Improving Activity Recognition without Sensor Data: A Comparison Study of Time Use Surveys

Marko Borazio
borazio@ess.tu-darmstadt.de

Kristof Van Laerhoven
kristof@ess.tu-darmstadt.de

Embedded Sensing Systems
Technische Universität Darmstadt
Hochschulstr. 10, Darmstadt, Germany
www.ess.tu-darmstadt.de

ABSTRACT

Wearable sensing systems, through their proximity with their user, can be used to automatically infer the wearer's activity to obtain detailed information on availability, behavioural patterns and health. For this purpose, classifiers need to be designed and evaluated with sufficient training data from these sensors and from a representative set of users, which requires starting this procedure from scratch for every new sensing system and set of activities. To alleviate this procedure and optimize classification performance, the use of time use surveys has been suggested: These large databases contain typically several days worth of detailed activity information from a large population of hundreds of thousands of participants. This paper uses a strategy first suggested by [16] that utilizes time use diaries in an activity recognition method. We offer a comparison of the aforementioned North-American data with a large European database, showing that although there are several cultural differences, certain important features are shared between both regions. By cross-validating across the 5160 households in this new data with activity episodes of 13798 individuals, especially distinctive features turn out to be *time* and participant's *location*. Additionally, we identify for 11 different activities which features are most suited to be used for later on activity recognition.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

time use surveys, activity recognition, wearable computing, rhythm modelling, probability model

1. INTRODUCTION

An activity recognition system is usually prototyped by first recording typical data from one or many different sensing modalities to create an optimal classifier with distinctive features from such data. The results have been found satisfying when classifying activities with clear links to the sensed modalities: Motion and posture sensors have for instance shown to be suitable for detecting activities like *walking*, *standing*, *sitting* [10, 13, 14] or *sleeping* [5, 11].

When considering high-level or composite activities like *having lunch* or *commuting to work*, finding distinctive sensing modalities and features is often as hard as designing an appropriate classification system. In these cases, having good and sufficient training data available is crucial to the later performance of the system. Especially the variability among study participants in how they execute the target activities, as well as in how long they tend to last and when they occur at different times, has been shown to be quite an obstacle as large sets of exemplar data are required [2].

The information from large time use surveys was suggested in [16] as a valuable instrument for designing activity recognition systems for which it is hard to obtain training data sets that contain a sufficient amount of variability. These national surveys capture for a substantial proportion of a population what activities are executed at what times during the day, and at which locations. This would allow for instance a classifier to use the current time and location to estimate what activity the user is most likely performing, based on the matching time use database's entries. For wearable sensing systems furthermore, one can assume that further user properties such as gender, profession or age, are readily available and can be used to refine this search. Although such estimates could be accurate enough for some applications, they would at the very least be promising as priors in a sensor-based activity recognition system.

In this paper, we investigate the possibilities of using data from a recent European time use survey with activity logs from 13798 participants for activity recognition in general, and compare the results to a previously published analysis from a similar-sized North-American time use study. Three contributions are made in particular:

- We give a qualitative and quantitative comparison of the US-based data analysis given in [16] with that of a comparable time use data from across the Atlantic - showing cultural differences. This includes a discussion of using large datasets and time use studies in wearable

idhh	idpers	ph01b2x	ph01c	idtag	...	zhc76	...	zvc76	...	zgc76	...
123	1	44	female	1	...	cooking	...	at home	...	listening to radio	...
123	1	44	female	2	...	going shopping	...	in car	...	listening to radio	...
123	1	44	female	3	...	cooking	...	at home	...	talking on the phone	...
123	2	46	male	1	...	eating	...	restaurant	...	talking	...
...	...	...	...	...	...	...	...	...	...	...	...

Table 1: Entries from the GTUS 2001/2002 dataset, displaying the household ID (idhh), person ID (idpers), age (ph01b2x), gender (ph01c), recorded day (idtag = {1,2,3}), main activity in time-slot 76 (zhc76), location or means of transportation in time-slot 76 (zgc76) and simultaneous activity in time-slot 76 (zgc76).

computing and the challenges in extracting data from these large data sets.

- We argue for a leave-one-household-out cross-validation methodology and perform such an evaluation in order to study the time use survey features for activity classification.
- We discuss results from single activities for a specific European region and evaluate different features that can be used for later on activity recognition.

Since the concept of time use survey is pivotal to the main idea of this paper, the next section will first describe current trends in using large data bases for mobile and wearable research studies, as well as the time use data sets used and their original application.

2. TOWARDS LARGE-SCALE DATA

Monitoring hundreds to thousands of participants over a longer time-span has in recent years become easier in mobile activity research. Some large-scale studies involving many participants being monitored continuously over weeks to months have been reported that are relevant in the context of this paper. For instance, Do and Gatica describe in [6] an experiment involving smartphone-based monitoring of 40 participants over a year to mine for human interactions. This study was widened recently in the framework of the Lausanne Data Collection Campaign[1] to 200 participants. Another work also uses data from mobile phones of 215 subjects over 5 months to analyse dwelling times, places, and mobility patterns [8]. As wearable and ubiquitous sensors are harder to deploy, similar studies in this area have had far less participants, though some studies have monitored their participants for several weeks [21]. A different approach was taken in [3] by crowdsourcing data annotation for a wearable activity and context recognition, using the mobile phones of the participants.

Several countries perform inquiries from which statistical information about the population can be derived. These time use data are usually obtained by keeping a diary for one or more days. Over the years, time use surveys have become more and more standardised so that they can be used internationally, enabling comparisons between different countries or fusing surveys to one big survey, like the Multinational Time Use Survey (MTUS[2]) or the Harmonised European Time Use Survey (HETUS[3]). The HETUS is being maintained by EuroStat, the statistical office of the European Union, and embeds time use data from 15 different European countries. Data analysis can be conducted directly on the HETUS web page, though displaying summary statistical information only. All participating countries are using the same database structure in order to fuse the data afterwards. Activity and location descriptors are abstracted: They are categorized in tiers and recorded for at least 24 hours. Participants are included according to a rigid selection process and financially recompensed, and for each a standard and anonymised set of demographic information is available.

Obtaining time use data can be challenging. Some databases are freely available, but might not be useful for our purposes, because they contain only summarized statistical information. In this section, we will have a closer look on the German Time Use Survey (GTUS) 2001/2002 data set that was used in this paper, as well as the American Time Use Survey (ATUS[4]) 2006 database that was used in recent work [16].

2.1 ATUS

The ATUS dataset is one of the few time use study databases that is freely available without restriction and is being updated every year since 2003, logging activities of US residents for 24-hours. It contains 18 different (Tier 1) activity groups and in total a distinction is made between 462 activities [18]. In addition to the activities, participants logged the time the activity started, how long it lasted and where it took place. Information about gender and age are also present, with further fields in the dataset anonymised after collection. Prior to 2011, elder care was not considered as an activity, in contrast to the GTUS, where this field is already present. For more information on the dataset we refer to the original work in [16].

This paper largely follows the study method of Partridge and Golle [16] that used the 2006 respondents version of the ATUS, which contains 12943 participants. The work contains a detailed description on how these data might be applied for activity recognition, and reports on a study using a 10-fold cross-validation analysis of the features, showing that the hour of day and location are the most useful features for activity estimation. The authors note in their paper that studies performed by different nations vary in terms of par-

[1] `http://research.nokia.com/page/11367` [02/2013]
[2] `http://www.timeuse.org/mtus/` [last access 02/2013]
[3] `https://www.h2.scb.se/tus/tus/` [last access 02/2013]
[4] `http://www.bls.gov/tus/datafiles_2006.htm` [02/2013]

ticipant behaviour (observed for instance in response rates) and constructs (motives range from quantifying unpaid work to measuring exposure to environmental pollutants). This is exactly the motivation for this paper, as it reports on studies and comparisons with the time use data from a large European country. We will now introduce the time use survey that is used for this work.

2.2 GTUS

The GTUS was first surveyed in 1991/1992, being updated every 10 years and is only accessible by regional government employees, after going through a formal admission process. The data acquisition takes usually a year (therefore it is labelled 1991/1992), in order to compensate seasonal bias and also to capture certain population groups (e.g., a single mother or father). We used the survey from 2001/2002, since the data from the current measurement period have not yet been made available.

The 2001/2002 GTUS consists of data from 13798 participants, all older than 10 years, who kept a detailed diary for three days each, writing down which activity they performed in 10-minute slots. The diary also keeps account of the location where the activity took place, as well as whether a secondary activity was performed (e.g., *watching TV* while *eating*) and who was present at the time (e.g., a household member). Additionally, personal information like relationships between household members are available. In total 272 single activities have been distinguished and allocated to three hierarchical tiers, with Tier 1 containing generic descriptions such as *personal care*, *household activities* and *mass media*, Tier 2 including a more precise description of the activity, like *sleeping*, *cooking* and *reading*, while Tier 3 contains the highest specificity, such as *sewing clothes*, *doing laundry*, and *traveling on a bus*.

Table 1 depicts an example of such a dataset (of the household with the ID no.123), displaying the first two household members and their performed main activities, as well as the locations and simultaneous activities. Also stored are information like age and gender, which were used as features later on in Section 3. Within a household, each member is allocated to an ID (idpers), whereas each household is assigned to a unique ID (idhh). The data set was created from two different sets as provided by the national statistical office, fusing the information into one table for feature extraction. We will now compare the general characteristics of both the ATUS and GTUS data sets, pointing out important differences.

2.3 ATUS vs. GTUS Overview

The GTUS can be compared easily to other European countries' time use surveys since the data is similarly structured. ATUS on the other hand is built up differently, not logging activities, location and simultaneous activities in 10-minute slots, but when the activity started and how long it lasted. Nevertheless, research groups like the Centre for Time Use Research (CTUR[5]) are maintaining the MTUS in order to create a huge time use survey, including to this day ATUS and HETUS.

Table 2 lists the basic properties in terms of the type and the amount of data that was included for the ATUS and GTUS sets: both data sets are similar in the amount of participants and time monitored. GTUS identified more activity episodes, which can be explained by the fact that more days per participants are present and that activity episodes are logged in 10-minute intervals. A previous research study on the ATUS data set [16] considered hour-of-day (60 mins) as a time interval for simplicity.

Property	**ATUS 2006**	**GTUS 2001/2002**
participants	12943	13798
households	12943	5160
# activities per tier	18 / 110 / 462	10 / 48 / 272
# locations	27	8 / 21
time interval (mins) data set / study	1 / 60	10 / 10
period monitored	1 day	1-3 days
activity episodes	263,286	356,910

Table 2: Comparison of the basic properties of the ATUS 2006 taken from [16] and GTUS 2001/2002. The time interval for ATUS was remodelled in [16] from minutes to hour-of-day, displaying therefore not the original intervals (duration of an activity in mins).

A further property of the GTUS is that it keeps track of the time use for all members in a household above the age of 10, whereas ATUS explicitly chooses single participants from one household. As mentioned previously, the ATUS is being updated every year, interviewing participants over the phone and keeping track of their activities for one day. Therefore, the data set is always up-to-date in contrast to the GTUS, which is being refreshed every 10 years. The following section will display quantitative results for the GTUS, as well as the ATUS, highlighting promising features from the data sets.

3. EVALUATION

The contributions of this work are threefold: 1. We will discuss a demographic comparison between GTUS and ATUS. 2. We show quantitative results for the GTUS data set and 3. we highlight important activities from the GTUS for probable mobile and wearable research.

3.1 ATUS vs. GTUS Demographic Analysis

In order to perform a graphical comparison, a few conversion steps had to be taken. A first hurdle is the difference in categorization and hierarchy of activities for both data sets: The most relevant Tier 1 and Tier 2 activities of the GTUS were translated to the corresponding 18 Tier 1 activities of ATUS. Furthermore, to reflect the 10-minute segments in the GTUS data set, all entries from the ATUS were converted to 10-minute time slots.

The resulting more in-depth visualization of both datasets is shown in Figure 1. Displayed are all 18 Tier 1 activity groups from the ATUS 2006, showing per time-of-day the performed activity of the participants in per cent. The figure shows some differences that exist in the activity reporting: Where the ATUS dataset contains significant digit bias (i.e., a bias of participants rounding off start and stop times of reported activities toward full or half hours, visible as jagged

[5]`www.timeuse.org` [last access 02/2013]

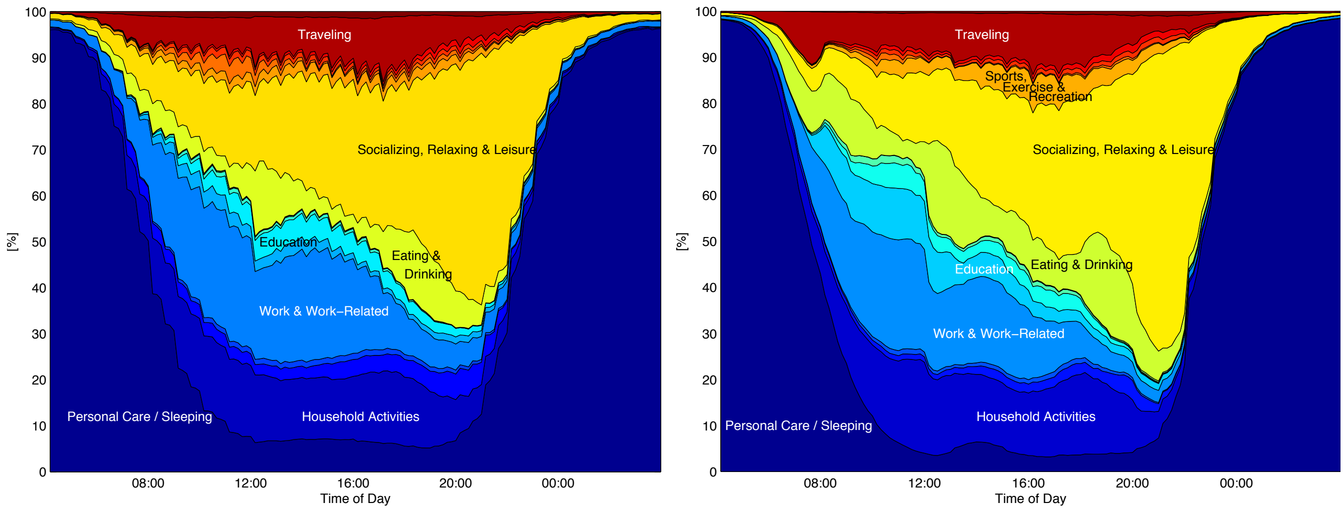

Figure 1: A visual comparison of both ATUS 2006 (left) and the GTUS 2001/2002 (right), depicting per time-of-day the normalized occurrences of common Tier 1 activities over all participants. Slight cultural differences appear in digit bias (jagged edges on left) and activity clusters (e.g., breakfast, dinner on right).

edges in the plot), this is less pronounced in the GTUS even though the reporting time intervals are relatively small for both. A second more cultural difference that is visible from this visualization is that the GTUS dataset contains stronger time dependencies for particular activities (see the larger increases around the times for breakfast, lunch and dinner for instance), including a sharp rise in leisure activities after 20:00. The plot displays significant differences in both US and European regions, pointing to the importance of using time use data for specific regions only, since demographic characteristics have to be considered. We will now have a closer look at the GTUS data, describing how time use data is being used to infer which activity has been performed.

3.2 GTUS Activity Recognition

Methodology. We use all 5160 households from the GTUS for a cross-validation analysis: Each of the 5160 households is left out to calculate the maximum likelihood of an activity for the rest of the 5159 households. For each member of the left-out household we compare the activity to the most likely one from all the other households. For each household a confusion matrix is being stored, from which precision, recall and accuracy are calculated for later analysis. The residents in a household vary from 1 (total count: 15432) to 8 (total count: 9), where 1 and 2 household members provide the majority (27050 from 35691 members in total). We believe that within a household, participants tend to the same activities when spending time together, leading to biased results. Therefore, we exclude a complete household and observe the activity distributions for the rest of the households, constructing maximum likelihood classifiers. For this we use as input different features $f_1, ..., f_n$, from which the classifier derives for a target activity A the maximum conditional probability $P(A|f_1, f_2, ..., f_n)$.

Features. Many different features can be extracted from the GTUS, like gender or even location. In Section 2.3, we remodelled the GTUS to fit the ATUS. Here, we use the GTUS as it is, changing only minor things (the changes will be discussed later on in this section). To evaluate the features for activity recognition from time use data, we consider different aspects similar to [16]. Therefore, as features we detect:

1. **Time.** Time is a significant feature for activity recognition as derived in [7, 9], since people tend to their behaviour patterns. The fine granularity as given by the GTUS for logging activities in 10-minute slots will confirm that. Therefore, we will not use hour-of-day as [16] did.

2. **Prev. act.** A previous activity is an activity that took place prior to a different activity. We are not considering the activity prior to a time-slot, which would lead to a biased result when considering for example sleeping: prior to sleep we most likely would be sleeping.

3. **Location.** The idea that knowing the location might infer to the activity that is being performed, was mentioned in [7, 1]. For our work, this feature needs some remodelling, fusing all 20 different means of transportation (e.g., *in the bus*, *by foot* or *in car*) into one *transportation* variable to simplify the use of this feature and to minimize classification runtime. We receive 10 different locations out of 29 from the original data set.

4. **Gender.** Recently, researchers in [15, 12] used gender to infer which activities are being performed by participants of the same sex. Other physical characteristics have been added as well, but we will focus on gender itself as a feature.

5. **Age.** The age in the GTUS varies from 10 to 80 years, which is why we divided the datasets into 5 years age groups (10-14, 15-19, ..., 75-80), just as in previous work [16], not only to simplify the calculation, but also to sustain a significant amount of participants per age group when using the classifier.

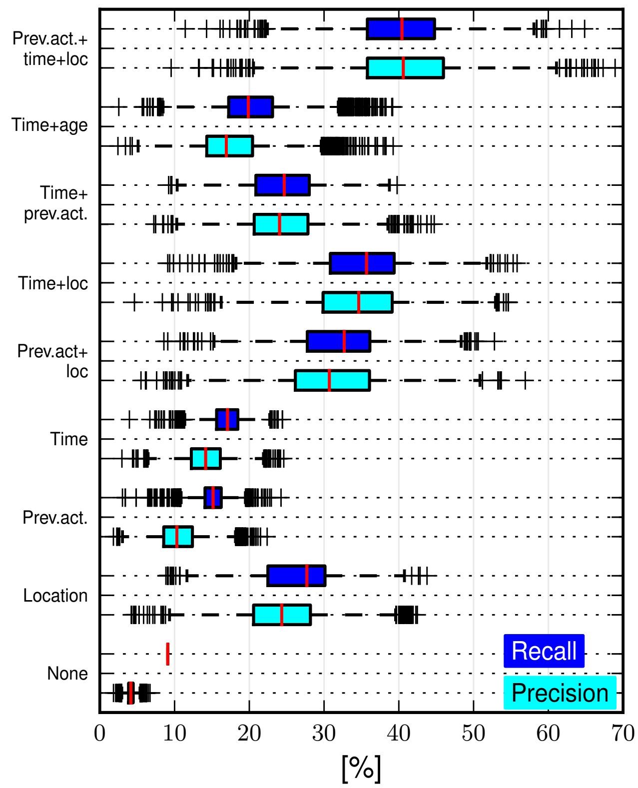

Figure 2: Precision and recall results for different features and their combinations from the GTUS, plotted as box plots with the median (red line), upper and lower quartile (75% and 25% respectively, end of boxes), whiskers for highest and lowest values and fliers ('+') as outliers. Combination of *prev.act.*, *time* and *location* yield the highest results.

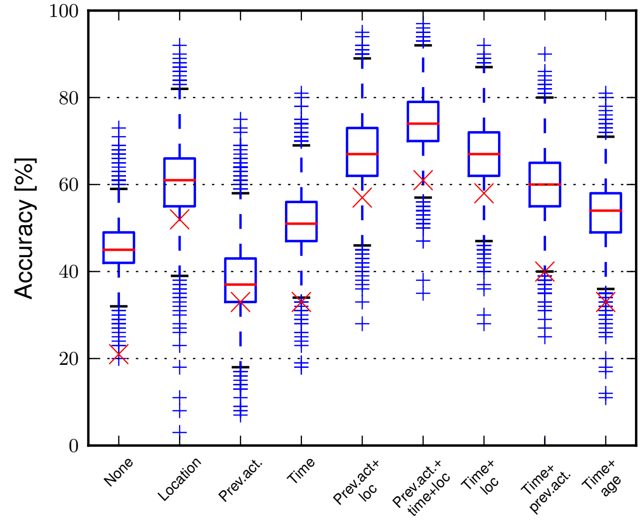

Figure 3: Box plots of the accuracy results per household from the GTUS and overall average results from the ATUS (values taken from [16]) marked as a red 'X' for different features and their combinations. Here, the GTUS results exceed the ATUS results, marked by the median box plot (red line in each box).

In the process of evaluation, we do not only use the single features but also all combinations of the features, e.g., by combining *time* and *location*. A list of these combinations can be observed in Figure 2, left side.

In the course of this work, we considered *traveling* (e.g., *traveling between the home and the office* or *from school to the home*) as an important activity, which is present in all of the 10 Tier 1 activity groups of the GTUS as individual activities (e.g., within the Tier 1 group *work*, *going to work* would be an activity). To detach *traveling* from the Tier 1 groups, we creat an 11^{th} Tier 1 activity group *travel* to which we relocate all travel codes from the other Tier 1 groups. A complete list of the GTUS Tier 1 activity groups as used in this paper is shown in Table 3, left side.

Results. Figure 2 shows precision and recall for different features and feature sets as box plots. We see the median being displayed within the boxes as a red line, the upper and lower quartile (75% and 25% respectively), the whiskers which show the whole range of the data, as well as the fliers ('+'), displaying the outliers within the results. Starting with *none* we can observe how the values for precision and recall vary as we add more and more features to the classifier. *None* yields a very low precision and recall (4.17% and 9.09% respectively), since the results are biased by the activity *sleep* (this is the case also for *age* and *gender*, leading to the same overall results as *none*). For single features, precision and recall are below 25% and 28% respectively, whereas the highest results are gained by using *prev.act.+time+location* (precision of 40.79% and recall of 40.25%), showing also outliers in the 60% region.

Location has the highest impact on the results: For *location* alone the results are already quite high, but can be even increased by adding *time* as a feature, resulting in 34.58% precision and 35.24% recall. Some improvement can still be achieved by adding *prev.act.* to the feature set. *Time* seems to have the highest impact here, since the results are not far apart, whereas *prev.act.* contributes little but significant to the results.

We showed that a combination of features yields the highest probable result for inferring the right activities over a large-scale data set from the GTUS, highlighting especially *time*, *location* and *prev.act.* as proper features. We will now have a look at the results taken from recent work [16] for the ATUS data set and compare them to the GTUS results, also showing accuracy for different features.

GTUS vs. ATUS. In order to perform a quantitative comparison between the GTUS and ATUS, we display the results for different features for both in Figure 3. Note here, that from [16] only the results presented in a barplot graph were available. Unfortunately, both of the authors did not have the original result sets, which is why we read off the approximate results from the paper.

Figure 3 shows for different contextual variables the distributed accuracy per household as box plots, highlighting the median (red line in the box). Additionally, a red 'X' marks the accuracy from the ATUS data set for the same features and their combinations. We can observe that the accuracy results for the GTUS exceed the ATUS results, especially for the highest accuracy, when *prev.act.* is being combined with *time* and *location* (the median is always above the ATUS accuracy).

Prev.act. showed promising results in recent work [16], which is why we used it in combination with *location*, resulting in an accuracy of 67.7%. Compared to *location* on its own, a slight increase of +6.54% is achieved. The same can be observed for the ATUS. Adding now also *time* to the feature set *prev.act.+location*, accuracy is rising to 74.78%, leading to the highest results. In contrast to that, if we use *prev.act.* with *time*, we achieve a better result than *time* on its own (51.76%). We can conclude that *previous activity* is rather a weak feature, but combined with *time* and *location*, the results are promising. Compared to previous research, similar findings have been reported for the ATUS. *Time* combined with *location* yields an accuracy of 67.72%, again increasing the accuracy by +15.96% compared to *time* itself. Adding *age* or *gender* to *time*, a marginal increase of 1% to 2% in contrast to *time* by itself is perceptible. The results indicate that activity inferences perform better for the GTUS.

Note here, that the duration of an activity has not been considered. As shown in recent work [16], a duration weight for the activity can increase the accuracy. Since we consider a 10-minute interval of the given activities as sufficient, we left out a duration weighting. In the next section we will evaluate the results for single activity groups of the GTUS data set, highlighting important features for recognizing certain individual activities.

3.3 Results for GTUS Tier 1 Activities

In order to perform a more in-depth analysis, we again use the maximum likelihood classifier in a leave-one-household-out cross-validation (the same method which was used in Section 3.2) to calculate for the best performing feature combinations from Figure 2 the precision and recall for single Tier 1 activity groups of the GTUS. We include *travel* here, thus resulting in 11 groups in total. Table 3 displays the results for precision and recall. Also shown are 'average duration per day' for the activities, showing how participants spend their time during the day on average. The hyphens in the table indicate that an activity group was never chosen by the classifier, because other activities were preferred. Results for the ATUS, which are taken from [16], are also shown in the table. Activity groups *hobbies*, *mass media* and *unknown* do not occur as single Tier 1 activity groups in the ATUS, which is why 'n/a' was put in the corresponding rows in the table. Note here, that we are not comparing the ATUS to the GTUS, but rather displaying the results as a reference in the table.

Immediately visible in Table 3 is that *location* yields high results for specific activities. *Personal care* achieves a high recall of 97.7% and a precision of 60.2%. *Location* results are again biased by the activity *sleeping*, which is a Tier 2 activity of *personal care*, and since participants usually sleep at their home, this activity is being preferred by the classifier. This explains also the poor results for *household activities* for *location*, although one might expect that this activity is being mostly performed at home. Adding *time* raises the results for *household activities.* Note here, that while *household activities* gains in recall, for *personal care* it is dropping, showing that misclassifications are being corrected. On the other hand, recall is dropping for *personal care.*

Adding features leads for some activities to a raise in precision and recall, achieving the highest results again by the combination of *location* with *prev.act.* and *time* in almost all the activity groups. Thus, *volunteer activities* are predicted at all, whereas a drop in the results of *education* is perceptible. Therefore, adding features does not always yield better results. *Travel* is a striking example here, since precision and recall exceed both 90%, although *prev.act.* is not adding much to the results. *Location* and *time* are the preferred features here. *Travel* as well as *sports* are an example of activities that cannot be inferred by the feature *time.* Additional features correct that, especially *location* has an impact on the results for *sports*, which seems to be performed at specific locations.

Rather poorly detected are *volunteer activities*, *education* and *hobbies*, for which none of the features yield high results in recognizing the activities. For the latter two activities we believe that activity recognition systems can help raising the recognition rate, whereas *volunteer activities* could be hard to predict. A key factor here is the time spent for performing the activity: *volunteer activities* are carried out on average 10 minutes per day, for which the 10-minute slots of the GTUS could lead to other activities that are being preferred by the classifier.

We conclude that for specific activities, a combination of features yield the highest results for recognizing the activities. Such information could be used as a prior for activity recognition systems. Note here, that with the GTUS a demographic analysis for the activities is being performed. We believe that using such information in combination with a classification system that considers a user's activity pattern, even higher recognition rates could be achieved. We will now summarize the findings of Section 3.

3.4 Results: Summary

Having analysed the GTUS 2001/2002 data set in detail, we conclude the following from the results shown throughout Section 3:

- The two time use survey databases experimented on (ATUS 2006 and GTUS 2001/2002) have fundamental differences in composition and structure that make it challenging to apply time use surveys across regions. Especially for activity recognition systems, it would be important to have more unified data sets with identical activities and survey data collection approaches.

- Demographic differences were found between the data from North-America (ATUS 2006) and Germany (GTUS 2001/2002), which is why it is important to use time use data that was taken from the same region as where the system is to be deployed.

- We detected *location* and *time* as the best performing features in the GTUS data set, which is in line with previous work in this area. Activities such as *personal care* or *travel* can best be represented by *location* and *time*. *Prev.act.* was also high-lighted during the evaluation, but made only minor contributions to the overall results. Similar results were found in recent work [16], with the exception that *prev.act.* had a higher impact on the results.

Activity	Average hh:mm per day		time		time + age		prev.act. + time		location		location + time		location + prev.act. + time	
	GTUS	*ATUS**	*GTUS* Pre Rec	*ATUS* Pre Rec	*GTUS* Pre Rec	*ATUS* Pre Rec	*GTUS* Pre Rec	*ATUS* Pre Rec	*GTUS* Pre Rec	*ATUS* Pre Rec	*GTUS* Pre Rec	*ATUS* Pre Rec	*GTUS* Pre Rec	*ATUS* Pre Rec
Personal care	11:00	9:23	68.8 86.7	74.5 87.2	72.5 78.3	77.8 85.8	80.8 89.2	84.0 88.2	60.2 97.7	56.8 100	80.3 86.6	82.3 89.0	85.7 92.1	86.6 89.1
Work	2:00	3:27	22.6 2.6	30.2 61.3	22.9 32.2	36.7 68.5	32.0 42.8	44.8 74.5	53.3 60.8	93.7 87.9	53.3 60.8	93.6 87.9	56.1 60.4	95.0 87.6
Education	0:40	0:27	-	-	10.7 15.2	32.9 39.0	-	44.1 4.4	-	70.0 64.7	-	70.0 64.3	6.9 1.2	72.2 59.8
Houshold activities	2:50	1:49	23.9 53.9	-	26.4 38.0	-	34.7 52.5	28.6 11.6	37.8 16.0	52.3 0.1	37.1 55.4	31.5 14.2	50.2 71.9	34 22.2
Volunteer activities	0:10	0:07	-	-	-	-	-	15.7 0.8	-	-	0.1 0.0	2.9 0.0	6.2 1.8	24.7 2.9
Socializing and pleasure	1:30	4:31	-	39.3 52.1	4.3 0.4	39.8 57.5	24.5 11.3	41.6 61.0	-	47.8 14.3	36.9 23.5	47.1 71.2	45.2 28.5	49.0 73.8
Sports	0:30	0:18	-	-	-	-	-	16.8 0.4	19.4 23.8	48.5 36.4	24.9 19.0	47.2 35.5	35.8 29.6	48.8 31.3
Hobbies	0:30	n/a	-	n/a	4.2 5.5	n/a	-	n/a	-	n/a	0.1 0.0	n/a	3.4 0.7	n/a
Mass media	2:30	n/a	40.7 43.7	n/a	38.2 44.2	n/a	54.1 54.2	n/a	-	n/a	49.7 43.0	n/a	60.2 57.1	n/a
Travel	1:20	0:11	-	-	5.6 0.4	-	41.0 19.4	49.4 31.8	98.0 99.3	97.0 96.0	98.0 99.3	96.8 95.9	98.4 99.3	96.1 94.8
Unknown	0:00	n/a	-	n/a	-	n/a	-	n/a	-	n/a	-	n/a	0.5 0.2	n/a

*The values and results for the ATUS are taken from [16]

Table 3: Our Tier 1 precision and recall results for the GTUS time survey database, alongside the results from the ATUS time survey Tier 1 activity results for reference (as mentioned in [16]). The best performing features from Figure 2 are used for displaying the results for the 11 activity groups. A hyphen indicates that the activity was never predicted, and 'n/a' indicates that the activity group is non-existent in the ATUS time use database.

4. CONCLUSIONS

We investigated the feasibility of using time use databases for wearable activity recognition systems, contributing with the analysis of the GTUS specifically. We compared the results from this study with previous work in [16] and discussed the use of these time use data in wearable activity recognition. The comparison of a US and European time use survey revealed that demographic differences need to be considered. Results of feature analysis for both data sets show that *time* is a very important feature, and that even when considering more fine-grained time slots of 10 minutes, the activity estimation with just *time* is still 50% on average, without using any sensor data sets to train from. Another import feature is *location*, which shows strong affiliation to certain activities, for example when considering *traveling* or *sleeping*. The highest results were achieved when using a combination of the features *location*, *prev.act.* and *time*, showing an overall accuracy across all tier one activity classes of almost 75%.

Combining the results for *time* and *location* with a probabilistic model for activity classification, like the Hidden Markov Model (HMM) [17, 19], better recognition rates could be achieved. Recent work [4] about detecting sleep segments relied on a HMM classifier, training on several days to set up a personal pattern for the recorded user, using time as a prior. Such work could profit from knowledge of time use data sets. Knowledge transfer can be used for other activity recognition systems as well. Making use of prior knowledge was also introduced in [20], showing how a sensor network system can profit from knowledge of other sensor network systems in a similar environment. Time use data reflects how common activities for a certain region are being performed by its residents, which can be used in other activity recognition systems as well.

Future projects could be using the findings and data from this work to construct prior models or first estimates for activity recognition systems in wearable activity recognition deployments. A project of recording dozens of participants with a wearable activity sensor is in progress to examine whether we can increase recognition rates of daily activities with the knowledge derived from the GTUS. Furthermore, we will be using the current GTUS 2011/2012 for future studies, expecting different insights in activity patterns and daily routines.

This paper's experimental scripts were written in Python and can be applied immediately to the ATUS and GTUS dataset versions mentioned to reproduce the results. They are publicly available with supplemental information[6].

[6] `http://www.ess.tu-darmstadt.de/datasets/ah_2013`

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[2] 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.

[3] M. Berchtold, M. Budde, D. Gordon, H. Schmidtke, and M. Beigl. Actiserv: Activity Recognition Service for Mobile Phones. *14th International Symposium on Wearable Computers (ISWC)*, pages 1–9, 2010.

[4] M. Borazio and K. Van Laerhoven. Combining Wearable and Environmental Sensing into an Unobtrusive Tool for Long-term Sleep Studies. In *Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics (IHI)*, pages 71–80. ACM, 2012.

[5] L. de Souza, A. Benedito-Silva, M. Pires, D. Poyares, S. Tufik, H. Calil, et al. Further Validation of Actigraphy for Sleep Studies. *SLEEP*, 26(1):81–85, 2003.

[6] T. Do and D. Gatica-Perez. GroupUs: Smartphone Proximity Data and Human Interaction Type Mining. In *15th IEEE International Symposium on Wearable Computers (ISWC)*, pages 21–28. IEEE, 2011.

[7] N. Eagle and A. Pentland. Reality Mining: Sensing Complex Social Systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[8] N. Eagle, J. Quinn, and A. Clauset. Methodologies for Continuous Cellular Tower Data Analysis. *Proceedings of the 7th International Conference on Pervasive Computing*, pages 342–353, 2009.

[9] 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.

[10] T. Huynh, M. Fritz, and B. Schiele. Discovery of Activity Patterns Using Topic Models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 10–19. ACM, 2008.

[11] 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.

[12] G. Krassnig, D. Tantinger, C. Hofmann, T. Wittenberg, and M. Struck. User-friendly System for Recognition of Activities with an Accelerometer. In *4th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 1–8. IEEE, 2010.

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

[14] J. Lester, T. Choudhury, and G. Borriello. A Practical Approach to Recognizing Physical Activities. *Proceedings of the 4th International Conference on Pervasive Computing*, pages 1–16, 2006.

[15] T. Maekawa and S. Watanabe. Unsupervised Activity Recognition with User's Physical Characteristics Data. In *15th International Symposium on Wearable Computers (ISWC)*, pages 89–96. IEEE, 2011.

[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] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] K. Shelley. Developing the American Time Use Survey Activity Classification System. *Monthly Labor Review*, 128:3–15, 2005.

[19] T. Starner, J. Weaver, and A. Pentland. Real-time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375, 1998.

[20] 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*, pages 1–9. ACM, 2008.

[21] K. Van Laerhoven, D. Kilian, and B. Schiele. Using Rhythm Awareness in Long-term Activity Recognition. In *12th IEEE International Symposium on Wearable Computers (ISWC)*, pages 63–66. IEEE, 2008.