Experiments in the Wild: Public Evaluation of Off-Screen Visualizations in the Android Market

Niels Henze
University of Oldenburg
Oldenburg, Germany
niels.henze@
uni-oldenburg.de

Benjamin Poppinga
OFFIS - Institute for
Information Technology
Oldenburg, Germany
poppinga@offis.de

Susanne Boll
University of Oldenburg
Oldenburg, Germany
susanne.boll@
uni-oldenburg.de

ABSTRACT

Since the introduction of application stores for mobile devices there has been an increasing interest to use this distribution platform to collect user feedback. Mobile application stores can make research prototypes widely available and enable to conduct user studies "in the wild" with participants from all over the world. Previous work published research prototypes to collect qualitative feedback or to collect quantitative attributes of specific prototypes. In this paper we explore how to conduct a study that focuses on a specific task and tries to isolate cause and effect much like controlled experiments in the lab. We compare three visualization techniques for off-screen objects by publishing a game in the Android Market. E.g. we show that the performance of the visualization techniques depends on the number of objects. Using a more realistic task and feedback from a hundred times more participants than previous studies lead to much higher external validity. We conclude that public experiments are a viable tool to complement or replace lab studies.

Categories and Subject Descriptors

H.5.2 [**Interfaces and Presentation**]: User Interfaces - Interaction styles

General Terms

Design, Human Factors, Experimentation

Keywords

Experiment, game, off-screen, Android Market, map navigation

1. INTRODUCTION

With the introduction of mobile application stores such as Apple's App Store and Google's Android Market a new way to conduct user studies became available to the mobile HCI research community. The Android Market, in particular, enables to publish an application in a few minutes without any review process. By publishing applications in mobile application stores, researchers benefit from a potential worldwide audience. They gain access to participants with various cultural backgrounds and different contexts.

We assume that those public studies (i.e. studies where virtually everybody can participate) can complement the common HCI lab study. The external validity of public studies can be much higher than lab studies which are usually affected by a lack of resources. E.g. the number of participants is low (e.g. n<20), participants have the same background (because they are students and colleagues from the lab), and are of similar age. The mobile HCI community, for example, usually conducts studies in the lab even though a mobile or natural context would influence the outcome [5].

So far, public studies that exploit mobile application stores are used to collect qualitative feedback [12] or usage data for a particular prototype [8]. We assume that public studies can also be used for experiments that try to isolate cause and effect much like controlled experiments in the lab. Just as controlled experiments in the lab are often advantageous, public studies using mobile application stores have their very own advantages which have not been explored yet.

In order to get a better understanding of experiments "in the wild" we use the well-defined "off-screen problem" to provide a proper illustrative background. Most off-screen visualizations have been developed to show the position of geographic objects which are currently beyond the segment of a digital map that is visible on the screen. In the following we describe the design and the results of a public study that compares three visualization techniques (see Figure 1).

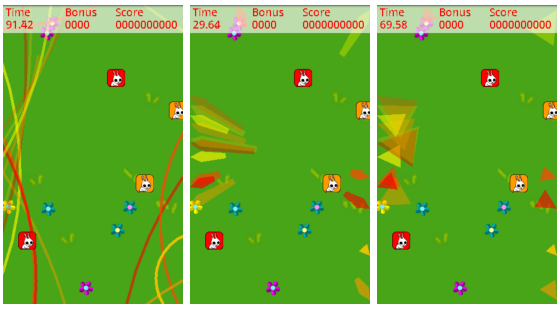

Figure 1: In-game screenshots of the three visualization techniques Halos, stretched arrows, and scaled arrows.

NordiCHI2010 October 16-20, 2010, Reykjavik, Iceland.

Using a game published to the Android Market we show that the visualizations are more or less suited depending on the number of shown objects. We assume that the external validity of the conducted study is much higher compared to similar controlled lab experiments and conclude that games are a viable tool to conduct public experiments.

2. RELATED WORK

Visualizing off-screen objects has received some attention for interaction with digital maps on mobile devices. Zellweger et al. [11] introduced City Lights, a principle for visualizing off-screen objects for hypertext. An extension of the City Lights concept for digital maps is Halo [1]. For Halo circles that intersect the visible area shown on the device's display are drawn around the object. Users can interpret the position of the POI by extrapolating the circular arc. Baudisch et al. showed that users achieve better results when using Halo instead of arrows with a labelled distance [1]. Burigat et al. [2] reviewed these results by comparing Halo with different arrow types e.g. by visualizing distance through scaling the arrows. They found that arrow-based visualizations outperform Halo, in particular, for complex tasks. Other off-screen visualization have been developed (e.g. Wedge [3]) but it has not been shown that these outperform existing approaches. The previous work conducted studies with static maps that participants had to interpret. E.g. they did not consider tasks where users can dynamically interact with the map by panning it. Furthermore, our knowledge about off-screen visualization techniques is based on studies conducted with less than 17 participants which share similar backgrounds (e.g. computer scientists).

Controlled experiments are the tool of choice to test hypothesis. E.g.: Users archive a higher performance using Halos than using scaled arrows. HCI researchers (even mobile HCI researchers [5]) usually conduct those studies in the lab. However, conducting studies in the field can reveal unforeseen aspects and Nielsen et al. argue that studies in the field are actually "worth the hassle" [9]. Supervised studies, in particular field trials, are expensive in terms of resources. Therefore, the number of participants is usually low and they often share a similar background. These aspects limit the external validity and make the results less generalizable.

An approach to overcome these limitation is successfully used to compare different variations of websites [6]. Because a large number of users is needed this technique is only used to answer specific questions interesting for a particular high frequent website. In contrast researcher recently began to exploiting mobile application stores (e.g. Google's Android Market or Apple's App Store) to gather feedback from a larger number of users. Pielot et al. report that they started the evaluation of a tactile navigation system by publishing the system in the Android market [10]. Zhai et al. published a text entry application for the iPhone and reported from 556 reviews written about their system [12]. McMillan et al. report from a very large scale study involving almost 100.000 user with the aim "to push the upper limit on the number of participants as far as [they] could while still combining quantitative and qualitative approaches in ways that usefully and efficiently fed into the redesign process" [8]. To our knowledge, however, previous work (beside our own preliminary work [4]) conducted formative studies that do not allow the identification of cause and effect.

3. DESIGN AND APPARATUS

In order to conduct a public experiment that tries to isolate cause and effect we selected the visualization of off-screen object on mobile devices as our domain. We identified the following research questions that have not been answered before:

- How do different techniques for visualizing off-screen objects perform in a more realistic (i.e. interactive) task that involves panning the background.
- How do the visualization techniques scale if the number of shown objects increases?
- How easy are the visualization techniques to learn and do users understand the meaning of the respective visualization without lengthy instructions?

Different off-screen visualizations have been proposed. In order to make our results comparable with previous work (e.g. [1, 2]) we decided to compare the three visualization techniques Halos, stretched arrows, and scaled arrows shown in Figure 1.

To compare the three visualization techniques we aimed to conduct a "controlled" experiment. This leads to the three conditions Halo, stretched arrows and scaled arrows. A repeated measurement design reduces the effect of the individuals compared to an independent measurement design. In a public experiment one cannot control important aspects such as the selection of participants, used devices and the participants' context which is why we decided for a repeated measurement design. In order to investigate the scalability of the visualization techniques multiple tasks with different numbers of objects are used.

It is crucial for public studies to motivate people to participate. Even though the visualizations have been designed for maps it would be difficult to force a mobile user looking for a hotel to repeat the same task with a different visualization technique. Therefore, we decided to use a mobile game which enables to naturally confront participants with variations of the same task. Thereby, it can be assured that participants repeat the same tasks while only the independent variable (i.e. the visualization technique and the number of visualized objects) is varied. However, as the game has to be installed and played by users at their own will it is necessary to find a balance between validity of the study and fun of the game.

We decided to use an increasing level of difficulty to motivate players. A game starts with a stage of three levels each containing 30 objects, represented by "cute" rabbit icons. The objects are randomly distributed on plane that can be

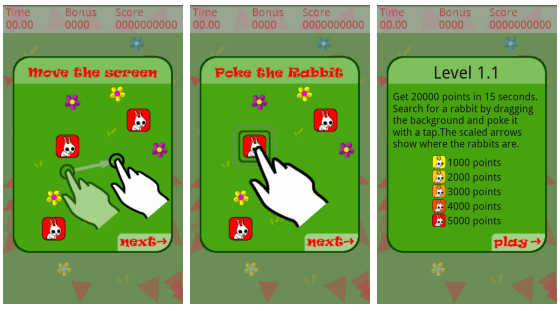

Figure 2: Screenshots of the intro screens.

paned much like a digital map. Each level uses a different offscreen visualization (see Figure 1). The task of the player is to "poke" as many objects as possible by tapping them with the finger in a certain time frame. Once an object is poked it fades to gray and a new object appears. If a player finishes the three levels he or she goes to the next stage where 20 objects are used and afterwards to a stage with 10 objects. The visualizations are randomized within a stage to reduce sequence effect. After finishing three stages the game starts from the beginning with more time to complete a level but also with more objects needed to successfully finish a level.

We implemented the game for the Android platform[1]. The visible area covers the same fraction of the complete field on different devices by scaling a fixed fraction to the whole screen. It is slightly affected by different devices' aspect ratio. A short explanation (see Figure 2) is shown each time a game is started. Furthermore, the player gets scores each time a rabbit (i.e. object) is tapped. A bonus is added if the player taps multiple rabbits in a row. To increase the motivation we implemented a local and a global high score list which can be accessed from the main menu. Furthermore, we added music that is played during the game. Each time a level is finished the number of tapped rabbits and the particular level is transmitted to our server. We also log the device's time zone, the selected locale, the device's type, and an anonymized device id.

4. USER STUDY

The describe game was published in the Android Market on the 14th of April 2010. We did not actively advertise the game among our friends and colleagues. In the following we report the results derived from data collected until the 25th of June 2010. According to the statistics provided by Google it has been installed approx. 5000 times. In total we collected samples from 3934 accounts. These samples came from 40 different types of devices. The devices cover most of the diversity of the currently available Android phones. E.g. the most frequent Sholes (alias Motorola Droid) runs Android 2.1 and has a 3.7" (854x480px) screen while the second most frequent HTC Hero running Android 1.6 has a 3.2" (480x320px) screen. The most frequent locale is en_US with 68.3%. In total English locales accounted for 76.5% and more than 92.3% have a western language. While users can freely select the used locale the results are very consistent with the observed time zones.

4.1 Results

We analyzed the effect of the visualization technique on the players' performance if different numbers of rabbits are present. Since different levels have different durations we normalized the number of poked rabbits to "hits per minute" (hpm). Furthermore, we pre-processed the raw data by removing incomplete samples and samples where players did not poke a single rabbit. The analysis of variance shows that the visualization technique significantly affected the players' performance for 30, 20, and 10 rabbits (all $p<.05$). The average performance is shown in Figure 3. With 30 rabbits and using scaled arrows (ø=38.41hpm) the players archived a higher performance (both $p<0.01$) than using halos (ø=37.33hpm) or stretched arrows (ø=37.26hpm).

[1]An updated version of the game can be found in the Android market by searching for "net.nhenze.game.offscreen".

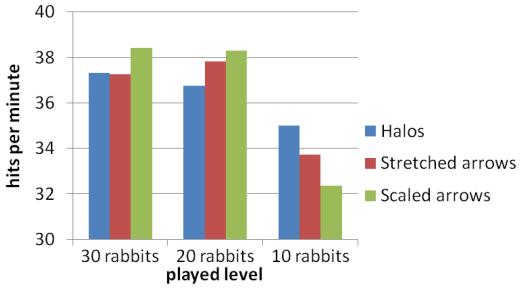

Figure 3: Performance for different numbers of objects.

When 20 rabbits are used players achieve a lower performance with halos (ø=36.75hpm) than with stretched arrows (ø=37.82, $p<.05$) or scaled arrows (ø=38.29, $p<.01$). If only 10 rabbits are used the order of the visualizations is reversed. If using Halos (ø=35.33) players perform better than using stretched arrows (ø=33.52, $p<.001$) or scaled arrows (ø=32.18, $p<.001$). The difference between stretched arrows and scaled arrows is also significant ($p<.05$).

We expected that the learning curve for the three visualizations differ. In particular, we assumed that the arrow based visualizations are more intuitive and novice players perform better with those techniques than with halos. The design of the experiment does not allow a systematic analysis. However, the players' performance after playing a respective number of levels shown in Figure 4 suggest a general tendency. The trend lines of the three techniques are very similar and we therefore assume that their learnability is also surprisingly similar.

Due to the nature of the study we could not control which device the participant uses. The large number of different devices (40) makes Type I errors (i.e. we believe that there is an effect, when in fact there is not) very likely if we do a pair wise comparison of all devices. Furthermore, the numbers of samples from the devices are very different and devices with a low number of samples should not be considered. In addition, it is possible that players with a low performance (partly induced by the used device) quit playing the game early which would make the differences between devices look larger than they actually are. As we did not define a procedure beforehand (e.g. how many samples are needed from each device) it is likely that extensive results would be error-prone. Therefore, we only compared the two

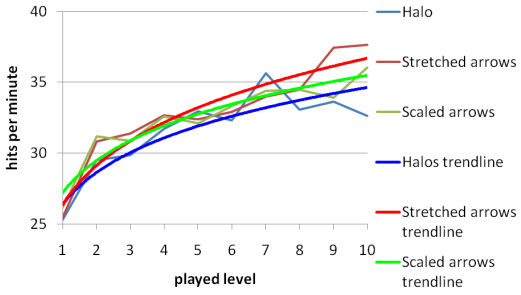

Figure 4: Average performance after playing a particular number of levels. Only samples where players poke at least one rabbit are considered.

most often observed devices. The average hits per minute for the Sholes is 39.37hpm (n=2205) and 34.57hpm for the HTC Hero (n=1134). Even with a conservative Bonferroni correction the difference is significant ($p<10^{-9}$)

4.2 Discussion

In summary, the results show that the visualization techniques scale differently. For 30 objects scaled arrows are more suitable and for 10 objects player perform better with Halos. The difference between the visualization techniques regarding learnability is presumably small. As expected, the used device does affect the players' performance.

For a large number of objects the results are consistent with literature results described for complex tasks and a low number of objects [2]. In contrast, our results suggest that Halos perform better than the arrow-based approaches for a low number of objects. This, is consistent with [1] which used a very low number of objects to compare Halos and arrows with labelled distances. However, our study analyzed the effect of the off-screen visualization if the user dynamically interacts with the objects while in previous studies the participants used static maps and more complex tasks. Thus, our results are particularly relevant for systems with a high interactivity.

The study treated internal validity for external validity. Due to the large number of participants with different background, devices, and contexts our results are more generalizable than studies involving 12 [1] or 17 [2] participants, which use the same device, perform the tasks in the same room, and live in the same region. Even though we tried to address users from all over the world most players originate from the US or at least from a western country. It might be possible to attract more players from other cultural backgrounds by internationalizing the game and its description in the Android Market. The experiments internal validity is limited because we had little control over external factors and the data is heavily affected by noise. This is one of the reasons why we can conclude little about learnability and differences between devices.

5. CONCLUSIONS

This paper described a public experiment with thousands of participants that compares three visualization techniques for off-screen objects. It is shown that the performance of the visualizations depends on the number of objects. We showed that public experiments can successfully be used to answer research questions. Even though the results are affected by noise we assume that the study has a much higher external validity compared to experiments in the lab.

We use a game to motivate player to take part in the study. While games have been used before to study HCI questions (e.g. [7]) and are widely used in psychology we assume that games are especially useful to study cause and effects in public experiments. Games naturally allow a repeated measurement design and task repetition. Compared to applications games can abstract from real world tasks. Therefore, games are often easier to design and the results are less affected by additional functionalities that are needed to make an application useful but do not contribute to or affect the results.

Like lab experiments public experiments must be carefully designed. The high number of participants does not help to overcome design flaws. E.g. we cannot conclude much about learnability because this aspect has not been addressed in the design carefully enough. The large amount of collected data entices to compare every possible combination of variables (e.g. the 40 different device types we observed). If these tests are not defined before starting the experiment and if the statistical tests are not used in a precise way, this will likely lead to false results.

In our future work we will use public experiments to study further research questions. In particular, we believe that public experiments are a viable tool to replicate lab studies to validate their results at a larger scale. In order to determine advantages and disadvantages of different apparatus designs (e.g. games, interactive tutorials and fully-fledged applications) we will use these different designs to study similar questions.

Acknowledgments This paper is supported by the European Commission within the projects InterMedia (FP6-038419) and HaptiMap (FP7-224675). We thank Sascha Hornauer and Fadi Chehim for their support.

6. REFERENCES

[1] P. Baudisch and R. Rosenholtz. Halo: a technique for visualizing off-screen objects. *Proc. of CHI*, 2003.

[2] S. Burigat, L. Chittaro, and S. Gabrielli. Visualizing locations of off-screen objects on mobile devices: a comparative evaluation of three approaches. *Proc. of MobileHCI*, 2006.

[3] S. Gustafson, P. Baudisch, C. Gutwin, and P. Irani. Wedge: clutter-free visualization of off-screen locations. *Proc. of CHI*, 2008.

[4] N. Henze and S. Boll. Push the study to the app store: Evaluating off-screen visualizations for maps in the android market. *Proc. of MobileHCI*, 2010.

[5] J. Kjeldskov and C. Graham. A review of mobile hci research methods. *Proc. of Mobile HCI*, 2003.

[6] R. Kohavi, R. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. *Proc. of SIGKDD*, 2007.

[7] J. Looser, A. Cockburn, J. Savage, and N. Christchurch. On the Validity of Using First-Person Shooters for Fitts' Law Studies. *Proc. of British HCI*, 2005.

[8] D. McMillan, A. Morrison, O. Brown, M. Hall, and M. Chalmers. Further into the Wild: Running Worldwide Trials of Mobile Systems. *Proc. of Pervasive*, 2010.

[9] C. M. Nielsen, M. Overgaard, M. B. Pedersen, J. Stage, and S. Stenild. It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. *Proc. of NordiCHI*, 2006.

[10] M. Pielot, B. Poppinga, and S. Boll. PocketNavigator: Vibro-Tactile Waypoint Navigation for Everyday Mobile Devices. *Proc. of MobileHCI*, 2010.

[11] P. T. Zellweger, J. D. Mackinlay, L. Good, M. Stefik, and P. Baudisch. City lights: contextual views in minimal space. *Proc. of CHI*, 2003.

[12] S. Zhai, P. Kristensson, P. Gong, M. Greiner, S. Peng, L. Liu, and A. Dunnigan. Shapewriter on the iPhone: from the laboratory to the real world. *Proc. of CHI*, 2009.