Unsupervised User Observation in the App Store: Experiences with the Sensor-based Evaluation of a Mobile Pedestrian Navigation Application

Benjamin Poppinga, Martin Pielot
OFFIS - Institute for Information Technology
Oldenburg, Germany
{poppinga, pielot}@offis.de

Niels Henze, Susanne Boll
University of Oldenburg, Germany
{niels.henze, susanne.boll}@informatik.uni-oldenburg.de

ABSTRACT
Traditional methods to observe a participant during a field study are often not very scalable and obtrusive. Given the facts of more and more available smart phones and mobile distribution channels, e.g. Apple App Store, the emerging logging observation method gains an increasing attention. In this paper we report on our experiences of conducting a user study in the Android Market by relying on the logging methodology, and thus on sensors of a common mobile smart phone. Based on our preliminary findings we identify the major challenges a researcher needs to face, when an in-market study should be conducted.

1. INTRODUCTION
To observe the mobile user experience various observation techniques exist. For field studies often ethnographic observation techniques, like shadowing, are used. In shadowing an experimenter follows a participant and takes notes on the observed behaviour. Shadowing is known to be highly situated [3, 5]. However, this technique doesn't scale very well. Additionally, because of its obtrusiveness, it might change the observed participant's behaviour.

To overcome the disadvantages of low scalability and high obtrusiveness, new observation methods are developed. In theory, passive automated logging through sensors seems to reach almost the same situatedness, while being scalable and unobtrusive [3, 5]. In practice logging has been rarely applied for mobile observation during the last years. One reason for this might be that suitable data sources, e.g. sensors, were not available on a common mobile device and needed to be self-build [1]. While these self-build sensor systems reduce scalability, they are able to infer users' everyday situations [2].

Nowadays a commercial off-the-shelf mobile smart phone, like the iPhone, has a variety of sensors integrated. Thus, principles were earlier specialized hardware was required, can now be ported to the phone (e.g. a pedometer). McMillan et. al. [4] successfully applied logging in the large scale in a mobile game which they submitted into the App Store. Given all these sensors makes logging more and more interesting as scalable, unobtrusive, and situated observation technique.

However, while there are some well-known concepts, like e.g. a pedometer algorithm, available and ready for instant application, a holistic view on how to use, combine, and apply sensors to log a specific user action is missing. In this paper we present our approach towards unsupervised in-market studies and identify three major challenges based on our preliminary findings.

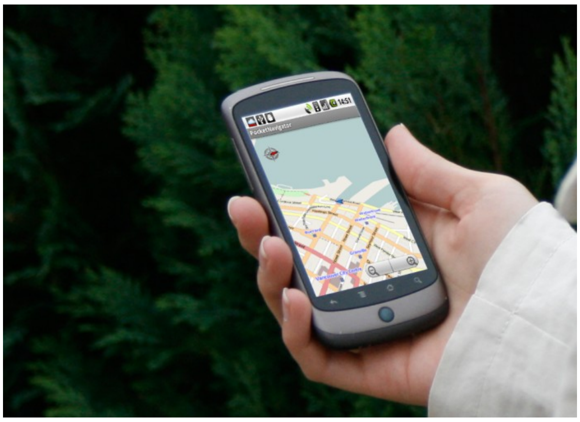

Figure 1: The PocketNavigator is a mobile pedestrian navigation application. Our integrated sensor-based observation technique is invisible for the user. However, the participation within the user study is defined as opt-in to maintain ethical correctness.

2. EXPERIMENT DESIGN
Originating from the interest to provide tactile feedback as additional navigation aid, we developed the PocketNavigator[1]. The PocketNavigator is a personal navigation application, available for free in the Android[2] market (see Figure 1). Designed as traditional map-based application, a map surface, the user's location, and a waypoint-based route towards an arbitrary destination can be provided [6].

However, in addition the application is complemented by a concept that encodes the direction towards the next waypoint in vibration patterns. If the waypoint is straight ahead of the user, two vibration pulses of equal length are shown.

[1] `http://www.pocketnavigator.org/`, last visited August 31, 2010.
[2] `http://www.android.com/`, last visited August 31, 2010.

If the next waypoint is on the right, the duration of the second pulse increases. The same happens to the first pulse, if the waypoint is on the left. If the waypoint is behind a user, three pulses are shown.

The additional values we assumed for the tactile feedback are that a user will need to watch on the display less often, will do less navigation errors, and will be less often disoriented. These three assumptions serve as hypothesis for an experiment we decided to conduct remotely and unsupervised in the Android Market. If a concrete research question should be answered, it is recommended to define the hypothesis right before any sensor data is gathered.

Then, for each hypothesis the observable values need to be identified. Therefore one should think about what are observable events, supporting or not supporting the hypothesis. The own imagination or personal, field-related experience are a good entry point for these definitions. However, often comparable studies in literature already propose a definition how a specific parameter can be observed. In case of the PocketNavigator, we decided to measure e.g. if the user looks at the display be using the roll and pitch angle, as there is no eye tracking available.

In the last step the to be measured values will be assigned and represented through available sensors. In the exemplary case if the user is watching the display we decided to use the accelerometer, which is able to provide the required values roll and pitch. As one can imagine, every matching of an hypothesis to an observable behaviour and then to a set of sensors induces some noise and inaccuracy. Thus it is necessary to design and validate the sufficient representation of a to be observed behaviour iteratively. At some time if the selected representations are reasonable accurate, the experiment can be released to the market.

3. IDENTIFIED CHALLENGES

The PocketNavigator is still available and the study (i.e., the logging) is still ongoing. Until now we can report of 500 people who participated in the study. In this section we transfer our experiences into general challenges which need to be approached to further establish sensor-based observation in mobile applications. We identified three challenges: recruiting, analysis, and the question on internal validity.

3.1 Recruiting

In the participant recruitment process, the very first aspect is that a good application title and description needs to be provided in the market to attract participants. Further, a nice application icon and some screenshots can also attract users. Without question the application should provide the advertised functionality and should be robust and reliable.

To fulfil the ethical requirements of the society or the projects requirements, where the application is developed in, the study needs to be announced to the user in a sufficient and apparent way. Thus, the mentioning of the study in the application's general terms and conditions is ineligible. More, a separate menu entry should clarify the purpose and frame of the study, as a traditional informed consent does. Obviously the participation in the user study should be an opt-in instead of an opt-out. Like in a traditional field study, a user should be able to withdraw at every time.

Early releases of the PocketNavigator presented the study in a separate info view, selectable through the application's menu. If interested in participation, the user must explicitly check a checkbox. However, under this condition the acquisition of participants proceeded quite slow. In an updated version, we proactively announce the study through a simple and short pop up dialog. If the user disagrees to participate in the study, a more detailed info screen on the study is shown, trying to convince the user. This approach leads to a participation rate of about 5 to 10%.

3.2 Data Analysis

The recording of sensor values within the application is one thing. However, the gathered data of each client must be available to do analysis. Therefore we used a custom made server, to which each client connects via sockets and transmits the gathered data in chunks. Alternatively a script, running on an existing server can be used, like e.g. PHP. This can also be easily combined with encryption algorithms, like SSL. To avoid loss of any data, a backup and watchdog is recommended.

Once the application is in the market and the participants are sending their data, it's possible to do some analysis. From our personal experience we recommend to do the analysis on a regular basis, to identify overlooked aspects or strange application behaviours, which can be solved by adapting the logging algorithms. With every adoption it is important to monitor the version a participant is using to not confuse different types of data during analysis.

The actual analysis is done by custom made tools, as universal analysis tools most probably doesn't exist for a specific use case. In case of the PocketNavigator we build one application which does a summary over the data of all participants and prepares an output file, which is readable by e.g. Microsoft Excel, to do some further analysis. Second we build an application which is able to replay the behaviour of an individual user by displaying the values of the sensors in real time. The first tool is more suited for quantitative analysis, while the second tool can give insights in individuals situations, which can be treated as qualitative data.

3.3 Internal Validity

In controlled experiments internal and external validity are two contrasting aims. Internal validity is the validity of the inference of causal relationships, or how confident the observed effects can be attributed to the experimental manipulation. External validity is the validity of the generalisation of experimental findings, or how confident the observed findings can be generalised beyond the experiments setting.

Typically, experiments (especially those conducted in the lab) focus on internal validity. The disadvantage of this approach is that the experimenters often can only carefully generalise their findings to actual usage scenarios. Studying applications in ”real” use by making them available to a wide range of users - as we did with the PocketNavigator - stresses external validity at the expense of the internal validity.

In the case of the PocketNavigator we identified two factors that threaten the internal validity: the design as quasi-experiment and the unpredictable usage.

3.3.1 Experiment vs. Quasi-Experiment

In a true experiment, conditions get allocated randomly. As we are studying the effect of the vibro-tactile feedback technique, in a true experiment, half of the participants would be chosen to use the tactile feedback and the other half not.

However, in our actual study design we allowed the participants to choose for themselves if the tactile feedback should be turned on or off. We were afraid that people get annoyed by the tactile feedback, giving the application bad ratings in the Android Market, and in consequence deterring potential future users.

Thus, the experiment is not a true but a quasi experiment. Due to the lack of randomization it is harder to rule out confounding variables and unsystematic variance. In our case, people that decide to use the tactile feedback could have certain traits or be in certain situations which favour or disfavour the usage. For example, if only people with lots of experience use the tactile feedback, because they are more open to new innovations, their navigation performance could be disproportionally better than average because of either their experience or the tactile feedback.

3.3.2 Unpredictable Usage

Another problem that turned up is the unpredictable usage of the application. In a typical experiment the task is well-defined and well-known to the person analysing the data. In the case of the PocketNavigator we neither have a way to dictate a certain usage pattern to the users nor can we completely understand the usage at a certain time. In the following we give a few examples of unpredicted usage patterns that could have threatened the internal validity if we had not identified them:

Example 1: Lying on table. In the first stream of data we received from our participants we had many situations where no navigation at all took place. Having a close look at the data, the accelerometer indicated that the device was oriented parallel to the surface and the GPS signal showed no walking speed. From these data we inferred that many users might be testing the application indoors first, leaving the device on the table and probably keep running the application in the background.

Example 2: Car Driving. At a later stage we were investigating the effects of the tactile feedback on the average walking speed. However, we were surprised by the huge variance in the walking speed averages. Taking a closer look at the individual data we found that some walking speeds were unnaturally high (e.g. $>$ 70km/h in average) for pedestrians, so we inferred that people had used it in their cars or any other vehicle.

Example 3: Background idling. Android offers parallel and background executing. As the PocketNavigator is expected to run in the pocket we designed it to continue running when the screen saver is activated or another application is pushed to the front. The problem is that the Android OS does not really terminate applications but only pushes them into the background until the resources are needed otherwise. Thus, in a few cases the application kept running in the background producing nonsense data.

4. CONCLUSION

In this paper we report on our experiences on applying a sensor-based virtual observer to the Android Market. We identify three major issues, which need to be considered and approached in future developments: recruitment, data analysis, and internal validity.

In our future work we want to extend and apply the in-market observation methodology for true experiments, as well as for more open research questions, which can not be answered within an experiment. Additionally we want to apply logging as observation method in a traditional field study to prove the validity of the method. Finally we are interested in the advantages, disadvantages, and limitations of the virtual observer in different settings.

5. ACKNOWLEDGEMENTS

The authors are grateful to the European Commission which co-funds the IP HaptiMap (FP7-ICT-224675). We like to thank our colleagues for sharing their ideas with us.

6. REFERENCES

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

[2] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay. Activity sensing in the wild: a field trial of ubifit garden. In *Proc. of CHI*, 2008.

[3] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proc. of MobiSys*, pages 57–70, New York, NY, USA, 2007. ACM.

[4] D. McMillan, A. Morrison, O. Brown, M. Hall, and M. Chalmers. Further into the wild: Running worldwide trials of mobile systems. In *Proc. of Pervasive*, volume 6030 of *LCNS*, pages 210–227, Helsinki, Finland, May 2010. Springer.

[5] I. Mulder, H. ter Hofte, and J. Kort. Socioxensor: Measuring user behaviour and user experience in context with mobile devices. In *International Conference on Methods and Techniques in Behavioural Research*, 2005.

[6] M. Pielot, B. Poppinga, and S. Boll. Pocketnavigator: Vibro-tactile waypoint navigation for everyday mobile devices. In *Proc. of MobileHCI*, 2010.