp****(sub_c, M_i), equals* $\frac{cov(sub_c, M_i)}{l}$.

For example, Col.3 in Table 2 shows the length-3 subsequence [t t w] is present in two amongst the three (l=3) instances of HA_1. Hence, we have cov([t t w],HA_1)=2 and $supp$([t t w],HA_1)=2/3.

• *Duration-preserving (SA-TD):* Given a minimum support threshold Θ_0 and a maximum sub_c size K_{max}, this strategy discovers the set of all sub_cs of length $[2, 3, \ldots, K_{max}]$ that have $supp(sub_c, HA_i) \geq \Theta_0$. For example, Col.3 in Table 2 shows that the sub_cs of length 3 selected with $\Theta_0 \geq 0.6$ for HA_1 are {t t w}, {t w w}, {w w t}. The SA-TD algorithm finds the *union* of all such qualifying sub-

Table 2: Running example of feature selection in the Locomotive Signature Space in the 2-$Tier$ approach.

Col. (Column) 1		*Col. 2*	*Col. 3*			*Col. 4*	*Col. 5*	*Col. 6*			*Col. 7*
MA Streams of 2 Types HA_1: Office_break HA_2: Office_lunch *[w:walk, s:sit, t:stand]*		OO Features *[w, s, t]*	SA-TD Patterns Subseq (size: 3, θ≥0.6)			SA-TD Features *[w, s, t, ttw, tww, wwt, tts, tss, sst]*	T-P Seq	SA-TP Patterns Subseq (size: 3, θ≥0.6)			SA-TP Features *[w, s, t, twt, tst]*
			sub_c	cov	supp			sub_c	cov	supp	
HA_1	$MS_1^{(1)}$: [ttttttwwwwt]	[4, 0, 7]	[ttt]	1	1/3	[4,0,7,1,1,1,0,0,0]	[twt]	[twt]	2	2/3	[4,0,7,1,0]
	$MS_2^{(1)}$: [tww]	[2, 0, 1]	[ttw]	2	2/3	[2,0,1,0,1,0,0,0,0]	[tw]				[2,0,1,0,0]
	$MS_3^{(1)}$: [ttwwtt]	[2, 0, 4]	[tww]	3	3/3	[2,0,4,1,1,1,0,0,0]	[twt]				[2,0,4,1,0]
			[www]	1	1/3						
			[wwt]	2	2/3						
			[wtt]	1	1/3						
HA_2	$MS_4^{(2)}$: [ttssssst]	[0, 5, 3]	[tts]	2	2/2	[0,5,3,0,0,0,1,1,1]	[tst]	[tst]	2	2/2	[0,5,3,0,1]
	$MS_5^{(2)}$: [wwttsst]	[2, 2, 3]	[tss]	2	2/2	[2,2,3,0,0,1,1,1,1]	[wtst]	[wts]	1	1/2	[2,2,3,0,1]
			[sss]	1	1/2						
			[sst]	2	2/2						
			[wwt]	1	1/2						
			[wtt]	1	1/2						

sequences across all HAs in the training data. The resulting features are appended to the OO features to create a longer OO+$Sequence$ feature vector, with the i^{th} element of the vector corresponding to the frequency of occurrence of the corresponding feature (Refer Col. 4 in Table 2).

• *Transition-Preserving (SA-TP):* Different from SA-TD, SA-TP preserves only the transitions between *distinct, adjacent* MA instances, by removing (or collapsing) the run-length of consecutively repeating MA symbols for each HA instance. E.g., Col.5 in Table 2 shows the TP sequence of $MS_1^{(1)}$ is transformed to [t w t]. By focusing purely on the sequence of *transitions* among distinct MAs, the approach ignores slight variations in the duration of an individual MA and produces the transition structure.

3 Experimental Results on HA Detection

Data Collection: We recruited 5 users; each one was provided with a Nokia N95 phone with embedded Python scripts that sampled the accelerometer sensor at 30Hz. The data (along with an inferred GPS-location - *Home* or *Office* - by [8]) was periodically transmitted to a back-end server.

- *MICRO-SHORT:* Each user was asked to perform a set of 7 MAs: {sit, sit active, walk, loiter, bursty move, stand, using stairs}, consecutively for 7-10 mins each, resulting in a per-user study duration of 50-60 mins. These MAs were chosen based on user's feedback of locomotions commonly associated with their daily-life at home and office.
- *SEMANTIC-LONG:* Each user carried the phone in their preferred position as they went about performing their daily lifestyle activities. Alongside, they maintained a separate diary where they tagged *all the HAs* performed, at their office and home locations. This longitudinal data was gathered for 8 weeks on working days, with gaps due to individual variations in lifestyle routines (see Table 3).

Table 3: Summary of Semantic Activity Dataset

	User1	*User2*	*User3*	*User4*	*User5*
# of Days	27	31	39	32	23
# of unique HAs	30	64	25	41	65
# of HA instances	194	215	372	167	228
# of HA instances selected	186	203	356	165	192

Tagging Process & Principles: Each user recorded the tag tuples: *[activity_start_time, activity_tag]*. As the activities were sequential, the end time of an activity was derived from the start time of the next tag. In total, we obtained 152 days of data, with each day having 4-15 tags/person. This data was cleaned by applying a per-user manual process of normalization and information summarization, resulting in a total of 1102 HA instances across all users; some detailed tags are shown in Table 4.

Table 4: Examples showing Activities collected (right column) and corresponding normalized Tags (left column)

HA Label	Examples of User Tags
O_work	office_work, work_work, office_work_TA, office_work_check_printer
O_break	office_break, office_break_walk around office_break_talk,
O_coffee	office_coffee_break, office_break_tea, office_short_break_coffee
O_toilet	office_break_toilet, work_break_toilet, office_short_break_toilet
O_meet	office_meet, office_meet_lab, office_meeting, office_meet_NRC
O_lunch	office_lunch, work_lunch, office_lunch_desk, office_break_lunch
H_work	home_work, home_work_move, home_work_on_computer
H_relax	home_relax, home_relax_freshen_up, home_relax_movie
H_break	home_break, home_break_shopping, home_break_coffee
H_cook	home_cook, home_cooking, home_clean_dishes, home_wash_dishes
H_eat	home_eat, home_lunch, home_dinner, home_eat_dinner
H_baby	home_baby_routine, home_baby_routine_eat_with_baby

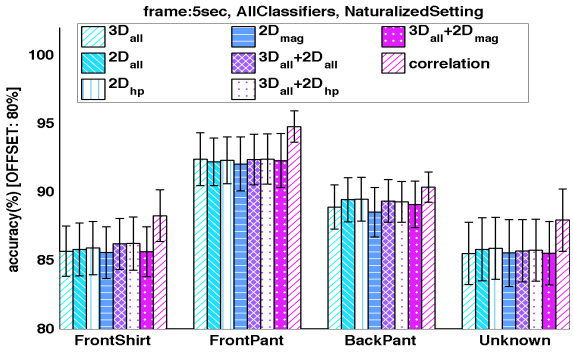
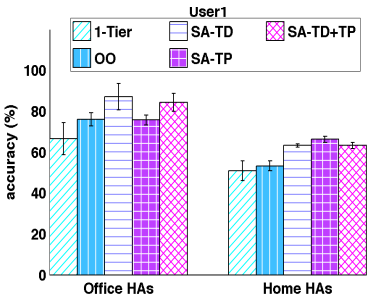


Figure 2: MA Classification accuracy for *User1* with naturalistic (varying) phone orientations (T_f=5 secs) and with unknown body positions. 'FrontShirt'=shirt pocket in the chest, 'FrontPants'=front pocket in the pants, 'BackPants'=back pocket in the pants. 'Unknown'=body position is mixed and not given.

Results of MA Inference: We present the results using 10-fold cross validation. We tested many classifiers[2] and evaluated discriminatory power of different feature choices. Fig. 2 plots the MA accuracy on $User1$ with 9 feature choices and 4 phone positions. The plot shows the average accuracy (& the standard deviation) over all of the classifiers. In Fig. 2, $3D_{all}$ (or $2D_{all}$) implies the use of all 3D-features (or 2D-features); $2D_{hp}$ refers to features computed

[2] *Decision tree – J48*, *Naive Bayes*, *Bayesian network*, *LibSVM*, and *Adaptive Boost* (Adaboost) using J48 as the meta learner



HA	No.	1-Tier Method Confusion Matrix & Avg. Accuracy						2-Tier Method (SAMMPLE) Confusion Matrix & Avg. Accuracy						Gain
O_work	30	.900	.100	.000	.000			.900	.033	.067	.000			
O_break	17	.176	.824	.000	.000		57.53%	.000	.941	.059	.000		87.14%	29.61%
O_meet	15	.467	.467	.067	.000			.143	.071	.786	.000			
O_lunch	11	.000	.818	.182	.000			.000	.111	.111	.778			
H_work	36	.833	.033	.100	.000	.033		.931	.000	.000	.034	.034		
H_cook	21	.143	.619	.190	.048	.000		.000	.789	.158	.000	.053		
H_relax	25	.053	.526	.316	.000	.105	45.54%	.000	.316	.368	.105	.211	66.33%	20.79%
H_break	14	.071	.571	.214	.000	.143		.000	.071	.143	.429	.357		
H_eat	17	.176	.000	.471	.000	.353		.118	.059	.000	.235	.588		

Figure 3: Performance of 1-$Tier$ vs SAMMPLE ($K_{max} = 4$; $\Theta_0 = 0.7$; MA frame-size $T_f = 5secs$) on $User1$. 'Office HAs' and 'Home HAs' indicate office and home semantic activities. (Left: accuracy comparison; Right: confusion matrix.)

on the projected orientation-independent frames, including both gravity ($\vec{p}$) and its plane perpendicular ($\vec{h}$); $2D_{mag}$ refers to features computed over magnitudes of $\vec{h}$ and $\vec{p}$; 'correlation' refers to using correlation-based feature selection. We observe that *a)* the MA classification accuracy is higher when the phone is placed in the lower part of the body (an observation previously made with multiple body-worn sensors [1]); *b)* the choice of feature classes result in performance differences of $\sim$5%, and using feature selection is the best in classifying such naturalized usage data; *c)* the classification accuracy for the "unknown" case, which best reflects naturalistic usage conditions, is at an acceptably healthy $\sim$90%.

Results of HA Inference: Fig. 3 plots the HA classification accuracies for the first $User1$. The HA classification accuracies were obtained by 8-fold cross validation, as some HA instances did not have enough samples to perform 10-fold. Further, we compare four feature extraction strategies in SAMMPLE – OO, $SA\text{-}TD$, $SA\text{-}TP$ and the combined feature sets $SA\text{-}TD{+}TP$. We applied a variety of classifiers (*decision tree – J48*, *Adaptive Boost - Adaboost*, *LibSVM*, *Bayesian Network* and *Naive Bayes*). The plots show mean & standard deviation of the accuracies across these classifiers. The right of Fig. 3 shows the confusion matrices obtained using the locomotive signatures (2-$Tier$) and the statistical features (1-$Tier$).

Across 5 users, locomotive signatures results in an improvement in the classification accuracy ranging from 7-30%, compared to the 1-$Tier$ approach. Due to the different dynamics of lifestyle activities of different users, the absolute accuracy values are user-dependent. A salient observation is that even the OO approach, with a slim feature vector dimension (7 MAs), mostly out-performs the 1-$Tier$ approach which uses $\sim$70 statistical features (with correlation-based feature selection). We also note that sequence-based features provide an additional but variable (4-15%) amount of improvement in the classification accuracy. Typically, the sequential features provide better performance improvement on home activities, but not much on office ones. Nevertheless, using both OO (set) and SA (sequential) features establish the superior quality of locomotive signatures, compared to their statistical counterparts.

We also studied the sensitivity of choosing different parameters, e.g., K_{max}, i.e., the maximum possible sequence length considered in SA-TD & SA-TP. We observe that relatively-short MA sequences possess the highest discriminatory power (e.g., K_{max}=4).

4 Conclusions

This paper evaluated the power of locomotive signatures to infer semantic activities (HA) in realistic environments, using data from a single phone-embedded accelerometer. Our investigation finds that locomotive and postural features as low-level, reliably extractable events achieve 16.37% accuracy gain, over semantic activity structures extracted directly using statistical features. Overall, the average accuracy achieved was 77.14% with many activities exhibiting over 85% accuracy. We plan to release the longitudinal activity data set for further research by the community.

References

[1] L. Bao and S. Intille. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive*, pages 1–17, 2004.
[2] U. Blanke and B. Schiele. Remember and transfer what you have learned - recognizing composite activities based on activity spotting. In *ISWC*, pages 1–8, 2010.
[3] T. Hunh, M. Fritz, and B. Schiele. Discovery of Activity Patterns using Topic Models. In *UbiComp*, 2008.
[4] H. Lu et al. SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones. In *MobiSys*, pages 165–178, 2009.
[5] H. Lu et al. The Jigsaw Continuous Sensing Engine for Mobile Phone Applications. In *Sensys*, pages 71–84, 2010.
[6] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *International Conference on Multimodal Interfaces*, pages 3–8, 2002.
[7] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity Recognition from Accelerometer Data. In *AAAI*, pages 1541–1546, 2005.
[8] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. In *EDBT*, pages 259–270, 2011.
[9] J. Yang et al. Physical Activity Recognition with Mobile Phones: Challenges, Methods, and Applications. In *Multimedia Interaction and Intelligent User Interfaces. Advances in Pattern Recognition*, pages 185–213, 2010.