Using Time Use with Mobile Sensor Data: A Road to Practical Mobile Activity Recognition?

Marko Borazio Kristof Van Laerhoven

Embedded Sensing Systems
Technische Universität Darmstadt, Germany
{borazio,kristof}@ess.tu-darmstadt.de

ABSTRACT

Having mobile devices that are capable of finding out what activity the user is doing, has been suggested as an attractive way to alleviate interaction with these platforms, and has been identified as a promising instrument in for instance medical monitoring. Although results of preliminary studies are promising, researchers tend to use high sampling rates in order to obtain adequate recognition rates with a variety of sensors. What is not fully examined yet, are ways to integrate into this the information that does not come from sensors, but lies in vast data bases such as time use surveys. We examine using such statistical information combined with mobile acceleration data to determine 11 activities. We show how sensor and time survey information can be merged, and we evaluate our approach on continuous day-and-night activity data from 17 different users over 14 days each, resulting in a data set of 228 days. We conclude with a series of observations, including the types of activities for which the use of statistical data has particular benefits.

Categories and Subject Descriptors

H.1.m [**Models and Principles**]: Miscellaneous

General Terms

Measurement; performance

Keywords

Time use surveys; activity recognition; wearable computing; mobile devices; probability model

1. INTRODUCTION

As we carry our mobile devices with us while performing our daily activities and tasks, having a mobile device that is aware of our activities would result in a variety of benefits for its user. It has been suggested that interaction with such activity-aware mobiles could be alleviated since the devices know more about their users' status [20]. Tracking the user's activities over longer periods with a mobile device is in particular interesting in various healthcare scenarios [15, 28]: Many mobile and wearable systems[1] that monitor the user's movements and fitness activities [2] are commercially available. Mobile phone platforms are being used as hubs to gather such data or collect information from onboard sensors to determine what the user is doing [5, 24]. Recognition performance for activities like walking or sitting from mobile sensors are already promising [3, 10, 12]. More diverse activities, however, such as having lunch or vacuum cleaning, are less pronounced: Their complex nature requires not just performant machine learning techniques and considerable amounts of training data, but also person-specific information. Especially in battery-driven and resource-limited mobile platforms, an additional problem is that data acquisition comes with a substantial cost.

MUM '13, December 02 - 05 2013, Lulea, Sweden.

http://dx.doi.org/10.1145/2541831.2541850

For this purpose, we looked into promising research [16] that investigated the use of prior knowledge on time use to improve activity recognition on mobile and ubiquitous systems. Additional information, such as an approximation of the user's typical schedule, could be combined with real-time sensor data from a mobile phone. Time use information is already available through national time use surveys, which are statistical surveys that contain information from thousands of individuals on where and when people are performing which activities on average. Recent work [4] used such data bases to infer the user's activity based on the activities of similar users within the survey. Results have thus far shown that time survey data, when used from within the same geographical and cultural region as the user, could enhance activity recognition systems considerably.

In this work, we investigate a method that makes use of prior knowledge like the time use data for activity recognition with a mobile device, concurrently to mobile sensors. We will have a look at a common classifier like the Support Vector Machine (SVM). We show which features from time use databases are important to consider and how our system outperforms common models, resulting in high precision and recall values for activities like sleeping and working. For this purpose, we obtained inertial data from 17 subjects over a recording period of 2 weeks each, resulting in a dataset of 228 days in total. We collected data from 11 activities that are usually performed by the users, obtaining the ground truth by keeping a diary. Results indicate that certain activities can profit from prior information such as time use surveys.

[1] such as www.actigraphcorp.com/, last access 10/2013

The remainder of this paper is structured as follows: First, in Section 2, we will give some insight into work related to our research, after which we describe in Section 3 the methods used to determine the activities of the test subjects, and describing the classifiers in Section 4. In Section 5, we show the results obtained from three different classification modalities. A short discussion about the experiment's findings will be held in Section 6. We conclude this work in Section 7, giving also an outlook to future studies.

2. RELATED WORK

Embedding additional information to improve the recognition rates has been done in several studies [13, 14, 22]. Usually, information about the environment is being used, like location or even time. For several years now, countries are gathering data about the population, which might also give some insight into the habits of the inhabitants. We will first have a look at the most important parts of this work, mobile activity recognition and time use studies, and will then continue on how prior information has been used in the area of wearable and mobile computing, concluding this section with a summarize of ensemble classifiers.

2.1 Mobile Activity Recognition

Activity recognition with mobile sensors has been investigated for some years now [5, 24], with researchers also analysing if it is feasible to use a mobile device for detecting activities [17]. In [5], basic human movements (walking, sitting, standing, climbing stairs) are detected in real-time on a mobile phone by analysing the accelerometer data with a high confidence. Similar results are obtained by researchers in [24], again detecting basic activities on a mobile phone, but also considering the orientation of the device. The performance of the system is high, stating that basic activities can be captured with a mobile device. Patel et al. [17] on the other hand investigated for how much portion of the day a mobile device (smartphone) is with the user. Interestingly, results indicate that half of the time the mobile phone is not with the user.

The idea of using mobile phones not only for real-time activity recognition, but also for gathering useful information about the user and the activities, has been investigated in [8], in order to develop rhythmic data that can be used to detect daily routines. We will have a look at such data in the following section, which has not been gathered by mobile devices, but by other means.

2.2 Time Use Surveys

Time use surveys (TUS) are inquiries performed by a countries government, asking participants to keep a diary of their activities over a period of one to three days. Depending on the region, time use surveys are being updated in different intervals. In this study we used the German Time Use Survey (GTUS) from 2001/2002, which is being updated every 10 years. The latest version from 2011/2012 will be available for researchers in 2015. In the GTUS, 13,798 participants were keeping a diary to log in 10 minute intervals what activity they performed. Additionally, location information of where the activity took place or what secondary activity has been performed (e.g. talking to friends while having dinner or watching TV while eating) were noted. Also logged are interrelations between household members, especially when activities were carried out in company with other household members.

Recently, researchers in [4] investigated the GTUS in regard to benefits for activity recognition, identifying features in the dataset that can be used to determine the activities that occurred within the time use dataset. They extracted activity histograms according to the investigated features, e.g., *time* or *location*. From the histograms probabilities are then calculated to infer the most likely occurring activity within the time use survey.

Similar to [4], the work in [16] investigates the American Time Use Survey (ATUS) and identifies time use surveys as a promising instrument for designing activity recognition systems. The ATUS can be freely obtained online[2] for further studies, while the GTUS is available for regional government employees only. More details about time use surveys in general with supplement information can be found online[3] or in [18].

2.3 Using Prior Probabilities

Predictions have been used in various scenarios for activity recognition [9, 11, 21, 27, 31]. Recently, researchers in [27] predicted a person's going-out behaviour in order to asses if the person is going to leave the home or not. A rhythm of a person's habits has been established by observing with a camera when a person is usually at home or not. This information was then used as a prior for a Hidden Markov Model (HMM). Here, time histories of people leaving or entering the home have been generated. Especially in healthcare such scenarios have to be considered, when for example elderly people living on their own are monitored in order to be able to respond to emergencies. Daily routines can help here by predicting what the user will be doing most likely. A similar idea is being pursued in [11], by gathering prior information by keeping a diary and additionally obtaining GPS information about a person being at home. The paper shows how the prior location improves the prediction.

In [9] on the other hand, individual and group behaviours have been investigated in a large mobile phone dataset. A probabilistic topic model has been used on the data, detecting routines in the data to determine behavioural patterns. Overall, the research community is interested in what people are doing next, predicting the behaviour not only from the same persons, but also from various individuals. Such information could then be used in different models that deal with sequential data where such prior probabilities or even posterior probabilities might improve the classification results. An example of such a classifier is the Conditional Random Field (CRF) [29], which is a temporal probabilistic model.

2.4 Ensemble Classifiers

The idea of fusing two or more classifiers to improve the recognition rate for activity recognition has been mentioned in several works [1, 19, 23, 26]. Zappi et al. [30] for example use multiple body-worn sensors for activity recognition in the context of quality assurance in a car assembly factory.

[2]http://www.bls.gov/tus/, last access 10/2013
[3]http://www.timeuse.org/, last access 10/2013

Using a discrete HMM, the results led to an improvement in the recognition rate. In [19], different ways of classifier fusion or ensemble classifiers are being discussed, like mixture of experts, bagging, boosting or algebraic combination rules. The latter are usually majority voting, sum and product rule. Researchers in [1] evaluated the two common combination rules (the mean and product rule) for fusing classifiers by their posterior probabilities, which will be used later on in Section 4.3.

In this study, the data we obtain is not high frequent, on the one hand because the battery consumption on a mobile platform can be reduced this way, on the other hand to reduce memory usage. We will show that low frequent data is sufficient for efficient activity recognition by embedding probabilities in the classification process. Researchers in [16] already mentioned that ubiquitous systems might benefit from time use data and that such data should be used in the process of developing mobile platforms. This way, information not only from the participants of this study is taken into consideration, but also from time use surveys, which represent a countries inhabitants. We show in the following sections how time use survey data is being used to evaluate data obtained from a mobile system.

3. METHODOLOGY

For recording inertial data we decided to use a system that is already commercially available, is comfortable to wear over a long time-span, but also enables researchers to get direct access to the sensor data. We will first introduce the sensor used for this study and then explain how we performed activity recognition with a common classifier on sensor data only and with time use survey data.

3.1 Mobile Sensors

For this study, we used the SenseWear Armband shown in Figure 1 from BodyMedia[4]. The SenseWear is used to monitor one's activity, especially during workouts and while resting, e.g., sleeping [25]. It is worn comfortably on the upper arm, as shown in Figure 1, and can rest there continuously day and night. Additionally, it simulates data that could be obtained from a mobile system which are the main advantages to use this device. A graphical tool displays useful information to the user, like showing how data channels change during time. Additional information, such as how much activity has been performed or a step counter are also accessible, enabling the user to keep also track of his fitness status. The device is splash waterproof, which is why it can be used during work-outs.

The SenseWear Armband embeds a 2-axis accelerometer, a skin temperature, a galvanic skin response and a heat flux sensor. Sensor values can be stored in the internal storage in different intervals, from 32 samples per minute up to one sample every 10 minutes. Depending on the log frequency, the storage lasts for 2 hours only (32 samples per minute) or a little more than two weeks (one sample per minute). The power source is a common AAA battery, which needs to be replaced by the user after approximately one week, depending on how often the sensor was worn and how much the user moved. The sensor automatically starts logging when skin contact is being detected and stops logging as soon as the user takes off the unit. The recording frequency for our study was set to one minute, being the optimal trade-off for recording for a long time-span and not loosing too much sensor information. Had we increased the sampling rate, the device would have run out of memory after a few days only instead of being able to record for 14 days straight.

[4]http://www.bodymedia.com/, last access 10/2013

Figure 1: BodyMedia SenseWear Armband worn by a male (left) and female (right) participant. The Armband is usually worn on the upper arm.

3.2 Dataset Description

The SenseWear Armband was worn day and night by 17 test subjects for 14 days each, resulting in a dataset of approximately 228 days. While wearing the device, the subjects were asked to keep a diary of common activities for the entire recording period, usually recalling at the end of the day what activities they performed. Some subjects kept a diary by writing down the activity immediately after performance. The list of the activities is shown in Table 1. We obtained inertial data such as the average and longitudinal and transversal acceleration, as well as the Mean Absolute Difference (MAD) of the acceleration. A continuous dataset of 14 days for a male participant is displayed in Figure 2, showing in the top plot the MAD and the bottom plot the average of the longitudinal (blue) and transversal (red) acceleration. Each day starts at the numbering of the day along the x-axis. The nights are immediately visible by segments where the acceleration information is low. Additional sensor information like skin temperature has been logged, but were not considered for this study. For the evaluation, we used only the accelerometer data to infer the performed activity. Note here that the activities that have been logged are not equally distributed, especially the amount of personal care events appears rather low with 88 hours in total as can be observed in Table 1. This can be explained by the fact that the sensor was taken off for showering. Similar, sports occurs only 40 times throughout the whole datasets, either because the device was taken off during work-outs or because the participants are not very sportive. Also, when considering eating, this activity occurs very often, but does not take up as much time as sleep for example.

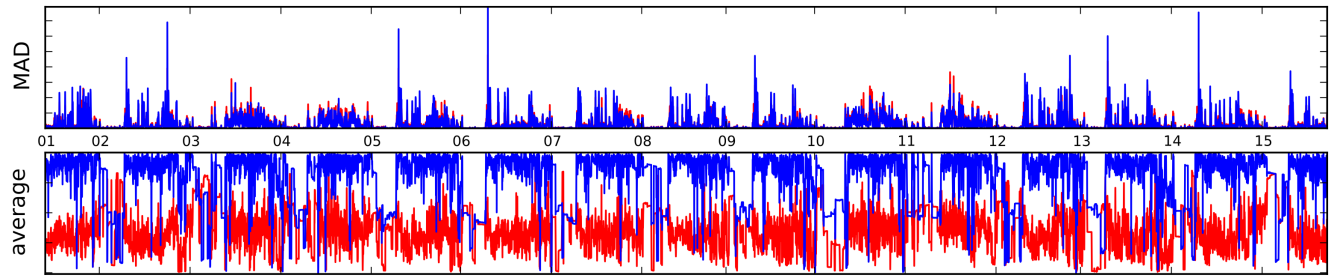


Figure 2: Inertial data from the SenseWear Armband's 2-axis accelerometer from one participant over 14 days before normalization. The top plot displays the Mean Absolut Difference (MAD), while the bottom plot depicts the average of the longitudinal (blue) and transversal (red) acceleration over 20,312 samples, each sample taken every minute.

Table 1: Activities taken from the time use survey that were logged by the test subjects, additionally displaying how often the activity occurred and for how long in total.

ID	activitygroup	occurences	duration [hrs]
1	Sleeping	247	1868
2	Eating	373	257
3	Personal care	223	88
4	Working	297	1047
5	Studying	67	215
6	Household work	215	284
7	Socializing	111	249
8	Sports	40	41
9	Hobbies	72	172
10	Mass media	205	397
11	Travelling	366	864

Table 2: The test subjects that participated in this study, along with additional information like gender and age. Also displayed is the amount of data obtained from the participant.

subject	gender	age	data [hrs]	comments
1	male	32	338	employee
2	female	28	334	student
3	male	31	333	employee
4	male	27	319	employee
5	male	28	284	employee
6	male	32	334	employee
7	male	30	320	employee
8	male	27	328	student
9	male	28	315	employee
10	female	35	346	housewife
11	female	29	321	employee
12	male	31	340	employee
13	male	26	274	employee
14	female	28	292	employee
15	male	21	316	student
16	male	25	334	student
17	female	48	352	housewife

In order to later on compare our activities to the time use activities, we established the list in Table 1 according to the given activities in the time use survey, with one exception: The activity personal care from the time use data includes sleeping, eating and other activities in the area of personal care, such as showering or dressing. We decided to split up these activities into the first three activities as shown in Table 1, especially to be able to catch sleeping and eating on its own.

We chose to use a high variety of test subjects as summarized in Table 2 to capture different data and especially different activity behaviours according to their profession to stress-test our algorithm. The subjects are between 21 and 48 years old of which 12 male and 5 female. The majority of the participants are common employees, working either at the university or in an office, but also students participated, who are known to have a completely different daily routine than employees. Additionally, we display the amount of data recorded by each participant in Table 2.

3.3 Evaluation Measures

A common measurement for describing the recognition rate is accuracy, which can be derived from a confusion matrix. A confusion matrix is a $n \times n$ matrix, where n is the amount of classes in the classification. The rows of the matrix represent the ground truth, i.e., the actual class, while the columns represent the predicted classes. A benefit of a confusion matrix is that it directly shows if the system is confusing two classes, i.e., commonly mislabelling one as another. Consider the confusion matrix in Figure 3 (left) with $n = 3$. Accuracy is the sum of the diagonal divided by the sum of all occurrences (here: $15/21 = 0.71$).

The overall accuracy is often not enough to reveal particular details of the system's performance. To gain this information, calculating per class performance values, namely precision and recall, are necessary. For this purpose, the confusion matrix of $n > 2$ has to be considered as a two-class matrix, by summing up the rows and columns outside the actual class. We obtain the confusion matrix in Figure 3 (right). *True positives* (TP) are the correctly predicted classes according to the ground truth and the *false positives* (FP) are the wrongly predicted classes in regard to the ground

actual class \ predicted class	$C1$	$C2$	$C3$
$C1$	6	1	1
$C2$	1	5	2
$C3$	0	1	4

actual class \ predicted class	$C1$	$C2'$
$C1$	TP	FP
$C2'$	FN	TN

Figure 3: Left: Example of an confusion matrix with three classes $C1$, $C2$ and $C3$. Right: Confusion matrix shown as a two-class matrix, with classes $C1$ and $C2' = C2\&C3$.

truth. The *false negatives* (FN) on the other hand are all the activities labelled as one class but do actually belong to another class. The *true negatives* (TN) are the sum of all the *true positives* of the other classes, i.e., the sum of all correctly predicted classes. Here, precision is the amount of correct labelled classes (TP) that was labelled as the activity in the ground truth:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (1)$$

From (1) we calculate for class 1 in Figure 3: $6/(6 + (1 + 1)) = 0.75$. Recall is defined as the proportion of the data originally labelled as an activity that was correctly classified as the activity:

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (2)$$

leading to a recall of $6/(6 + 1) = 0.86$ for class 1.

Most activity classes are being recognized by our system (see Section 5), which can be observed in the various confusion matrices we established and also in Figure 6. Nevertheless, when considering classes that have been very poorly or not at all detected, the overall accuracy for recognizing a participants activity is still pretty high, while precision and recall are rather low. In order to depict how well the data describes each class, we will focus on the precision and recall values only.

With a mobile sensing platform like the SenseWear Armband, we are able to record low frequented sensor data over a long time-span. For 17 participants in this study we will investigate how activities can be sensed in a mobile environment, evaluating the results by applying precision and recall on the obtained classifications. We will now have a closer look on the classification techniques and how the data was prepared for classification, explaining how probabilities have been calculated from the mobile sensor approach.

4. CLASSIFICATION

The method proposed in this work can be described in three steps: The inertial data is first being evaluated with a common classifier to determine the activities, after which we use the time use dataset only to infer the activity. Then, the results from the common classifier and the time use probabilities are fused to improve on the results from the first two steps. For Sections 4.1 and 4.2 we evaluate the activities by dividing the dataset into five equal folds per participants dataset, performing a leave-one-fold-out cross-validation per test subject to enable user-specific activity recognition.

4.1 Mobile Sensors Only

For detecting the activities within the sensor data, we used a Support Vector Machine (SVM) [7]. The implementation of the classifier was done completely in Python. For this purpose we used the sklearn[5] package, which embeds an SVM library that is based on LIBSVM [6]. We use a linear SVM, since we are dealing with large datasets and have a multi-class problem at hand. As a strategy, the one-vs-the-rest method is applied, which basically trains a SVM for one class and tests it against the rest of the classes. Before starting the SVM training, we normalize the dataset. Then, we balance the training set by randomly choosing data rows from labelled features and duplicate them to receive a conform dataset with an equal amount of samples per activity. As shown in Table 1, the occurrence of activities is unbalanced, especially sleep and work dominate the datasets. We perform a five-fold cross-validation to estimate the optimal penalty parameter C on a small subset of the training data, which is being used in the training phase of the classification process. After training the SVM, we estimate the classes for the test set. Additionally, we calculate the softmax output for the testing data, receiving a likelihood estimation for the input data. The softmax output is defined as

$$\sigma_{prob} = \frac{1}{1 + e^{-2*d}}\ , \quad (3)$$

where d is the decision function from the SVM, being described by the Support Vectors of the dataset. The SVM outputs for each datapoint $x_1, ..., x_i$ the decision function $f(x_1), ..., f(x_i)$, describing the distance to the calculated hyperplanes of the SVM. The likelihoods will be used later on when fusing the mobile sensors and time use results.

4.2 Time Use Only

The time use classification technique uses a maximum-likelihood estimation for determining which activity took place. For this purpose, we make use of features f within the time use survey, like *time*, *age* and *gender*. In [4], different features have been evaluated for their use in activity recognition, identifying *time* and *location* as useful features. For this study, *location* information was not logged, since the used sensor is not equipped with a sensor like GPS to infer the location at a given time. Having the user log additionally where the activity took place would have increased the effort for keeping a diary. Therefore, we consider *time* combined with other, available information, adding *age* and *gender* as features.

We calculate from the time use dataset a histogram for each of the given features *time*, *age* and *gender* and the 11 activities, obtaining a 4D histogram of the shape $[144, 5, 2, 11]$. The shape corresponds to 144 10-minute time-slots per day, 5 age-groups[6], 2 gender types (male and female) and our 11 activites. According to the 3-tuple $(time, agegroup, gender)$ we sum up all the occurrences for each of the 11 activities. We then calculate the distribution of the 11 activities for each 3-tuple to obtain the probabilities.

[5]http://scikit-learn.org, last access 10/2013

[6]Note here that for downsizing reasons and to obtain a representative histogram, we divide the time use survey into 5 years age groups (20-24, 25-29, 30-34, 35-39, 45-49), according to our participants.

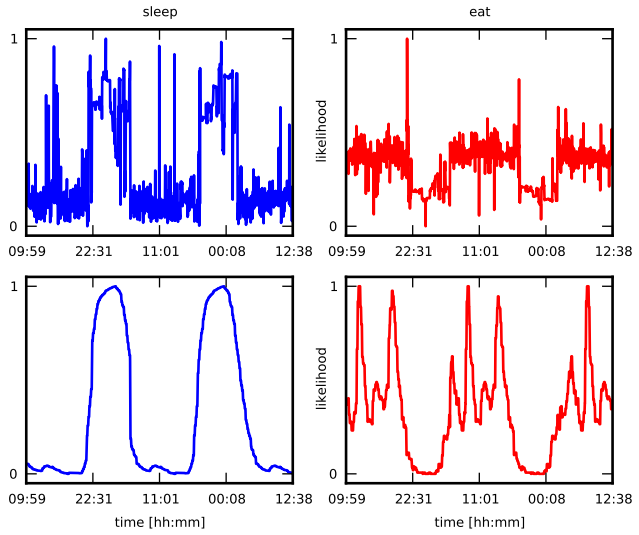


Figure 4: Example of likelihoods estimated from the sensor (top plots) and the time use (bottom plots) approach for a male subject in his thirties, displaying the activities sleeping and eating.

The maximum-likelihood estimation calculates then the probability $P(c_i|f_1, ..., f_n)$ for a target class c_i, $i \epsilon [1, ..., 11]$ and the features $f_1, ..., f_n$, classifying the activity by the highest probability according to the 3-tuple given in the sensor dataset.

4.3 Ensemble: Sensors and Time Use

In this section we describe the combination of likelihoods from the mobile sensors and the time use survey, resulting in a new likelihood-table that is used to determine the activities. The equation

$$c_i = argmax_i \left(\frac{P(c_i|x) + P(c_i|TUS)}{2} \right) \quad (4)$$

describes the procedure of estimating the class c by applying the mean rule [1] on both likelihoods. We scale the likelihoods from both mobile sensors and time use survey for each activity class c_i, i.e., we calculate for all likelihoods of activity c_i the scaling by the equation

$$P_{c_i} = \left(\frac{P_{c_i} - abs(min(P_{c_i}))}{max(P_{c_i}) - abs(min(P_{c_i}))} \right) \quad (5)$$

to avoid the domination of bigger likelihoods over those in smaller numeric ranges. Note, that the likelihoods could be weighted additionally, depending on how the overall classification behaves. The weighting would be applied to equation (4). The overall likelihood $P(c_i)$ would be calculated for the mobile sensors after the training phase and multiplied with each class probability $P(c_i|x)$ in the test set. The same would be done for the time use dataset, obtaining the overall probabilities $P(c_i|TUS)$.

Figure 4 shows the likelihoods for one participant at the age of 32 after scaling of the mobile sensors (top plots) and time use (bottom plots) likelihoods. Displayed are the activities sleeping and eating over approximately two days, showing how likely they are to occur at a specific point in time. The rhythmic nature of the probabilities for the time use data can be observed here, as well as for the likelihoods for the mobile sensors which exhibit much more noise.

Being able to allocate likelihoods to the results of the mobile sensors offers a possibility to combine the likelihoods with the probabilities calculated from the time use survey. This way, we are able to not only consider information from mobile sensor data of the participants, but also additional information from participants of the time use survey, which describe the usual habits of the regional inhabitants. We will now take a look at our results and special findings in the next section.

5. RESULTS

In this work we compare three different classification techniques for the same dataset. We will be discussing the results individually, highlighting the differences for each modality. We will first visually inspect the outcome, then we are going to discuss the results quantitatively.

In Figure 5 we see a qualitative evaluation of all three modalities we used. Displayed are in different colors the activities from a male participant in his early thirties for 14 days, showing from top to bottom: the ground truth, i.e., what the user was actually doing, the estimated activities from the mobile sensors, the time use only and the ensemble of the two modalities. When observing the ground truth, we discover that a certain pattern or even rhythm is visible in the recurrence of the activities throughout the 14 days. Such rhythm could help in the classification process when known. A rhythmic behaviour is also visible in the mobile sensors plot, which is riddled with small detection episodes of different activities. The time use exhibits the most clear rhythmic activity representation by displaying for each day the same activity sequence. The time use classification takes into consideration only the information about *time*, *age* and *gender*. Therefore, we observe here what a male person between 30-34 years is usually doing. Would we imply other features like day-of-the-week, the plot would surely differ. Remarkably, the sensor based approach shows a high variety in the results, except for the activities that seem to take up more time during a 24-hour period, such as sleeping and working. We observe a high variety in the data, because the influence of the mobile sensors is taken into account here. We additionally notice how other activities are being detected more often, as for the ensemble in Figure 5 socializing and eating. We will now summarize the quantitative results, starting with the mobile sensors only.

5.1 Mobile Sensors Only

When using a Support Vector Machine (SVM) classifier to detect the 11 activities from the sensor data, we observe that it is feasible to catch all of the activities but with mostly low recognition rates. Overall, the most confident precision and recall results are obtained for sleep (88.7% and 85.35%) and working (30.3% and 43.68%), which is followed by travel (24.32% and 23.2%) as can be observed in Table 3. The rest of the activities perform rather poorly. A reason for that is that the training data for some activities is not distinguishable, since we balance the sensor values for each class as described in 4.1. Classes with a larger occurrence within the dataset do have an advantage over the other classes.

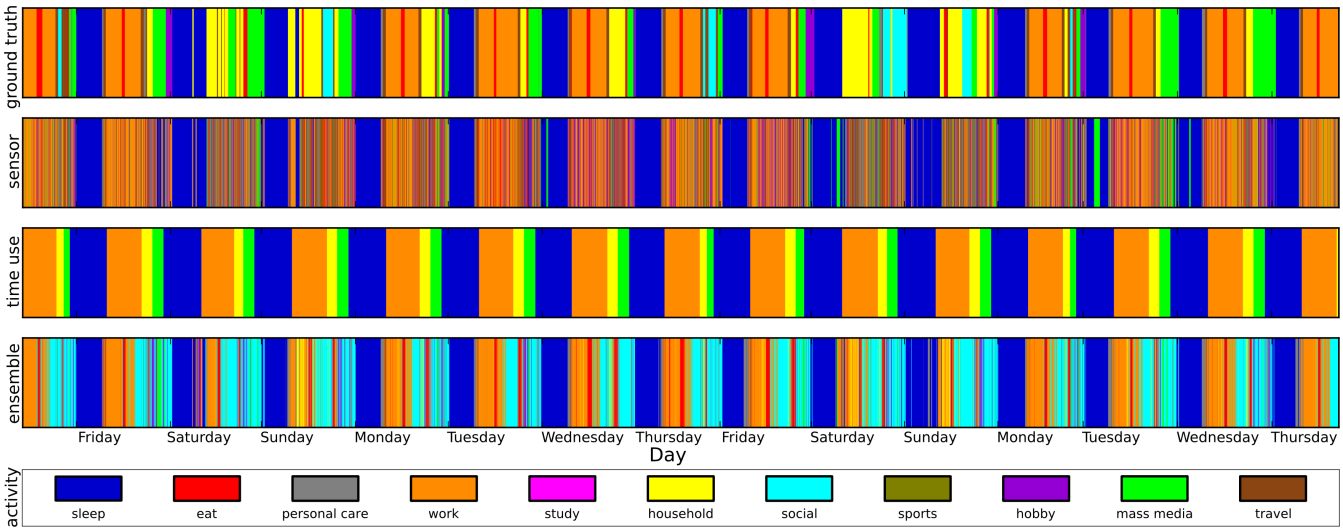


Figure 5: Example of classification results for a male subject (32), showing from top to bottom: the ground truth, estimated activities from the sensor, the time use and the ensemble of both modalities. Displayed are all 14 days from left to right, with barplots showing the activities in different colors. Immediately visible is the improvement when fusing both sensor and time use in the ensemble plot, exhibiting a higher variety of detected activities over the whole dataset.

Table 3: Overall precision (p) and recall (r) results for each activity for all three classification methods, sensor (sen), time use (tus) and ensemble (ens). The highest score for each activity is displayed in bold, showing how the ensemble method exceeds the results from the sensor and time use only based approaches for most of the activities.

activity	p_sen	r_sen	p_tus	r_tus	p_ens	r_ens
sleep	88.59	85.36	**92.87**	82.64	83.22	**95.62**
eat	5.98	13.91	0.0	0.0	**30.45**	**19.76**
pc	13.67	6.29	0.0	0.0	**15.11**	**14.1**
work	30.0	43.65	55.35	46.06	**56.38**	**48.74**
study	3.77	**9.34**	0.0	0.0	**11.5**	9.13
hw	11.34	**17.71**	**34.18**	14.25	7.76	20.87
socialize	12.3	9.43	0.0	0.0	**30.43**	**13.84**
sports	**12.65**	4.03	0.0	0.0	2.39	**7.73**
hobbies	5.11	6.93	0.0	0.0	**11.92**	**9.49**
mm	20.74	14.42	**55.2**	31.06	30.97	**44.53**
travel	**24.31**	23.24	0.1	0.0	6.68	**35.61**

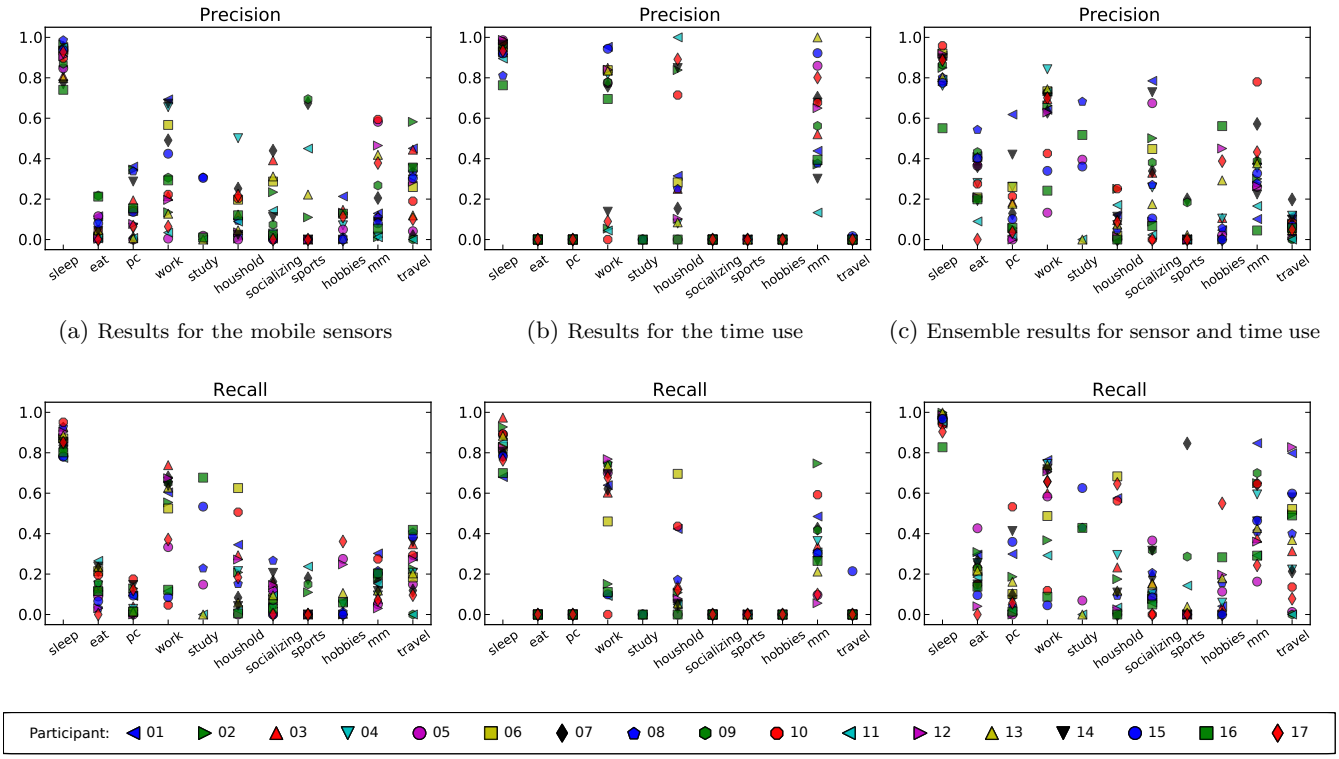
The data of these classes exhibit a higher variety of features, which is a benefit for the training process. Note here that the classification process is completely independent from the duration of each activity, which could lead to different results when considered and used in a sequential classification model.

In Figure 6(a) we can observe the precision (top) and recall (bottom) values and how they distribute for each user over all the activities. Activities that were not performed by the user are therefore never detected, since there is no training data for these activities. Unsurprisingly, sleep was detected with a high confidence for all the participants. Work varies quite a lot, depending also on the participant and the amount of work phases that have been logged in the diary. Students for example do work, but only a few hours per week, like participant 15. Precision is high (43.4%) while recall degrades to 8.25%, which means that many of the working events were unidentified. Participant 1 for example is an employee, working 8 hours every day, which is being displayed in the results of precision and recall both being in the range of 60%. We note here that the data representation plays a significant role for the classification process.

The overall results for all the participants for the mobile sensor approach lead to a precision of 20.42%, and a recall of 21.11%. We can conclude that it is possible to detect certain activities with a high confidence over a large dataset which contains inertial data that is low frequent (one minute intervals), using a common classifier like the SVM. However, since the activities are discriminated rather poorly, we need to improve the recognition rates with the help of additional information embedded in the classification process.

5.2 Time Use Only

We observe in Table 3 and Figure 6(b) that with the time use data only, we detect four out of 11 activities from the dataset, which are sleeping, working, household work and mass media usage. The precision and recall scores for the time use based approach for each activity and user are depicted here individually. As features for the time use dataset we use as much information as possible, namely *time, gender* and *age*. Overall, we reach a precision and recall of 23.76% and 17.4% respectively. The overall results for travel can be neglected, since precision and recall are below 1%. Remarkably, travel was detected for a 21 years old male participant only, who is travelling home on the weekends for several hours. This coincides with a small portion of the estimation from the time use dataset. Even though the other



(a) Results for the mobile sensors (b) Results for the time use (c) Ensemble results for sensor and time use

Figure 6: Precision (top) and recall (bottom) results for three different classification techniques from left to right: mobile sensors only, time use only estimation using the features *time*, *gender* and *age* and the combination of both modalities (ensemble). Each participant is displayed in a different sign and color. Immediately visible is that the ensemble exhibits a high variety of detected activities and is improving on recognizing more activities in comparison to mobile sensors and time use.

participants are travelling, the activity takes up only a few minutes, which is why it is difficult to detect it. Also, depending on the time use histogram, travelling is often not the most likely occurring activity.

Overall, precision for all four activities range from 34.18% to 92.87%, exceeding the results from the sensor based approach for these activities (see Table 3). Nevertheless, the rest of the classes are not recognized by the system, since the likelihoods are too small and therefore exceeded by the likelihoods from the other activities. Our findings show that the participants most common activities are detected with a high confidence with the time use dataset only.

5.3 Ensemble: Sensors and Time Use

Combination results for the mobile sensors and time use dataset are displayed in Table 3 and Figure 6(c), showing again for all participants, coded in a different symbol and color, the classification results with the ensemble of mobile sensors and time use data. We see that in contrast to time use only, we now capture all the activities appearing in the datasets. In regard to mobile sensors only, we detect certain improvements for precision and recall, but also some degrading results for a few activities. For sleep for example, we obtain a lower precision than when applying each of the other two models for participant 16 (green cuboid). Even though the results for precision were the lowest for mobile sensors and time use, we still wonder why it is dropping to 55%. After investigating the ground truth and estimated classes, we discovered that participant 16 is exhibiting a very unusual sleeping pattern, e.g., sleeping between 8pm and 10pm, watching TV until midnight and then going to bed again, waking up quite early the other morning. It seems that the ensemble is confusing too many classes here, which is why precision is dropping. For the overall results for all activities, we obtain precision and recall of 28.01% and 28.38% respectively.

The results in Table 3 indicate that adding time use data after the classification process of the mobile sensor data, it is feasible to improve on the recognition rates of the mobile sensors only. We will now discuss in detail our findings, highlighting important results that were observed during the evaluation process.

6. DISCUSSION

Having evaluated 17 datasets consisting of a total of 228 days of mobile sensor data leads to several interesting results, which are being summarized and discussed in the following paragraphs:

Time use surveys are highly useful for improving the recognition rates for activities that make up a significant portion of the user's day, which in this study were mostly *eating*, *working*, *socializing* and *mass media*. The ensemble approach leads to better results than just using the mobile

sensor data or the time use dataset to infer the activity. These results confirm what was mentioned in [16], that the use of time use data could enhance certain activity recognition systems. Additionally, the recognition rate benefits from activities that occur more regular in the time use survey for the inspected features, e.g., a male subject in his late twenties will be most likely working in the afternoon.

Time use statistics fit mobile devices. We benefit from the size of the time use database, which is below 1MB. Therefore, the data can be immediately pre-loaded on a ubiquitous mobile device. Combined with a common classifier, activity recognition can be improved directly on a mobile environment. Additionally, time use data incorporates information about the habits not only from the mobile user, but also from the people from the time use survey. A large number of people (here: over 10,000) are present in the time use database, along with their usual routines.

It is important to note that we inherently exploited the knowledge from the time use survey data, as we classified low-frequent sensor data only. It is not trivial to detect activities with such low frequent data using a common classifier, but we nevertheless recognized certain activities such as *work* with a high confidence. Also *sleep* can be detected very accurately with 2D inertial data sampled over 1 minute.

Time use surveys are less useful for detecting activities that occur only for a small portion of the day, e.g., *travelling*. Although the sensor classifier was quite confident in detecting *travelling* (3rd best recognition score for mobile sensors only, see Table 3), adding the time use survey information led to a drop in precision. Note here that *travelling* in our study includes usually activities like going home, going for lunch, taking the bus, etc., which occur not regularly and take up a small amount of time.

Even though our results are very promising, we believe that some aspects could still be improved. First, the ground truth was gathered by participants keeping a diary, where it is not clear how accurate the participants entered the activity events in the notebook. Another way of gathering the ground truth could be to let the user keep a diary directly on the mobile device if possible (a smartphone would be well suited for that task). Secondly, we might benefit from knowing the participants location, which is an evaluated feature of the time use survey, as mentioned in [4]. GPS information is already available on most of the mobile devices, which is why recording this information additionally is feasible.

7. CONCLUSIONS

This paper presents a novel approach of improving activity recognition on a mobile platform by simulating mobile usage with the SenseWear Armband and combining time use information with a common classifier. Making use of additional information in a classification process on a mobile device has many advantages. First, the sampling rate of the sensor data can be reduced. We showed how recognition rates vary with just using a common classifier and adding time use information after the classification process. Second, we improve the results for certain activities with an ensemble model. Precision for activities like eating, socializing and hobbies have been increased by approximately 25%, 18% and 6% respectively in contrast to using mobile sensors only. We discussed certain advantages of using time use survey data and identified the limits of embedding such data in the classification process.

In this paper we simulated sensor data that could have been obtained on a mobile device. For this purpose we used a SenseWear Armband device, which embeds a 2-axis accelerometer. The next step would be to put time use databases on mobile platforms such as a smartphone or a smartwatch, to perform real-time activity recognition on the device. Further, we will investigate whether the location can improve on the recognition rates.

Another interesting aspect is to use the time use data as a prior within the classification process. We would like to analyse the usage of Conditional Random Fields as a sequential probabilistic model, weighting the sensor data with the time use probabilities before training the data.

8. ACKNOWLEDGEMENTS

We would like to thank the BodyMedia team and Jonny Farringdon for providing us with 10 SenseWear Armband devices for this study. We thank all participants for wearing the device and keeping a diary for such a long time-span.

This work was sponsored by the project **Long-Term Activity Recognition with Wearable Sensors** (LA 2758/1-1) from the German Research Foundation (DFG).

9. REFERENCES

[1] L. A. Alexandre, A. C. Campilho, and M. Kamel. On Combining Classifiers Using Sum and Product Rules. *Pattern Recognition Letters*, 22(12):1283–1289, 2001.

[2] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga. Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey. In *Proceedings of the 23rd International Conference on Architecture of Computing Systems (ARCS), 2010*, pages 1–10. VDE, 2010.

[3] L. Bao and S. Intille. Activity Recognition from User-annotated Acceleration Data. *Proceedings of the 2nd International Conference on Pervasive Computing*, pages 1–17, 2004.

[4] M. Borazio and K. Van Laerhoven. Improving Activity Recognition Without Sensor Data: A Comparison Study of Time Use Surveys. In *Proceedings of the 4th Augmented Human International Conference*, pages 108–115. ACM, 2013.

[5] T. Brezmes, J.-L. Gorricho, and J. Cotrina. Activity Recognition from Accelerometer Data on a Mobile Phone. In *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing and Ambient Assisted Living*, pages 796–799. Springer, 2009.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] N. Cristianini and J. Shawe-Taylor. *An Introduction to*

Support Vector Machines and other Kernel-based Learning Methods. Cambridge university press, 2000.

[8] K. Farrahi and D. Gatica-Perez. What did you do Today?: Discovering Daily Routines from Large-scale Mobile Data. In *Proceedings of the 16th ACM International Conference on Multimedia*, pages 849–852. ACM, 2008.

[9] K. Farrahi and D. Gatica-Perez. Discovering Routines from Large-scale Human Locations Using Probabilistic Topic Models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):3, 2011.

[10] G. Jean-Louis, D. Kripke, W. Mason, J. Elliott, and S. Youngstedt. Sleep Estimation from Wrist Movement Quantified by Different Actigraphic Modalities. *Journal of Neuroscience Methods*, 105(2):185–191, 2001.

[11] J. Krumm and A. B. Brush. Learning Time-based Presence Probabilities. In *Proceedings of the 9th International Conference on Pervasive Computing*, pages 79–96. Springer, 2011.

[12] S. Lee and K. Mase. Activity and Location Recognition Using Wearable Sensors. *IEEE Pervasive Computing*, 1(3):24–32, 2002.

[13] M. Li, V. Rozgic, G. Thatte, S. Lee, A. Emken, M. Annavaram, U. Mitra, D. Spruijt-Metz, and S. Narayanan. Multimodal Physical Activity Recognition by Fusing Temporal and Cepstral Information. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(4):369–380, 2010.

[14] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner. Recognizing Workshop Activity Using Wody Worn Microphones and Accelerometers. In *Pervasive Computing*, pages 18–32. Springer, 2004.

[15] J. Pansiot, D. Stoyanov, D. McIlwraith, B. P. Lo, and G.-Z. Yang. Ambient and Wearable Sensor Fusion for Activity Recognition in Healthcare Monitoring Systems. In *Proceedings of the 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, pages 208–212. Springer, 2007.

[16] K. Partridge and P. Golle. On Using Existing Time-Use Study Data for Ubiquitous Computing Applications. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 144–153. ACM, 2008.

[17] S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat, and G. D. Abowd. Farther Than You May Think: An Empirical Investigation of the Proximity of Users to Their Mobile Phones. In *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp)*, pages 123–140. Springer, 2006.

[18] W. E. Pentland. *Time Use Research in the Social Sciences*, volume 11. Springer, 1999.

[19] R. Polikar. Ensemble Based Systems in Decision Making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45, 2006.

[20] A. Schmidt and K. Van Laerhoven. How to Build Smart Appliances? *Personal Communications, IEEE*, 8(4):66–71, 2001.

[21] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar. PreHeat: Controlling Home Heating Using Occupancy Prediction. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, pages 281–290. ACM, 2011.

[22] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele. ADL Recognition Based on the Combination of RFID and Accelerometer Sensing. In *Second International Conference on Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008.*, pages 258–263. IEEE, 2008.

[23] Y. Su, S. Shan, X. Chen, and W. Gao. Hierarchical Ensemble of Global and Local Classifiers for Face Recognition. *Image Processing, IEEE Transactions on*, 18(8):1885–1896, 2009.

[24] L. Sun, D. Zhang, B. Li, B. Guo, and S. Li. Activity Recognition on an Accelerometer Embedded Mobile Phone with Varying Positions and Orientations. In *Ubiquitous Intelligence and Computing*, pages 548–562. Springer, 2010.

[25] M. Sunseri, C. B. Liden, J. Farringdon, R. Pelletier, M. LC, S. Safier, J. Stivoric, A. Teller, and M. SureshVishnubhatla. The SenseWear Armband as a Sleep Detection Device. *BodyMedia internal white paper Healthy MDD Sleep duration min Sleep Quality Score Sleep Quality Score*, 2009.

[26] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler. Combining Multiple Classifiers by Averaging or by Multiplying? *Pattern recognition*, 33(9):1475–1485, 2000.

[27] S. Tominaga, M. Shimosaka, R. Fukui, and T. Sato. A Unified Framework for Modeling and Predicting Going-out Behavior. In *Proceedings of the 10th International Conference on Pervasive Computing*, pages 73–90. Springer, 2012.

[28] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate Activity Recognition in a Home Setting. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*, pages 1–9. ACM, 2008.

[29] H. M. Wallach. Conditional Random Fields: An Introduction. *Technical Reports (CIS)*, page 22, 2004.

[30] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster. Activity Recognition from On-body Sensors: Accuracy-power Trade-off by Dynamic Sensor Selection. In *Wireless Sensor Networks*, pages 17–33. Springer, 2008.

[31] Z. Zhu, U. Blanke, A. Calatroni, and G. Tröster. Prior knowledge of human activities from social data. In *Proceedings of the 17th Annual International Symposium on Wearable Computers (ISWC)*, pages 141–142. ACM, 2013.