Diary-Like Long-Term Activity Recognition: Touch or Voice Interaction?

Philipp M. Scholl, Marko Borazio, Martin Jänsch and Kristof Van Laerhoven
Embedded Sensing Systems
Technische Universität Darmstadt
Email: scholl,borazio,jaensch,kristof@ess.tu-darmstadt.de

***Abstract*—The experience sampling methodology is a well known tool in psychology to asses a subject's condition. Regularly or whenever an important event happens the subject stops whatever he is currently involved in and jots down his current perceptions, experience, and activities, which in turn form the basis of these diary studies. Such methods are also widely in use for gathering labelled data for wearable long-term activity recognition, where subjects are asked to note conducted activities. We present the design of a personal electronic diary for daily activities, including user interfaces on a PC, Smartphone, and Google Glass. A 23-participant structured in-field study covering seven different activities highlights the difference of mobile touch interaction and ubiquitous voice recognition for tracking activities.**

I. Introduction

Ecological Momentary Assessment (EMA) [1] is a type of psychological study design concerned with the sampling of experiences throughout the course of a day. In contrast to study designs like End-of-Day diaries, EMA designs are less prone recall, recency, peak and summary bias [2]. Mainly because the sampling happens on a regular basis by the use of electronic diaries. These can be either watches, palm computers, Smartphones or any other mobile computing device, which remind the participant that an input is necessary. This "manual" sampling of experiences/activities is dual to gathering self-reported ground truth for wearable activity recognition systems.

Wearable activity recognition with the goal of detecting for example leisure activities [3] or certain habits [4], can serve as triggers for feedback in EMA study designs and replace regular sampling. However, during the validation and training phase of a wearable activity recognition system a sample for each detected activity is necessary for semi-supervised approaches [5]. Similar to the regular sampling in an EMA approach, this has to be done manually by the participant. Depending on the complexity of the detected activity the system needs to be retrained and ground truth resampled for each user. This kind of interaction loop is depicted in Figure 1. Interaction with a mobile phone provides the labelled data for an activity recognition on wearable sensor data, which in combination give a momentary activity assessment. One important aspect is the design of the mobile interaction in order to minimize

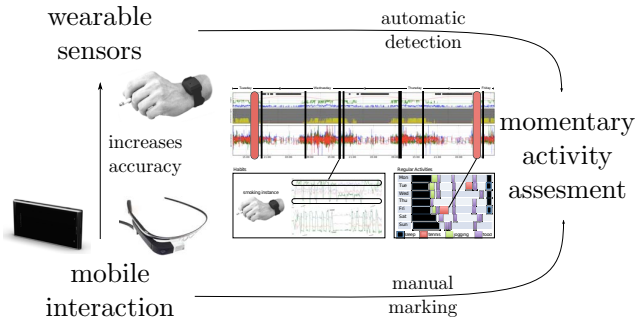

Fig. 1. An interaction loop aiming at tracking/assessing a user's activity. The explicit interaction with a mobile phone or wearable device needs more effort than the implicit tracking by wearable sensors. However, since the explicit interaction is more exact it lends itself as ground-truth for detecting activities automatically from wearable sensors.

the overhead of gathering activity annotations, as well as the technical design of data collection and management.

The approach of letting users annotate their data themselves during a study has been proposed by Bao et. al. [6]. Using a paper-based approach they were able to show the feasibility and comparable performance of activity recognition based on this user-provided ground truth. Considering body-sensor networks it has been shown [7] that gathering user annotated data over a long-term with a Smartphone is feasible. And that a semi-supervised activity recognition approach, in which only sparse annotations are used, can lead to practical results [8].

In this paper we present a system to explore the possibility of using a Smartphone and Google Glass to record the ground truth for wrist-movement data gathered with the HedgeHog [3] sensor. The system does not only record the manually entered ground truth but can also serve as a hub for the recorded sensor data from a Smartphone, Google Glass and HedgeHog unit. We hypothesize, that due to a smaller interaction time, an interaction based on an always-ready head-worn display and voice input is preferred over touch interaction on a Smartphone for "manually" recording activities, especially for activities which require the user's hands to be completed. A structured in-field user study with 23 participants shows that this hypothesis does not hold in this generality.

978-1-4799-6136-8/14 $31.00 © 2014 IEEE
DOI 10.1109/BSN.Workshops.2014.12

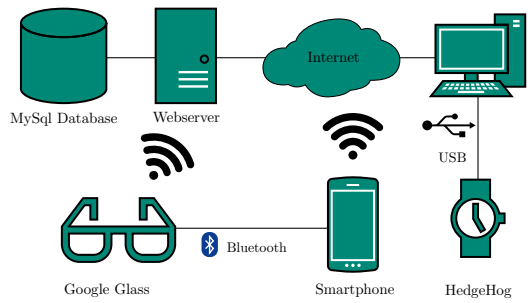

Fig. 2. Overall system architecture of the system. Smartphone and Google Glass are connected via Bluetooth to synchronise interaction events, letting the other device know that the user started or stopped an activity. Sensor data is aggregated and merged via timestamps on the webserver (through Wifi or USB).

II. Design and Implementation

The system we designed can be coined as a personal activity recognition tool (PART). Similar to a PC-based activity tracker application, in which current activities are manually tracked by marking their start and stop times, a PART allows for the same type of interaction but additionally controls available worn sensors and records their data. This optional manual marking of activities and automatic recording of sensor data allows to train activity recognition algorithms (e.g. [8]) on the worn sensors. This creates a loop as depicted in Figure 1, where the PART is used to automatically detect the context of its user after manual marking of activity instances.

A. Interaction

Three functions are supported by the PART system: starting, stopping and choosing an activity. An activity is defined as a string defined by the user, like walking, sitting, cycling etc.. These strings or labels can be defined by the user either through a web-, touch- or speech-interface on a PC, Smartphone or Google Glass.

The web-interface is optimized for a PC-based interaction, i.e. point-and-click with a mouse and keyboard. It displays the raw sensor data which has been gathered, together with the timeframes of the marked activities. Furthermore the activity labels can be moved, resized, created and deleted after a recording session. This allows to review activity labels later on, for situations where labelling could not be achieved or has been done imprecisely on purpose, for example using a Smartphone prior to washing hands.

The Smartphone interface, which has been optimized for touch interaction is depicted in Figure 3. Similar to the web-based interface it allows to add new activity labels and, depending on the screen size, displays at least five buttons which toggle the recording of its respective activity. Additionally, to pushing the buttons, activities can be toggled by swipe gestures, from the midpoint of the display to the direction of each button, which allows for less exact interaction compared to pushing a button. The application can either be accessed directly, as a widget on the homescreen or on the lockscreen. Activities can also be started/stopped through the same voice commands as used on Google Glass: "I am [x]" and "I stopped [x]", where [x] denotes labels like "sitting down", "brushing my teeth", "washing my hands".

While the system on Google Glass supports the same voice commands to start and stop activities its graphical user interface is completely different. Even though it is in principle possible to run the same application on both an Android Smartphone and Google Glass, we decided for two applications to support the input modalities of both devices properly. For Glass this means that mainly voice input is to be supported, an activity can be started with the "I am" command and displayed as in Figure 3.1. After selecting the specific activity Figure 3.2 is displayed to show the user the current activity. Whenever the user wakes up Glass again, the activity can either be stopped via voice command or by tapping the touch-sensitive part of the device. All actions started with voice commands can also be achieved through touch interaction.

In total a user is presented with six different input modalities on three devices. Point'n'Click can be used on the web-based application. The Smartphone supports touch interaction, drag interaction and speech recognition and Google Glass supports touch interaction and speech interaction. From an interaction effort point of view, speech recognition together with feedback on the displays is the only interaction type which does not require using ones' hands and therefore provides hands-free mobile interaction.

B. Synchronization

Since multiple devices are forming the system, synchronization of both interaction and sensor data is an important aspect. Fig. 2 depicts the connections of all components of the system. Google Glass and the Smartphone are connected via Bluetooth, for exchanging information about started and stopped activity, i.e. an activity can be started on the mobile phone and stopped on the Glass or vice versa. Additionally, the logging process is started on all devices when at least one activity is being recorded. The HedgeHog, i.e. a watch-like accelerometer logger, is running continuously. Sensor data is aggregated whenever a recording session (or by user choice) is terminated. Google Glass and the Smartphone automatically upload their sensor data to a central server under a given user identification. The HedgeHog data needs to be manually uploaded by visiting the web-application and uploading files found on its USB mass-storage emulation. The data from all devices are merged by their global timestamp, which requires that the clocks of all devices are synchronized. For the Smartphone and Google Glass this is achieved through their Bluetooth connection. However, since the HedgeHog does not have any wireless connection, its clock can only be synchronized when connected via USB and therefore the synchronization resolution is limited by the respective clocks' drift. Such a kind of data syn-

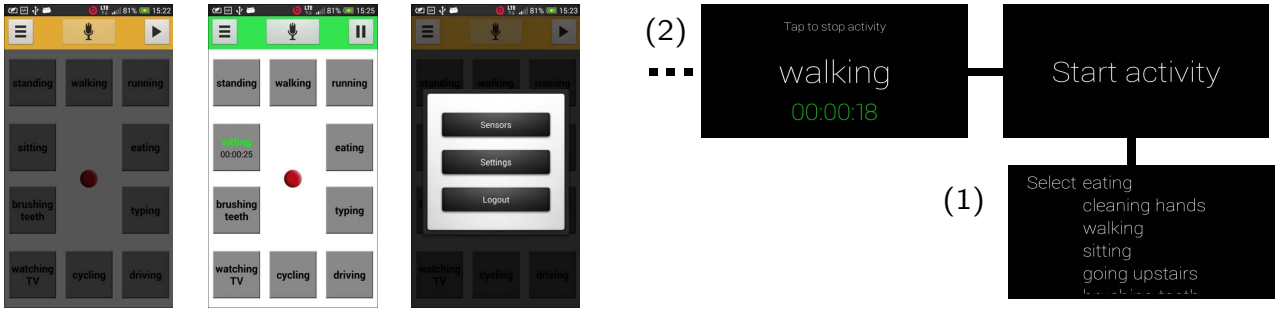

Fig. 3. User interfaces on the Smartphone and Google Glass. Both can be accessed via voice and touch interaction.

chronization allows the user to switch seamlessly between his preferred interaction modality (voice vs. touch, Glass vs. Smartphone vs. PC) and to record sensor data from multiple wirelessly connected and unconnected devices.

III. Experiments

A system like PART, which automatically detects the user's current activity, should require no explicit interaction to be most useful. However, to bootstrap or personalize activity recognizers at least a few explicit interactions are necessary, i.e. a few instances of to be recognized activities need to be manually marked. This in turn requires that marking these activities can be done as quickly and accurate as possible. We conducted a study to test the following two hypotheses: (1) given a shorter interaction time (due to not having to take out the Glass), user will prefer Google Glass for marking activities and (2) hands-free operation (i.e. voice input) is preferred for activities that require the hands to complete.

A. Study Design

For this we asked participants (10 female, 13 male, aged 22-33) chosen from the vicinity of the authors to perform 7 activities (Table I) in random order and use whatever input modality they prefer to mark the beginning and end of those activities. All participants were introduced thoroughly to each input modality on both Google Glass and the Smartphone. After this introduction, the participants were given one hour to perform all activities at least once. During the course we logged which interaction has been used, and how long each interaction took. The latter was achieved by taking the time the device was activated until an annotation was completed. This allows to test hypothesis (a), while the activities "brushing teeth", "cleaning hands" and "eating" were specifically chosen to test hypothesis (b). At the end of each session the participants were presented with a questionnaire to get their subjective opinion on the overall system. Especially we asked participants to *rate* if annotating data can be done faster with Google Glass, if the presented application could be used intuitively, if the voice recognition had an acceptable error rate and if they preferred touch over voice interaction on a standard Likert-scale. We furthermore gave participants the possibility to explain if and why they preferred a certain device, and wether they preferred touch over voice.

activity/new activity	instances p.P.	Glass/Phone	Voice/Touch
sitting/sitting down	3.3 ± 1.3	28/71%	19/80%
walking/taking a walk	3.6 ± 2.0	24/75%	15/84%
eating/eating something	1.5 ± 0.7	17/81%	13/85%
going/walking downstairs	1.7 ± 1.0	26/74%	20/80%
going/going upstairs	1.7 ± 0.7	31/69%	21/79%
cleaning hands	1.2 ± 0.6	41/59%	29/71%
brushing teeth	1.1 ± 0.3	40/60%	24/76%
	14 ± 3.2	30/70%	20/79%

TABLE I
List of recorded activities.

B. Results and Discussion

Table I shows the mean number and standard deviation of recorded instances per participant and activity. It shows which device and which input modality was used for annotating as gathered by our logging process. In total Google Glass has been used in 30% of all instances, the Smartphone in 70% of all cases and the web-application based interaction has been omitted since it has been hardly used. One possible explanation for the relatively low usage number of Google Glass could be that participants are more trained/comfortable using a Smartphone, which was confirmed by most interview answers. Looking at the mean interaction times (6.29 seconds for Smartphone, 10.95 seconds for Glass touch and 3.19 seconds), i.e. Smartphone interaction is generally faster than Glass interaction. This points to a problem with the voice input design - our voice input phrases have not been chosen carefully enough. Even though the touch interaction for Google Glass is about double the speed of the Smartphone it has been rarely used. This is because it can only been used to annotate the end of an activity, since this involved activating the Glass and tapping the device, while starting an activity on Glass required to navigate through the list of all applications and activities. Compared to that, the Smartphone application was always in front since there were no other applications active during the study. Most participants explained their Smartphone preference by the more reliable touch input and being less conspicious in public.

Interestingly, usage of voice input for activities that involved using hands (cleaning hands, brushing teeth)

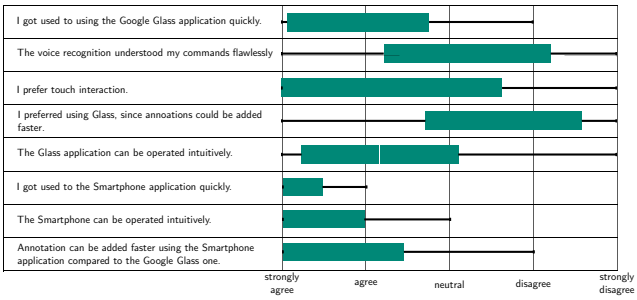

Fig. 4. Answers depicted as mean, standard deviation and range of answers given to facts on a standard Likert-scale.

increased only slightly, as can be seen in the modality row of Table I, where the first number represents voice input and the second touch input. Originally we assumed that voice interaction would be preferred, however interaction times show that it actually took longer for participants to interact with the system via voice. Most probably this is due to a non-optimal choice of input phrases, Table I contains the used voice input phrases during the study (first row). It can be seen that "sitting", "walking" and "eating" are phonetically similar and reduce the overall recognition reliability. To test this we conducted a post-study trial with 5 participants, in which we asked participants to toggle each activity via voice. One time the phrases in the first row of Table I were used, and the second time row two were used as voice input phrases. The overall recognition rate could be raised from 76.5% to 92%. This leads us to believe that with an increased reliability of the voice recognition, the interaction time can be reduced and annotating activities will become more useful.

The results also match the ones we got from the relevant statements of the questionnaire depicted in Figure 4. As can be seen there, some participants had problems using the Google Glass application, while no participants had problems with the Smartphone, again probably due to the fact that these devices are already commonly known to most people. Also most participants felt that adding annotations could be achieved faster with the Smartphone than with Google Glass, and we attribute that to a non-optimized interface. The recognition rate of the voice input was of major concern by the participants, reflected also in the questionnaire answers.

IV. Conclusion

We described a system (called PART) that allows to effortlessly record ground truth data for sensor data from wearable devices. This system can be used to quickly conduct activity recognition studies. A web-based tool aggregates all sensor data and can be used to modify ground truth data after a recording session, and select and export sensor and ground truth data for later analysis. Future work should optimize the user interface further and include tools to use the interaction with the devices as ground truth data. A combination of EMA studies and activity recognition from wearable devices could lead to a system where the regular sampling of experiences could be replaced by an activity-triggered sampling, i.e. instead of relying the user to explicitly make annotations, asking him wether the currently detected activity was correct.

A personal activity recognition tool (PART) needs to provide the means for a user to explicitly add annotations for activities. For the interaction design we recommened to follow these guidelines (minimize interaction time):

Voice select a keyword set that is seperable by the recognizer, minimize the number of spoken phrases

Touch minimize number of touch operations

Such an interaction needs to be as quick as possible to not interrupt the user in its current workflow. Based on a Smartphone such an interaction will always have the effort of taking it out, and blocking the hands. In contrast a head-mounted displays system, like Google Glass, allows for a hands-free interaction based on voice recognition. However, the presented study showed that Smartphone interaction is preferred for annotating activities. These results needs to be interpreted in the light of a sub-optimal voice recognition performance of 76.5%. Still, even with an increased performance, voice recognition can not always be used, for example in noisy environments or in social settings. Touch interaction needs to carefully designed also, the speed of input for the Smartphone is given by the fact that activities can be directly selected. Such a direct selection is currently hard to implement on the swipe-based interface of Google Glass.

Acknowledgment

This work has been supported by the German Research Foundation (DFG) through the Graduate School on Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments (GRK1362). The authors also like to express their graditude towards the voluntary study participants.

References

[1] A. a. Stone and S. Shiffman, "Capturing momentary, self-report data: a proposal for reporting guidelines.," *Annals of behavioral medicine*, Jan. 2002.

[2] N. Bradburn, L. Rips, and S. Shevell, "Answering autobiographical questions: The impact of memory and inference on surveys," *Science*, 1987.

[3] E. Berlin and K. Van Laerhoven, "Detecting leisure activities with dense motif discovery," in *Proc. of the 12th ACM Conf. on Ubiquitous Computing*, 2012.

[4] P. M. Scholl and K. V. Laerhoven, "A Feasibility Study of Wrist-Worn Accelerometer Based Detection of Smoking Habits," in *esIOT*, 2012.

[5] M. Stikic, K. Van Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," *IEEE ISWC*, 2008.

[6] L. Bao and S. S. Intille, "Activity Recognition from User-Annotated Acceleration Data," *Pervasive Computing*, pp. 1–17, 2004.

[7] M. Keally, G. Zhou, G. Xing, J. Wu, and A. Pyles, "Pbn: towards practical activity recognition using smartphone-based body sensor networks," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, 2011.

[8] M. Berchtold, M. Budde, D. Gordon, H. Schmidtke, and M. Beigl, "Actiserv: Activity recognition service for mobile phones," in *IEEE ISWC*, 2010.