Mining Complex Activities in the Wild via a Single Smartphone Accelerometer

Angshu Rai
EPFL
Lausanne, Switzerland
angshurai1@gmail.com

Zhixian Yan
EPFL
Lausanne, Switzerland
zhixian.yan@epfl.ch

Dipanjan Chakraborty
IBM Research
New Delhi, India
cdipanjan@in.ibm.com

Tri Kurniawan Wijaya
EPFL
Lausanne, Switzerland
tri-kurniawan.wijaya@epfl.ch

Karl Aberer
EPFL
Lausanne, Switzerland
karl.aberer@epfl.ch

ABSTRACT

Complex activities are activities that are a combination of many simple ones. Typically, activities of daily living (ADLs) fall in this category. Complex activity recognition is an active area of interest amongst sensing and knowledge mining community today. A majority of investigations along this vein has happened in controlled experimental settings, with multiple wearable and object-interaction sensors. This provides rich observation data for mining. Recently, a new and challenging problem is to investigate recognition accuracy of complex activities *in the wild* using the smartphone.

In this paper, we study the strength of the energy-friendly, cheap, and ubiquitous accelerometer sensor, towards recognizing complex activities in a complete real-life setting. In particular, along the lines of hierarchical feature construction, we investigate multiple higher-order features from the raw sensor stream (x, y, z, t). Further, we propose and evaluate two SVM-based fusion mechanisms (early fusion vs. late fusion) using the higher-order features. Our results show promising performance improvements in recognizing complex activities, w.r.t. prior results in such settings.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Sensor Data Mining]

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

activity recognition, accelerometer, complex activities

1. INTRODUCTION

After the "perfect storm" around 2008 with the onset of Android and iPhone app stores, there has been a tremendous increase in exploiting the sensing capabilities on the mobile [7, 14] for a variety of application segments – games, context-aware communications, healthcare, community sensing etc. Activity recognition using the mobile phone is an active area of research today. In this paper, we investigate recognition performance of *complex activities* like "having lunch", "cooking", "working at desk" etc, using solely a single smartphone-based accelerometer, *in the wild*. A "complex activity" is an activity that consists of multiple sub-activities from a given observation space. "In the wild" means real-life, unconstrained experimental setups that reflect reality as close as possible [12]. Though multiple sensors on the phone can be used for activity recognition, the accelerometer sensor is energy-friendly. Recent results show that it is possible to continuously sense the user's activities [20, 24] by keeping the accelerometer running, without causing a visible dent to the battery budget. This is unlike other sensors such as GPS and Microphone, which have much higher energy footprints, and often impose bigger privacy threats (e.g. microphone). Efficient recognition of complex activities (also referred to as activities of daily living – ADLs) is valuable for many applications ranging from building energy efficient systems to health care.

Our work is motivated by the broader goals of the Wattalyst Project[1] which aims at studying correlations between a user's activities and electrical apparatus she used, in order to reduce energy consumption. Here, we need mechanisms to continuously monitor a user, immersed in her daily activities in real life. For reasons cited above, the accelerometer on the smartphone is a perfect fit to investigate activity recognition in realistic, uninstrumented indoor spaces.

The accelerometer, when fixed to a body part, records ambulatory movements of it. In many prior works, the accelerometer has been used to study "simple activities" like sitting, standing, walking etc, that typically refer to locomotions and postures. Here, the underlying observation is reasonably periodic and predictable [15, 16]. Detection of complex activities is more challenging as the generated sensor data is aperiodic, unpredictable in nature [10]. Moreover, our restriction of using a single smartphone in the wild (which is arguably the real situation today), confounds the situation. Data from a single mobile phone accelerometer does not capture ambulatory movements of multiple body

SensorKDD '12, August 12, 2012, Beijing, China.

[1] http://www.wattalyst.org/

parts. The anchor point (location of the phone) guides the sensitivity to record movements, is largely in the pockets (pant, or shirt) and is user-driven rather than experiment design-driven. Moreover, the location can change dynamically. Finally, naturalistic data reflects far greater unpredictability and usage-dependent artifacts. For example, in our data, a user can receive a phone call while she is cooking, accept the call and talk for a few minutes (while shifting the phone position around) and then put the phone back in a different body position (or even on the kitchen counter).

Our focus in this paper is to investigate several feature classes from a single accelerometer observation space, that show robustness in detecting complex activities in naturalized settings. Most of the research on complex activity recognition has till date, focused on using *multiple* wearable accelerometers (along with a sensor platform – e.g., eWatch [17]) to get rich body movement signals and associated context (like [8, 21]). We cover a range of related work later (Sec. 7). With respect to investigations using a single accelerometer in naturalized settings, our recent prior work [23] shows that certain locomotive features such as walking, sitting, standing etc, learnt through supervision, form better features than their statistical counterparts, for complex activity classficiation. This is intuitively explained as locomotive and postural states are likely to be recorded relatively well from short observation windows. Complex activities are usually long running (of the order of minutes) and nonhomogeneous, resulting in temporal variations in the statistical features over the observation period.

However, learning locomotion features through a supervised process necessitates the subjects to go through a training phase. In real-life, there can be an arbitrary, personal set of movements and motions that characterize a complex activity. It is onerous to conduct such a survey and training for multiple users. In this paper we investigate the following questions: (1) Can we learn good features from short observation windows through unsupervised techniques such as clustering? (2) How much discriminatory power does such unsupervised learning of features have, towards detecting complex activities?

For investigation, we collected accelerometer records (sampled at 30Hz) from the primary mobile phones (Nokia N95) of 5 subjects continuously, as they went about performing indoor activities, for a span of approx. 8 weeks. We summarize our contributions:

- To perform complex activity recognition, we first empirically evaluate an unsupervised clustering based approach towards constructing higher order feature dimensions. This can reduce training time necessary for learning the locomotions and postures from users.
- We evaluate multiple higher order features (set, regression, topic models) of complex activity structures, obtained using sequences of data clusters.
- We propose and evaluate multiple support vector machine-based fusion approaches to learn mixed models from these higher order features.
- Our results suggest that, using our methodology, complex activities can be classified with very high accuracy (85-90%) for real-life settings.

2. ACTIVITY DATA COLLECTION

We now describe the process of collecting, cleaning and tagging our unique ground-truth annotated, longitudinal accelerometer dataset that provides valuable insights into both the activities of an individual and her/his usage of the smartphone, under completely-natural conditions.

2.1 Recruitment & Instructions

During data collection campaign, 5 users (3 Ph.D. students, 1 post doctoral researcher, 1 co-author) volunteered to carry a Nokia N95 phone, loaded with an application that sampled the accelerometer at 30Hz, continuously 24×7. Four users used it as their primary cellphone for the duration of our data collection. Users were instructed to tag their activities in a separate diary while they conducted chores, only in their home and office locations. The only specific instructions given were that the users should carry the phone with them, in their preferred way, while they were tagging activities, and that they should charge the phone only when they were not tagging themselves (typically in the night during sleeping). This longitudinal data was gathered over a span of 8 weeks on working days, with gaps due to individual variations in lifestyle routines.

The user tagging process and principle followed was *unconstrained* – users had their own discretion on what activities to tag. The users were provided an initial idea on what constituted a complex activity (e.g., work, break, lunch). Although not mandatory, users could provide additional detail for each activity (e.g., break_coffee, break_toilet, work_at_desk). 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. The last activity performed on a certain day at a certain location had an explicit end time registered by the user. Fig. 1 shows the tag cloud from our data collection campaign. There are 1284 tags in total, with 177 unique tags.

Figure 1: Tags of complex activities collected

2.2 Data Processing & Sanitization

The data was cleaned by applying a per-user manual process of normalization and information summarization: (1) Semantically equivalent tags (e.g., office_meet, office_meeting)

were converted to a standard notation. (2) Tags having additional context were collapsed to the corresponding root tag (for instance, office_meet_colleague → office_meet), unless the activity occurred very frequently, and vice versa, e.g., the activity office_break_toilet was separated from office_break for some users. Infrequent tags were subsequently removed from further investigation (e.g., home_freshenup).

In total, we obtained 152 days of data, with each day containing between 4-15 tags/person. Table 1 provides the person-specific summary of the collected data. The final, cleaned data contains records of a total of 1102 complex activities across all users.

Table 1: Summary of Complex Activity Dataset

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

3. INFERENCE METHODOLOGY

In this section, we formulate definitions and describe our problem and methodology of mining complex activities more specifically. Then we provide a primer of our inference approach, particularly focusing on the feature space explorations we perform in the rest of the paper.

3.1 Definitions and Problem Statement

Fig. 2 shows an example of complex activities in daily life. As shown in this figure, we first collect the raw accelerometer stream data (x, y, z), and then segment it into accelerometer frames of size τ. Statistical features are extracted from these frames. Several such frames combined together constitutes a complex activity (e.g. CA_1=cooking, CA_2=eating, CA_3=working).

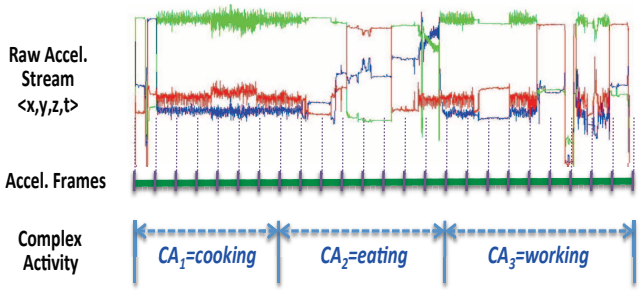


Figure 2: Complex activity recognition problem

We first present the necessary mathematical definitions needed to describe the steps in the rest of the paper.

Definition 1 (Accelerometer Data - $\mathcal{A}$). *A sequence of data points recording acceleration along 3 axes, i.e.* $\mathcal{A} = \langle a_1, a_2, \ldots, a_n \rangle$, *where* $a_i = (x_i, y_i, z_i, t_i)$ *is a tuple with accelerations* (x_i, y_i, z_i) *and timestamp* t_i. *We further assume that the accelerometer data is associated with a semantic location tag (e.g. derived from the processing of concurrently generated GPS samples)* $sl \in \mathcal{SL}(= \{SL_1, \ldots, SL_k\})$, *where* $\mathcal{SL}$ *is the set of distinct semantic locations we consider. In other words,* a_i *is a 5-tuple of the form* $(x_i, y_i, z_i, t_i, sl_i)$. *For this paper, we limit ourselves to two semantic locations* $(k = 2)$ *{"home", "office"}* [2].

[2] $\mathcal{A}$ was generated with a sampling frequency of 30 Hz, resulting in a sample size of $\sim$ 3MB per day

We do not recognize the complex activities directly from this raw accelerometer stream $\mathcal{A}$, but based on each group of accelerometer records in a time frame τ, where τ is typically small (e.g., 5 secs or 10 secs). We make a reasonable assumption that each such group corresponds to a specific locomotion/postural state of the user, e.g., sit, stand, arm-up. There are several advantages of using frames rather than the raw accelerometer directly for semantic activity recognition:

- *Robustness to outliers* – Our data collection is in naturalized settings where subjects use smart phones without any restrictions. The accelerometer sensor data $(\mathcal{A})$ is quite sensitive to various phone-specific usages, irrelevant motions and outliers, e.g., while changing the phone position, putting on the hands-free, checking the time etc. Therefore, temporally consecutive raw acceleration records (a) can vary, even though the activity (e.g. walking, cooking) remains invariant. Grouping the raw records a into frames reduces the effect of such outliers on the statistical features, and aids in achieving robust segmentation algorithms using the frames.

- *Storage Efficiency* – The raw data stream is huge, as accelerometer data is usually sampled at a reasonably high frequency (e.g., 30 Hertz). For example, in our experiment, accelerometer is sampling at 30 Hz, resulting in 1800 records/minute. However, by using a frame size of 5 secs $(\tau = 5)$, the ratio between accelerometer frames and accelerometer records is $\frac{1}{5*30}$, a remarkable compression before data analysis. Even though, such compression leads to information loss in theory, prior experiments like [12, 23] have established that frame-based features are more robust and leads to higher classification accuracy for different activities.

Definition 2 (Accelerometer Frame - $\mathcal{A}^{(\tau)}$). *A set of continuous accelerometer records, i.e.* $A^{(\tau)} = \{a_{i+1}, \ldots, a_{i+k}\}$ *in time duration* τ. *Basic features are calculated to describe the frame characteristics* $\langle f_1, \cdots, f_l \rangle$, *where* f_j *can be time domain features like mean, variance of the three axes* (x, y, z) *and frequency domain features like energy, entropy etc.*

In order to transform $\mathcal{A}$ into a sequence of frames, we use $\approx$ 70 statistical features from the time and frequency domain of the time series. For each accelerometer frame $A^{(\tau)}$, we assign a label $l_i^{(A)} \in \mathcal{L}_A$. $l_i^{(A)}$ can be inferred in the following two ways: (1) supervised learning using a training data representing well-known locomotions and postures to infer labels such as 'sitting', 'standing', 'jogging'. These labels are commonly used in many activity recognition literatures like [1, 23]. (2) unsupervised learning to infer the clusters of accelerometer data, then each frame is assigned a cluster label. Note that cluster labels represent frames that show similarity w.r.t. its statistical features. We do not know the specific locomotion or posture, a cluster corresponds to. Based on the inference on accelerometer frames $(A^{(\tau)})$, we can transform a sequence of raw accelerometer data into a sequence of $l_i^{(A)}$ labels for each complex activity.

Definition 3 (Complex Activity - CA). *A sequence of accelerometer frame labels,* $CA = \langle l_1, \cdots, l_n \rangle$, *where* $l_i^{(A)} \in \mathcal{L}_A$ *is the label corresponding to the accelerometer frame* $A_i^{(\tau)}$.

Therefore, the complex activity recognition task is to detect a high-level label (e.g., cooking, taking-dinner, working) from the sequence of low-level accelerometer frame labels (i.e., $l^{(A)}$). The complex activity label is noted as $l_i^{(CA)}$, corresponding to the accelerometer frame label $l^{(A)}$. We have $l_i^{(CA)} \in \mathcal{L}_{CA}$, which belongs to the tag cloud in Fig. 1.

3.2 Complex Activity Learning Approaches

Prior work in [23] shows that a two-tier complex activity learning approach, by converting the raw stream into a set of locomotive and postural states and extracting features from the resulting transformed stream, provides better accuracy than traditional one-tier approach where statistical features computed from the raw stream are directly used to classify complex activities. Moreover, due to the high sampling frequency of accelerometer data, such a two-tier approach provides better compute performance compared to inference using the raw accelerometer data (x, y, z) directly.

In this paper, we extend the two-tier framework along the following lines (see Fig. 3):

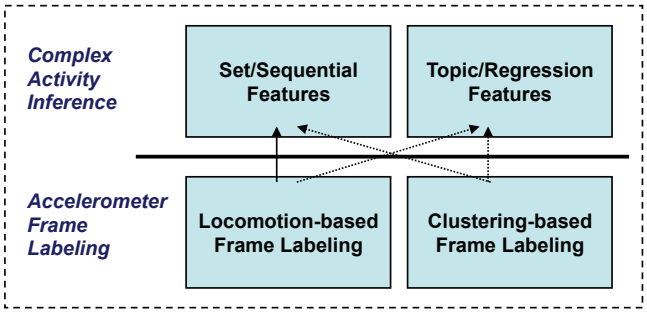


Figure 3: Activity Inference Approaches

- In the lower tier (*Accelerometer Frame Labeling*), besides relying on user-labeled training data of well-known locomotive and postural states (called "micro-activities"), we additionally evaluate an unsupervised clustering approach, to transform the raw data stream into a sequence of cluster labels. This eliminates needs for the tedious task of manually tagging data for such micro-activities. Moreover, this also avoids the tag errors due to intrinsic human errors given the invasive nature of such a tagging process. Finally, since clustering looks on data similarity, we discover several personalized movement patterns (which in turn can be used as latent features), which are impossible to determine through a manual survey-driven process for finding the representative micro-activities of each subject.
- In the higher tier (*Complex Activity Inference*), we study a variety of higher-order features from the sequence of labelled accelerometer frames. Such labels can be derived using the locomotion-based micro-activities, or the clustering-based labels. For higher-order features, we consider two groups: one is about set and sequential features, which are generated based on the counting statistics from the frame labels; the other group is to build advanced models and extract higher-order features, e.g. (1) regression model to represent the activity evolution and extract regression features, (2) LDA (latent dirichlet allocation) based topic model on the frame labels to extract topic features.

4. ACCELEROMETER STREAM LABELING

In this section, we discuss two types of lower-tier tasks to label accelerometer frames, in order to transform the raw accelerometer data stream into frame-level feature dimensions (e.g. locomotions, postures).

As mentioned before, one approach is to infer such frame-level features by using supervised learning. In this model, the user is asked to perform a set of well-known (or user feedback-based) locomotions (e.g. walk, sit, stand, recline, climb stairs etc). Data is collected and classification models are built. Using these models, an unknown stream of accelerometer frames are classified, to convert the raw stream to a sequence of symbols in the derived higher-order dimension. In our experimental study, we collect training data with 7 micro-activity tags (*sit, sit active, walk, loiter, bursty move, stand, using stairs*), build classification models, and use these tags to label each accelerometer frame.

Based on the study of using such user feedback-based locomotion tags to label the accelerometer frames[23], we observe some drawbacks:

- *Training Burden:* We need to request the users to spend some time tagging their locomotions and postures. This is an onerous task for the user. In addition, the user often needs to be shadowed, with a diary, in order to collect such data. Moreover, the user has to avoid using the smartphone for regular purposes during this data collection, in order to collect representative samples of the micro-activities.
- *Inability to capture complete reality:* A survey-driven method captures only a set of micro-activities performed by the user. Clearly, in real-life, while performing complex activities, users perform several other diverse micro-activities. For example, walk itself can have several variations, like walk slowly, walk briskly, walk with a strut etc. A user is expected to perform many kinds of such numerous real-life micro-activities. It is impractical and difficult to capture all of them through a supervised training set.
- *Personalization issues:* Users will have different personalized ways of performing micro-activities. For example, while "watching TV", an user might be "sitting and shaking legs" most of the time, while another might be sitting still. It is difficult to capture such personalized micro-activities, without an exhaustive monitoring of the user's complex activities, which is again impractical in real-life settings.

To overcome these problems, we label the accelerometer frames using an unsupervised clustering technique. We perform k-means clustering on the accelerometer data available from the semantic activity logs. In order to do this, the raw data in each accelerometer frame is represented as a vector of statistical features. Table 2 lists some of the features we used, including both time and frequency domain features. This approach provides a data-driven mechanism to group frames that intuitively represent similar (but unknown) micro-activities. This eliminates the training burden. Secondly, as derived from the same data stream as the complex activities, they represent reality. Thirdly, personal patterns are more likely to be captured using clusters.

We vary the number of clusters (K) to determine the best performing clustering configuration. In this process,

Table 2: Selected features used for activity recognition

Time Domain	Mean $(\bar{x}, \bar{y}, \bar{z})$, Magnitude $(\sqrt{x^2+y^2+z^2})$, Variance $\{var(x), var(y), var(z)\}$, Covariance $\{cov(x,y), cov(y,z), cov(x,z)\}$,
Frequency Domain	Energy $(\frac{\sum_{j=1}^{N}(m_j^2)}{N})$, m_j is FFT component Entropy $(-\sum_{j=1}^{n}(p_j * log(p_j))$, p_j is FFT histogram

we apply well-known clustering metric, i.e, DBI (the Davies-Bouldin index) K [9], to evaluate the clustering performance using different. We empirically determined the number of representative clusters for each user (between 6-14), and found the most suitable cluster number is $K = 10$. Therefore, we transform the original stream into a sequence of cluster identifiers (i.e., ten cluster labels) and use the transformed stream in this frame-level dimension for further analysis of complex activities.

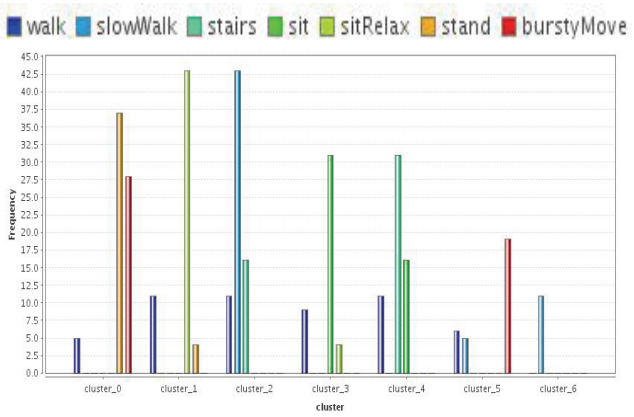


Figure 4: Locomotion tags vs. clusters

In addition, we test the clustering algorithm with the user-tagged micro-activity data with 7 tags (i.e., walk, slowWalk, stairs, sit, sitRelax, stand, bustyMove). As shown in Fig. 4 that plots the first 7 clusters (from *cluster*0 to *cluster*6) for one user, we observe the clusters are not necessarily consistent with the user-tags. We have similar observations for other four users, and all have such difference between user-tags and clusters.

5. COMPLEX ACTIVITY RECOGNITION

In this section, we exhaustively study different higher-order features as well as the learning methods, for inferring the complex activities in daily life.

5.1 Activity Features

Each complex activity (CA) instance is first transformed into a sequence of labels (see Definition 3), as a result of "accelerometer frame labeling", described in Section 4. Therefore, $mathitCA = \{l_1, \cdots, l_n\}$, where $l_i \in \mathcal{L}_A$ represents a locomotion tag label or a cluster identifier. It is worth noting that we do not know exactly to which specific motion or posture the cluster label corresponds. Based on this sequence of frame labels, we can calculate derived features.

5.1.1 Set and Sequence Features

The number of labels are finite (i.e., the size of $\mathcal{L}_A$). Given $|\mathcal{L}_A| = K$, we can compute a feature vector of size K from each CA instance. We call these "*Set Features*". To simply the explanation, let's assume that there are three clusters $\mathcal{L}_A = \{a, b, c\}$, and a CA instance has the sequence {acbbbaaaccbb} after frame labeling. Therefore, the set feature vector is $\langle 4, 5, 3\rangle$, indicating the number of a, b, c in this CA instance, respectively.

In addition to the set features that only focus on each individual label independently, the sequential features account for joint occurrences. In the previous example, we have the following sub-sequences: $\langle aa, ab, ba, ac, ca, bb, bc, cb, cc\rangle$. Thus, the sequential feature vector for {acbbbaaaccbb} is $\langle 2, 0, 1, 2, 0, 3, 0, 3, 1\rangle$ if we use an overlapping sliding window of size 2 that shifts by 1 step, and $\langle 1, 0, 1, 1, 0, 2, 0, 0, 1\rangle$ if we use non-overlapping sliding windows. Longer subsequence features like length-3 (aab, aba, $\cdots$) can be calculated similarly. The problem here is the exponential nature of the feature dimension. Therefore, discriminant sequential mining has been significantly studied in the literature to identify informative sub-sequences [5, 13, 23]. In [23], we identified that sequential features do not have significantly contributions compared to the set features for inferring semantic activities in the wild. Therefore, in this paper, we mainly concentrate on other higher-order feature spaces using advanced models. In particular, we investigate regression features and topic features. We discuss these next.

5.1.2 Regression Features

These features are based on the idea of building trend models to build discriminative features of complex activities and further make classification. The main intuition here is to learn the evolution of a certain CA in the underlying feature space K as the activity progresses in time. In our case, K indicates the number of clusters. We use the features from the activity evolution models to build classifiers. For testing an unknown CA instance, the evolution model of the test stream is compared with known models to predict the CA. The intuition is explained in Fig. 5.

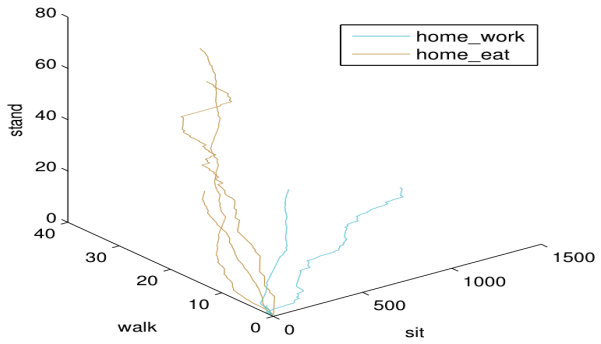


Figure 5: Regression features for activity learning

In Fig. 5, we observe the temporal evolution of five CA instances, with the feature space containing 3 locomotion labels (i.e., 'sit', 'stand', 'walk'). Note the trend difference between two activity types, i.e. "home_work" and "home_eat". Therefore, we can build the regression models of these CA instances, and use the regression models to infer CAs. For the example of 3 locomotion labels ('sit', 'stand', 'walk'), we can build a linear regression model (based on three variables) for an activity class:

$$x_1 a_1 + x_2 a_2 + x_3 a_3 + x_0 = 0$$

where a_1, a_2, a_3 indicate the number of 'sit', 'walk', 'stand' frames in each CA instance. Thereafter, the coefficients $\langle x_0, x_1, x_2, x_3\rangle$ can be used to train classification models. We

call them "*Regression Features*", and build *CA* classification models using them. Note that, we can also build higher-order polynomial regression features by fitting higher-order polynomials to the activity evolution.

5.1.3 Topic Features

Topic models are well-known approaches in discovering the abstract "topics" for a collection of documents, based on analyzing the statistical properties of the "bag of words" in the document. For implementing topic models, Latent Dirichlet allocation (LDA) is one of the most well-known instantiations of topic model [18, 3]. Topic models have successfully been utilized in several applications including activity recognition [10, 2]. In activity recognition literature, researchers have typically applied topic models over multi-sensor streams, typically in constrained wearable computing setups, where rich observation data about the activity is available. Our research has few unexplored angles: (1) We investigate a method to build latent topics from a single sensor stream, by transforming the stream to an observation space representing data clusters that present in the stream; (2) We use the emerging topic associations in turn as features, to build supervised models for studying complex activity recognition accuracy; (3) Discriminatory power of such higher order features has not been evaluated on complex activities measured in the wild.

In our problem, a topic t is a hidden latent activity. Elements in a given *CA* instance (i.e., the sequence $\langle l_1, \ldots, l_n \rangle$) are probabilistic occurrences due to $T(>=1)$ hidden latent activities (act_{topic}). A complex activity *CA* is a statistical combination of a group of such latent activities or topics $\langle act_{topic_1}, \ldots, act_{topic_j} \rangle$, and each latent activity act_{topic} is associated with a certain probability. The intuition is, from the accelerometer observations, *CA*s such as *office_break* can consist of 'getting off the chair', 'moving around', and 'sitting down on the chair' etc, *office_lunch* can include 'carrying the tray to the desk', 'sitting down', and 'eating' etc. Our aim is to investigate the power of the resulting topic distributions (as features) to classify the *CA*.

To do this, we apply LDA to infer the latent topics (act_{topic}) from the *CA* sequences representing cluster labels (tag_{frame}). Then, based on the act_{topic} distributions for each complex activity ($act_{complex}$), the actual *CA* label for each instance is to be inferred. Therefore, we have the following inference for T latent topics.

$$p(act_{complex}|tag_{frame}) = \sum_{k=1}^{T} p(act_{topic}|tag_{frame}) p(act_{complex}|act_{topic}) \quad (1)$$

Once we have learned the LDA model, we can use this model to estimate the distribution of latent activities act_{topic} in a given *CA* instance, i.e., the estimate of extent to which different latent topics are present in the *CA*. Therefore, for each training *CA* instance, we have a vector $\langle p_1 \cdots p_T \rangle$, representing the distribution of the T latent topics in the *CA* instance. This vector represents the "*Topic Features*".

5.2 Learning Algorithms

For learning classification models, we explore the support vector machine (SVM) method. In particular, we apply the LibSVM package which has a good performance in many classification applications [4]. Fig. 9 captures the overall procedure for our complex activity learning and classification. The feature dimensions (*Set, Regression, Topic*) are essentially representing information about a *CA* instance in different dimensions. We explore two vector machine variants for learning classification models: *early* fusion and *late* fusion.

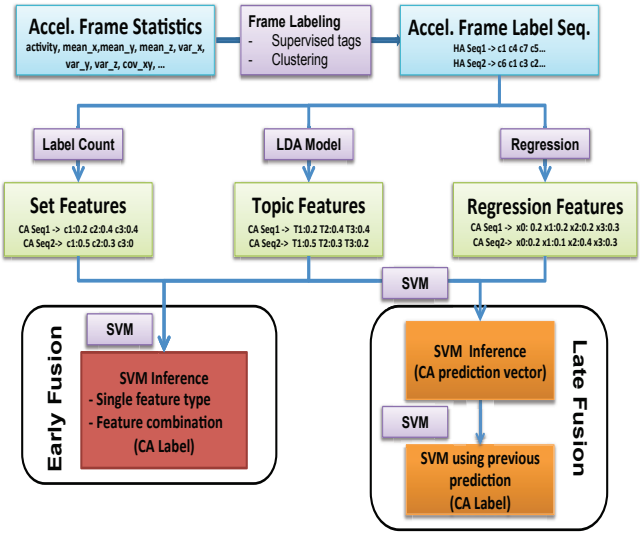


Figure 6: Learning procedure (Early fusion vs. Late fusion)

- *Early Fusion* - In early fusion, we directly build classifier (e.g., SVM model) on the *CA* features, and make prediction. Here, the *CA* features can be a single type of features (i.e, set, regression and topic); or a combination of these different types of features to form a single, master feature vector. From training data, these feature vectors are used to build a single SVM classification model. Intuitively, this approach learns the separation vector on the feature space, combining the multiple dimensions and can exploit correlations amongst features in the mixed feature space.
- *Late Fusion* - This approach exploits the discriminatory power of individual feature dimensions for classification, while compromising on the potential correlations present in the mixed feature space combining all feature extractors. Here, we learn *individual* SVM models for each feature dimension. Each SVM model outputs a prediction vector $[p_1, \ldots, p_n]$ for each *CA* instance, where p_i = predicted probability of the instance to belong to class i. Afterwards, the third vector plane is subsequently learnt using the prediction vectors of different *individual* SVMs, to predict the final class value of an unknown *CA* instance. Intuitively, this model accommodates variation in the predictive powers of multiple feature dimensions, and performs well when feature spaces are not necessarily correlated.

6. EXPERIMENT

To evaluate our approach on mining complex activities in the wild using a single smartphone accelerometer sensor, we provide a group of extensive experiments. The experiments are carried out using the data set described in Section 2.

6.1 Stream labeling: supervised vs unsupervised

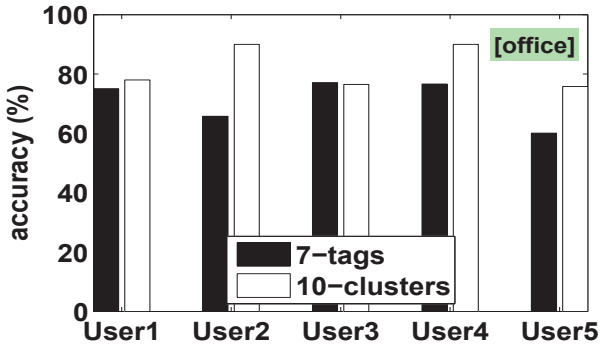
We first test the difference between labeling of the accelerometer stream with a supervised set of locomotion and posture labels, vs. unsupervised data-cluster based labeling of the same stream.

We collected ground truth data about 7 locomotion and postures of the same 5 subjects, with the phone in different preferred body positions. This was done via an exit interview after the data collection. The 7 labels are: 'sit', 'sit active', 'walk', 'loiter', 'bursty move', 'stand', 'using stairs'. These were chosen based on users' feedback of micro-activities commonly associated with their daily lifestyles at home and office. While most labels are self-descriptive, the non-obvious ones are described in Table 3. Classification models were trained and the accelerometer stream corresponding to the complex activities were labeled subsequently. This was compared against data-cluster based labeling of the streams, , obtained by using K-means algorithm with the number of clusters K=10 (Section 4).

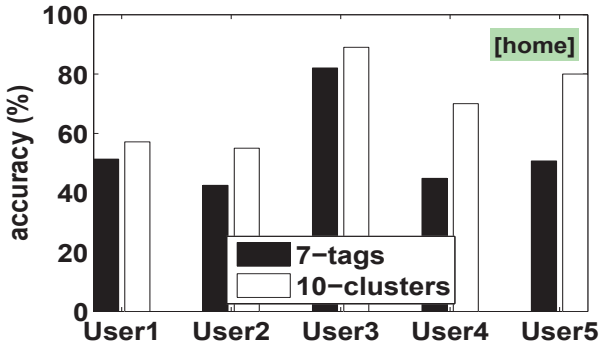
Table 3: Descriptions of some non-obvious activity labels

Labels	*Exemplary Description*
sitActive	sitting but being active (e.g., shaking legs, stretching, ..)
sit	sitting in a static way
loiter	walk at a slow pace with stops, walk inside office rooms
burstyMove	jerky movements (e.g., get up from chair, movements inside kitchen)

Fig. 7 compares the complex activity classification accuracy for five users with set features, after the stream was transformed with 7 supervised labels vs. 10 cluster-based labels. We observe that, for both home and office activities, we see performance improvements (up to 20%) using the cluster-based labels. The variation is non-uniform across users. For *User3*, we do not notice much improvement. We conjecture this is because the supervised labels perhaps captured most of the ambulatory patterns of that user, and cluster-driven labeling did not add much extra information. For subsequent experiments, we use the cluster-based labels as they are performing better to capture representative activity structures.



(a) office activities



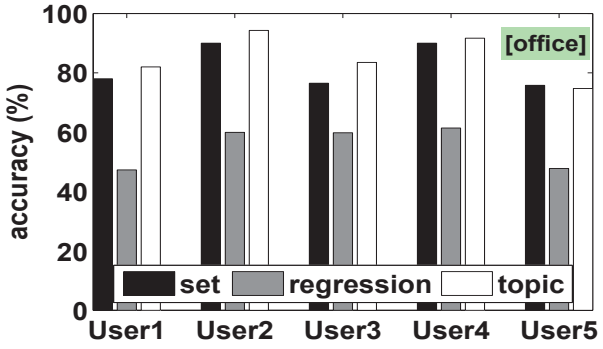
(b) home activities

Figure 7: Comparison of supervised labels vs. unsupervised data-cluster based labels, for complex activity recognition accuracy. (*Set* Features are used for this comparison.)

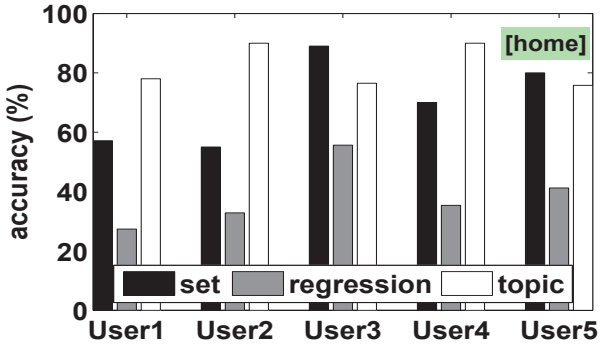
6.2 Higher Order Features

We build the second set of experiments to evaluate different higher-order features. Prior work [23] has evaluated discriminatory sequences as features, and no large improvement was found over *Set* features. Moreover, *Set* features have much less dimensionality than sequences. Hence, we compare *Set*, *Regression* and *Topic* features (Section 5.1) here.

Fig. 8 provides the comparison amongst the three feature types. We report poor performance in recognizing both home and office activities using *Regression* features. Such regression features are computed using simple linear regression models fitted to the activity evolution. Further, we also evaluated higher-order regression functions (e.g., quadratic regression). This resulted in some performance improvement compared to linear regression, but performed worse than Set features. Between *Set* and *Topic* features, there does not seem to be a clear winner, though *Topic* features perform better in most cases (7 out of 10), with *User*2's home activities recording the highest gain ($\geq$25%) using *Topic* features. Overall, the average percentage improvement achieved using *Topic* features surpasses the *Set* features.



(a) office activities



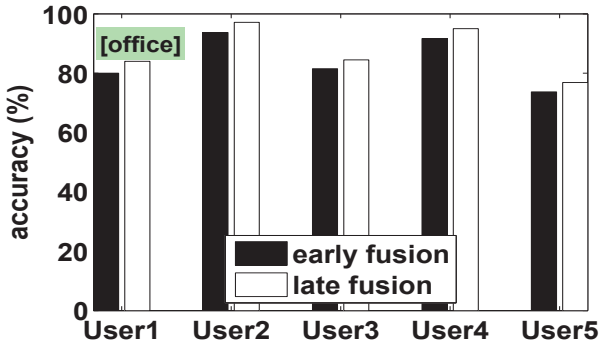
(b) home activities

Figure 8: Performance comparison of *Set*, *Regression* and *Topic* features. (10-clusters are used for this comparison.)

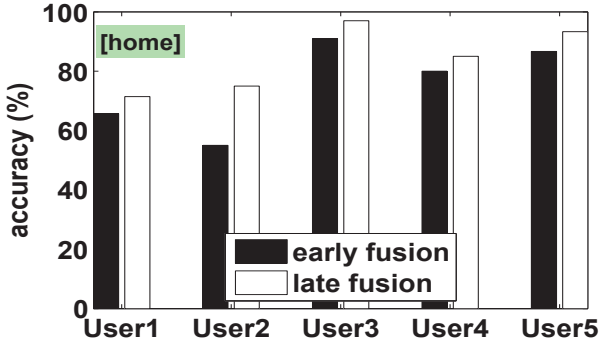
6.3 Learning: Early Fusion vs. Late Fusion

Finally, we evaluate and compare the two learning approaches, i.e., early fusion and later fusion, using the combination of multiple features. As *Regression* features cannot achieve good performance with the reality data, we mainly test fusion using the combination of *Set* features and *Topic* features.

As shown in Fig. 9, we observe that all home and office activities of five users can achieve good recognition accuracy using late fusion. The fact that late fusion works better indicates that perhaps not much correlation is present in the two feature dimensions (*Set*, *Topic*), for exploitation by the early fusion approach. The final accuracies computed by late fusion for home and office, are between 85%-95%, as compared to ≈65%-80% reported earlier [23]. The highest accuracy we achieved is 97%, which is about 10% better than the highest reported earlier. The average accuracy of 86.17% is also about 10% better than prior work.



(a) office activities



(b) home activities

Figure 9: Comparison of early fusion and later fusion

To further analyze the results from late fusion, Table 4 provides the confusion matrices of both home and office activities of 5 users. We observe that most of office activities can be more easily detected compared to home activities. In particular, all *office_work* complex activities from 5 users have an impressive ≥90% accuracy. For home activities, we observe the classifier gets confused for *home_relax* activity for $User1$. In our exit interview with $User1$, we subjectively found that $User1$ often performed this activity in ways that intuitively should look similar, when observed solely by the accelerometer. Since the tagging process is unconstrained, it is quite possible that some activities are similar in the observation space of an accelerometer. Additional sensory context (e.g., microphones) might be needed to obtain enough discriminatory features to disambiguate among such complex activities, which is out of the scope of this paper.

Our results provide evidence that the accelerometer is a powerful sensor to exploit, for complex activity recognition in the wild, with completely real-life scenarios. The late fusion approach shows promise for further evaluation of complex activities in naturalized environments.

	Activity	No.	Confusion Matrix					Acc
USER1	home_work	36	**0.88**	0.0	0.04	0.0	0.08	**71.4**
	home_break	14	0.0	0.4	0.2	0.4	0.0	**%**
	home_relax	25	0.066	0.0	0.667	0.133	0.133	
	home_cook	21	0.0	0.0	0.2	**0.8**	0.0	
	home_eat	17	0.2	0.1	0.2	0.0	0.5	
	office_break	17	**0.9**	0.0	0.0	0.1		**84.0**
	office_work	30	0.0	**0.96**	0.04	0.0		**%**
	office_meet	15	0.1	0.8	0.1	0.0		
	office_lunch	11	0.0	0.0	0.0	**1.0**		
USER 2	home_relax	21	**1.0**	0.0	0.0	0.0		**75.0**
	home_work	9	0.2	0.6	0.0	0.2		**%**
	home_baby	9	0.4	0.2	0.4	0.0		
	home_eat	12	0.1	0.1	0.0	**0.8**		
	office_work	80	**0.98**	0.2	0.0	0.0		**97.1**
	office_ toilet	33	0.067	**0.866**	0.067	0.0		**%**
	office_lunch	19	0.1	0.0	**0.9**	0.0		
	office_meet	20	0.4	0.6	0.5	**0.85**		
USER 3	home_relax	57	**0.98**	0.02	0.0	0.0		**97.0**
	home_cook	23	0.05	**0.9**	0.05	0.0		**%**
	home_eat	28	0.0	0.4	**0.96**	0.0		
	home_clean	11	0.0	0.2	0.0	**0.8**		
	office_work	122	**0.954**	0.09	0.181	0.0	0.181	**84.5**
	office_lunch	32	0.167	0.7	0.0	0.033	0.1	**%**
	office_coffee	35	0.2	0.04	0.56	0.16	0.04	
	office_ toilet	33	0.067	0.0	0.066	**0.867**	0.0	
	office_break	15	0.2	0.0	0.6	0.0	0.2	
USER 4	home_relax	15	**0.813**	0.062	0.125			**85.0**
	home_work	7	0.0	**1.0**	0.0			**%**
	home_cook	18	0.1	0.0	**0.9**			
	office_meet	11	0.6	0.2	0.2			**95.0**
	office_work	59	0.022	**0.978**	0.0			**%**
	office_break	41	0.0	0.1	**0.9**			
USER 5	home_cook	20	**0.95**	0.0	0.5			**93.3**
	home_work	6	0.333	0.667	0.0			**%**
	home_relax	11	0.0	0.0	**1.0**			
	office_work	65	**0.927**	0.018	0.018	0.036		**76.8**
	office_meet	11	0.0	0.667	0.333	0.0		**%**
	office_lunch	23	0.733	0.0	0.067	0.2		
	office_break	45	0.16	0.0	0.04	**0.8**		

Table 4: Confusion matrix of late fusion results

7. RELATED WORK

Early works on human activity learning (e.g., [1]) focused on activities such as walking, running, cycling (which we call MA, i.e, micro-activities) using data from *multiple* body-worn accelerometers under lab environments. Thereafter, several investigations have measured activity detection accuracies using accelerometers placed in different on-body position [8, 16] for different activity routines (e.g., gym activities). Recently, [14, 11, 22] addressed the problem of locomotion and posture prediction (i.e., MA classification) using cellphone-embedded accelerometers. Micro-activities exhibit recurring behavior over short time windows (seconds), making it possible to classify them with high confidence using $O(sec)$ test windows. However, complex activities typically extend over longer periods of time and consist of multiple types of unpredictable locomotions in the observation window. In this paper, we attempt to classify such high-level complex activities (CA), which have aperiodic behavior.

Approaches for mobile phone-based activity recognition principally aim to leverage upon multiple phone-embedded sensors to recognize the user's context [7, 14, 21, 15]. For example, [6] demonstrated that accelerometers and micro-

phones provided good features for activity recognition (e.g., walk, run, talk, cook, eat) while [14] developed an 'on-phone, continuous' recognition system to detect events using multiple phone sensors. The primary focus is to discover how a combination of phone sensors helps to improve the accuracy of activity detection. We however investigate a complementary question: *To what extent can the accelerometer sensor alone be used discriminate indoor complex activities?*

Exploitation of hierarchical approaches in recognizing complex activity is not new in activity recognition literature. [10] presented an approach towards automatic discovery of complex activities, based on the use of topic models to discover repetitive, aperiodic patterns of clustering among the underlying low-level activities. Such a hierarchical approach was also explored in [2], which employed a multi-layer discriminative approach using conditional random fields for detecting composite activities. Similar models for higher-layer activity detection have also been explored in the area of smart homes (e.g., [19]). All these approaches have relied on multiple body-worn sensors and infrastructure-based sensors (e.g. object interaction). In contrast, our investigation looks at activities in the wild, i.e., un-instrumented indoor spaces, using a single phone-based accelerometer, carried by the user in out-of-lab environments. We conduct a first-of-a-kind, thorough investigation of higher-order features for complex activity recognition under completely naturalized settings. Our results show that unsupervised learning of data clusters in the streams, and their subsequent usage to build higher-order features, show good results for complex activity classification. Our results of achievable accuracies indicate that many complex activities may be deciphered using only the accelerometer.

8. CONCLUSION

This paper investigated complex activity recognition performance using a single smartphone based accelerometer sensor in the wild. The primary focus was to investigate recognition performance of complex activities under naturalized settings, using a real-life data set collected from 5 users over a span of 8 weeks. Along the lines of hierarchical feature dimensions, we investigated several higher order features (e.g., set, regression, topic features), that can be extracted from the raw sensor stream.

While researchers have shown good performance of mining complex activities in controlled & multi-sensor settings, our paper makes the following new insight for smartphone-based complex activity recognition in real-life settings: *Features constructed using generative models, over a representation of the stream as a sequence of cluster labels, captures complex activity signatures very well, resulting in high recognition accuracy.* This also indicates that energy-friendly smartphone-based accelerometers can be a good choice to sense many complex activities in real-life.

9. REFERENCES

[1] L. Bao and S. 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] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[5] H. Cheng, X. Yan, J. Han, and C. Hsu. Discriminative Frequent Pattern Analysis for Effective Classification. In *ICDE*, pages 716–725, 2007.

[6] T. Choudhury, G. Borriello, S. Consolvo, D. Hähnel, B. L. Harrison, B. Hemingway, J. Hightower, P. V. Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. D. Rea, and D. Wyatt. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, 7(2):32–41, 2008.

[7] D. Choujaa and N. Dulay. Predicting Human Behaviour from Selected Mobile Phone Data Points. In *Ubicomp*, pages 105–108, 2010.

[8] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu. epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition. In *PerCom*, pages 1–9, 2009.

[9] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.

[10] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Ubicomp*, pages 10–19, 2008.

[11] J. R. Kwapisz, G. M. Weiss, and S. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explorations*, 12(2):74–82, 2010.

[12] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity Recognition using Cell Phone Accelerometers. In *Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data*, pages 10–18, 2010.

[13] J.-G. Lee, J. Han, X. Li, and H. Cheng. Mining Discriminative Patterns for Classifying Trajectories on Road Networks. volume 23, pages 713–726, 2011.

[14] H. Lu, Yang, J. Liu, N. Lane, T. Choudhury, and A. Campbell. The Jigsaw Continuous Sensing Engine for Mobile Phone Applications. In *Sensys*, pages 71–84, 2010.

[15] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. Eisenman, X. Zheng, and A. Campbell. Sensing Meets Mobile Social Networks: the Design, Implementation and Evaluation of the Cenceme Application. In *Sensys*, pages 337–350, 2008.

[16] M. Muehlbauer, G. Bahle, and P. Lukowicz. What Can an Arm Holster Worn Smart Phone Do for Activity Recognition? In *ISWC*, pages 79–82, 2011.

[17] A. Rowe, A. Smailagic, and D. Siewiorek. ewatch: a wearable sensor and notification platform. In *Workshop on Wearable and Implantable Body Sensor Networks*, 2006.

[18] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[19] E. Tapia, S. Intille, and K. Larson. Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors. In *Pervasive*, pages 158–175, 2004.

[20] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod. Accurate, low-energy trajectory mapping for mobile devices. In *NSDI*, pages 20–20, 2011.

[21] Y. Wang, J. Lin, M. Annavaram, Q. A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh. A framework of energy efficient mobile sensing for automatic user state recognition. In *MobiSys*, pages 179–192, 2009.

[22] G. M. Weiss and J. W. Lockhart. The Impact of Personalization on Smartphone-Based Activity Recognition. In *AAAI Workshop on Activity Context Representation: Techniques and Languages*, 2012.

[23] Z. Yan, D. Chakraborty, A. Misra, H. Jeung, and K. Aberer. SAMMPLE: Detecting Semantic Indoor Activities in Practical Settings using Locomotive Signatures. In *ISWC*, pages 37–40, 2012.

[24] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer. Energy-Efficient Continuous Activity Recognition on Mobile Phones: An Activity-Adaptive Approach. In *ISWC*, pages 17–24, 2012.