An Evaluation Framework for Mobile User Interfaces

Marco de Sá and Luís Carriço

LaSIGE & Department of Informatics, University of Lisboa, Portugal
{marcosa,lmc}@di.fc.ul.pt

Abstract. We present the extensions to a software framework which supports the construction and evaluation of mixed-fidelity prototypes for mobile devices. The framework is available for desktop and mobile devices and allows designers and users to test the prototypes on actual devices. Additionally, the extensions aim at allowing designers to gather usage information, both passively and actively, tailoring the used techniques to users or project's goals. It supports contextual and ubiquitous evaluation also including in-situ prototyping and participatory design on-the-go. We address the evaluation's features and their contribution to the field of mobile interaction design, presenting real-life case studies and achieved results.

Keywords: Mobile devices, In-Situ Evaluation, Ethnography, Prototyping.

1 Introduction

Designing for mobile devices is an increasingly demanding challenge. Besides the hardware constraints that are imposed by their size, interaction modalities, diversity and portability, their pervasiveness and multi-purpose functionality imply an entire new set of usage paradigms. Accordingly, specific design methods and techniques for mobile devices have been emerging and receiving special attention by researchers, leading to the appearance of different approaches for a wide range of problems [1,10,11,21,24]. Given the differences from fixed technologies, most efforts are directed to prototyping and evaluation. These suggest the need for detailed and carefully built prototypes that offer a more resembling picture of final solutions and their characteristics [9,19] , even at the earliest stages of design. Another important benefit that results from using such techniques and carefully built prototypes is the ability to use them in realistic settings. In fact, the majority of research results within this area point the necessity of conducting evaluation out-of-the-lab [1,7,8,17], on the actual settings where the final applications will be used, since the beginning of the design process [24]. Besides providing users with more realistic experiences, this approach allows designers and evaluators to see users interact with the devices within scenarios that will likely be real settings for the final applications. This leads to the detection of more design and usability issues and allows for a better understanding of how users interact with the applications in the real-world [24].

However, despite the increasing amount of attention towards in-situ contextual evaluation, support for this type of procedure is still relatively scarce [12,24]. Existing examples usually point guidelines on how to emulate real world settings within labs

[1,21] or conduct role playing simulations. This contrasts with the aforementioned solid evidence and body of work that emphasize the advantages of out-of-the-lab evaluation [4,8,11,17,19]. As some experiences have shown, the lack of in-situ real world evaluation is a result of the relative inexistence of appropriate techniques to support such endeavor [12,24] and, when adequate techniques are available, the added effort that these require.

These issues led us to conduct a series of experiments with low-fidelity prototypes and ubiquitous data gathering and evaluation techniques [24]. Throughout these, the learnt lessons pointed directions and techniques that proved to be adequate while supporting mobile design and, in particular, mobile evaluation [24]. Nevertheless, despite the exciting results, some of these techniques still required added effort and placed a heavy burden on users. Naturally, this provided the motivation, which, propelled by the inspirational results, set the ground for the design of tools to support and share ways on how to apply, on an easier fashion, the same techniques.

In this paper we present the evolution and added features of a framework that acknowledges the inevitable differences between fixed and mobile technologies and respective design requirements and aims at supporting mobile design. The framework provides the necessary features to support the creation and design of mixed fidelity prototypes and their usage on real devices for testing sessions with real users on real locations. However, now, and more importantly, it has been updated to cover the evaluation of mobile applications and design experiments in realistic settings using the set of techniques that had a strong and positive impact on the aforementioned experiences. Our contribution in this paper is the description of how the experiences and techniques that inspired the framework's upgrade were materialized and integrated with the prototyping stage, allowing designers to tailor the evaluation process to the project's need or the target end-users.

We discuss how the combination of emerging ubiquitous evaluation techniques were introduced within our tool, supporting shorter design cycles and experimentation at very early stages, mitigating the burden placed on users and designers. We succinctly describe the overall framework and its goals focusing on the evaluation features that it integrates. We proceed with the description of some of the case studies in which the framework has been used, highlighting the impact it had on the design process. Finally, we discuss results, present conclusions and draw future goals.

2 Inspiration/Motivation

The evaluation of user interfaces and interactive applications is a crucial part of the design process. It provides the opportunity to understand if the user interface is suitable to be used while users are trying to achieve certain goals within a certain context, opening doors for corrections and improvements. When it comes to mobile devices, the settings and contexts in which users interact with the applications while trying to get something done, may vary infinitely, depending on the device, the user's location and surroundings (e.g., lighting conditions, noise, social context, posture). Moreover, some activities span through various settings, which might affect how users interact with the device/application, even during a set of actions that lead to one

specific goal [24]. Consequentially, evaluation gains even more relevance in a mobile context and, as many experiences, including our own, have pointed, benefits greatly if undertaken in-situ, out of the lab, in the real world [17,19,21,23,24].

However, this introduces additional challenges, which are still difficult to overcome. The techniques that are usually applied for fixed technologies (e.g., user observation, WOz simulations) are demanding when applied away from laboratory settings, on moving contexts, sometimes even inadequate and frequently avoided [12,24]. Our experiences demonstrated that prototypes need to be realistic and to provide a tangible experience. Methods that require the designer's presence hinder the process by affecting the experience and because of the time/effort they imply. Nevertheless, techniques such as the Experience Sampling Method (ESM)[4] and diary studies[22], and even usage registration (e.g., actually writing and marking clicks/taps on the cards when interacting with them, provided good results. However, even these require user's cooperation and can be demanding during real-world tests.

Recent works, addressing these issues and aiming at further improving and supporting mobile evaluation, have introduced technological methods to gather usage data remotely through active (e.g., , ESM, Diary Studies) and passive modes (e.g., Logging), enabling evaluation on real settings. For instance, with close goals to our framework regarding evaluation, the Momento [3], and the MyExperience [8] systems provide support for remote data gathering. The first relies on text messaging and media messaging to distribute data. It gathers usage information and prompts questionnaires as required, sending them to a server where an experimenter manages the received data through a desktop GUI. On the second, user activities on Mobile Phones are logged and stored on the device. These are then synchronized depending on connection availability. The logging mechanism detects several events and active evaluation techniques can be triggered according to contextual settings. The Topiary system [15] follows a similar approach and aims at supporting the evaluation of context-based applications.

Our approach also addresses these issues and takes into account previous results, but inspired by lessons learned and the available literature [7,11,17,19,24] aims at integrating the in-situ prototyping and evaluation stages seamlessly, facilitating user involvement and the design process. None of the existing work integrates the in-situ prototyping and evaluation on real devices, also including means to adjust the prototypes while evaluating them or to analyze them (e.g., various alternatives to one user interface), individually or simultaneously, on intuitive and simple to review modes. Furthermore, most are focused on specific issues or domains and depend on server-client architectures, requiring a constant connection or frequent synchronizations. Additionally, none combines various techniques into the same tool, allowing designers to tailor evaluation to their needs.

Our goal is to focus the interactions that directly relate to prototypes on early design stages, facilitating their on-the-spot analysis through several mechanisms and to combine them with qualitative data easily gathered by end-users. For this, the evaluation methods are automatically integrated within a mixed-fidelity prototyping tool that conveys the two stages, bridging the gap between prototyping and in-situ evaluation. Our contribution over previous work is the integration of several data gathering and analysis techniques that can be used seamlessly on mobile devices, facilitating user involvement and the design process. We support the creation, use and

evaluation of prototypes on real devices, also including means to adjust them (e.g., low and high-fidelity with video and audio) in-situ, while evaluating or analyzing them (e.g., various alternatives to one user interface), on a video-like mode.

2.1 Background

To cope with early design stage difficulties, especially during prototyping and evaluation, the framework's features cover both stages, supporting an iterative and participatory design that facilitates the transition between the two stages. Its umbrella goal is to support the early design of applications for mobile devices. Rather than including complex features and supporting the creation of very detailed software prototypes, we aim at supporting rapid prototyping and in-situ evaluation, focusing techniques and features that have been proved to be highly effective in these settings [24]. Like existing tools [6,13,14] it provides designers with means to quickly create prototypes. At this stage, the framework's main features allow designers to create mixed-fidelity prototypes (e.g., sketch-based or visual component based), targeting different concerns, and configure their behavior (e.g., navigation between screens/cards, play audio and video files; allow end-users to interact with traditional components – text-boxes, combo-boxes). These prototypes can be easily copied into a mobile device where end-users can interact with them [25].
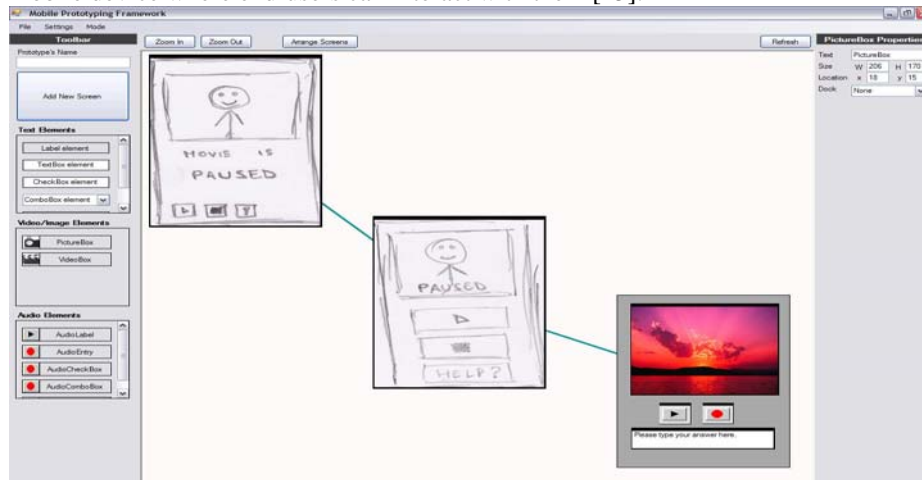

Fig. 1. Prototype editor. Cards/Screens can be dropped and moved inside the working area and composed by sketches (the first two screens) or interactive elements (right bottom – with a video container, two buttons and a text-box).

Additionally, direct prototyping on the mobile devices is also supported, enabling designers and end-users to update and re-arrange simple details, improving the prototypes on real settings, out of the lab. The major advantage of using actual devices, with components and interaction modalities available to the used device, is that problems regarding the device's characteristics (e.g., size, weight, screen

resolution, shape) emulation are solved [19], avoiding the cargo cult syndrome [9] (e.g., misleading users) and providing end-users with a much more tangible and realistic usage experience and better evaluation results.

Figure 1 depicts the framework's prototyping interface. Designers can visually create each screen/card using sketch based images and augmenting them with behavior (e.g., defining rules - click areas emulating buttons) which allow end-users to navigate through the prototype (without the user acting as a Wizard-of-Oz) on the mobile device. Alternatively, or in concert with sketches, graphical components can be dragged and dropped on each card (e.g., picture-boxes, labels, buttons, text-boxes, track-bars). Each prototype's specification is saved into an XML file. The counterpart for the prototyping tool (Figure 1) is the runtime environment that is responsible for recreating the prototypes on the targeted devices (Figure 2, left). Currently we have a runtime environment for Windows Mobile, Palm OS and SymbianOS. A Windows version was also created to allow testing on TabletPCs.
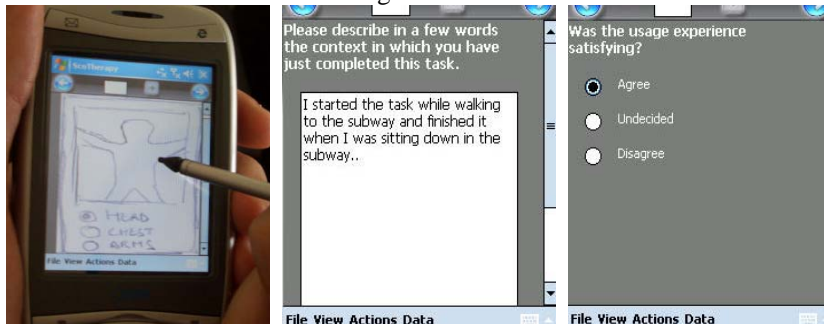

Fig. 2. a) User with an interactive low-fidelity prototype on an actual mobile device. b) Active data gathering. Questionnaires can be completed and browsed on the device.

3 Extensions and Requirements

In order to facilitate data gathering during real world evaluation sessions, using the techniques we applied on to the preceding experiences, our framework had to be extended, combining different approaches that could be used in concert or alternatively. Additionally, means to easily analyze the gathered data, even on-the-spot, during in-situ trials, were necessary. Overall, these goals can be supported by (a) retrieving reliable usage information without intrusive equipment, without the designer or usability engineer's presence and using seamless/passive techniques; (b) supporting the analysis of usage patterns and usability concerns through the visualization of the user's activities; (c) integrating alternative methods that provide qualitative data extending the scope of the evaluation process.

These goals emerge as a consequence from the framework's prototyping features and successful experiences they provided [24]. Directly embedding these techniques on the developed prototypes, created with the framework, should facilitate the overall design process flow. Globally, this can be achieved through the following features:

1.   Gather data through passive and active techniques. On the former, every action that the user takes is automatically logged with customized granularities. On the latter, the use of methods such as probing, ESM [4] and diary studies [22], integrated within the tool, provide another source of data and usability information.

2.   Contextual card tagging and sorting. Each screen/card can be tagged and numbered in-situ, defining a sequence that fits the user's needs in a particular setting/context. This allows designers to select optimal configuration and arrangements for the applications or user interfaces that are being designed.

3.   Analysis functionalities - The analysis of gathered data is paramount during the design and evaluation process. Accordingly, we include data analysis features on the desktop editor and on a mobile version for in-situ evaluation/analysis.

4.   Multimodalities are a key factor on mobile devices, providing redundant output/input channels that cope with the adversities of the usage context or offer support for disabled users. Since our framework supports the prototyping and simulation of multimodal applications (e.g., including voice capturing, audio and video elements) these modalities can play a paramount role for evaluation purposes.

4   Tailoring the In-Situ Evaluation

Besides combining the aforementioned techniques and integrating them with the prototyping features, particular care was directed into facilitating their utilization and adjustment. The following sections detail how they were applied and enhanced.

4.1   Passive Techniques – Intelligent Logging Engine

Integrated within the runtime environment, the tool in which users can interact with the prototypes, there is a logging engine which is responsible for the passive data gathering (i.e., without the user's explicit intervention) mechanisms. Its goal is to collect usage data without any intervention or even awareness by the user. Accordingly, it stores every event that is triggered by the user's interaction with the prototype and device or by the time constraints associated to each element/screen. Events range from each screen tap, each button press or even each character that was typed. Events are saved with a timestamp, allowing its reproduction for the re-enactment of the usage behavior. Details such as the type of interaction, location of the screen tap, etc., are also stored for every event. In summary, it generates a detailed and structured description of every occurrence while the prototype was used.

However, considering mobile devices' limited memory and battery, the granularity of the logged events can be easily configured both during usage and during the prototype's construction. Limiting the amount of logged events reduces the size of logs and the processing that logging requires. Moreover, the adjustment of the logging granularity can also be used to match the gathered data with evaluation and analysis' purposes. For instance, if the evaluator is particularly interested in understanding how the user navigates between the existing screens that compose the prototype, but has no interest in collecting data regarding the locations of taps on the screen, the latter event can be ignored. This mechanism creates logs that are focused to particular events.

These configuration options are also important when taking into account the several modalities that are available within a particular prototype. Here, the selection of events, which pertain to specific modalities (e.g., play, pause of an audio track or video, voice commands) is also paramount in order to facilitate analysis of usage logs (e.g., if the application is to be used by a visually impaired user, the designer might enable logging only for the audio modality and the user's voice commands). Overall, the logging engine supports configurable data gathering that can be focused on the evaluation goals without adding effort or requiring users' intervention.

4.2 Active Techniques

To support active data gathering users, while interacting with the prototypes assume an active role and are responsible for providing usage or context information for posterior analysis. This type of data gathering has been widely used on mobile devices, through techniques such as the ESM [4] or Diary Studies [23]. The main medium utilized to gather information with these techniques is questionnaires. For the former, at particular times, users are required to fill-in a questionnaire, responding to questions that pertain to the action that they are performing. On the second (diary studies) users are free to annotate any information that they find relevant such as their location, surroundings, etc. This provides qualitative data regarding the usability of the system that is being used. Naturally, users are usually required to carry paper questionnaires along with them while performing their activity, which often leads to users not remembering to complete the questionnaires or hinders the activity at hands [25]. Moreover, ESM questionnaires are generally directed to specific issues and diary studies are strongly dependent of what users find meaningful, which many times is not what designers are interested in. Here, the digital support plays a key role since questionnaires can be prompted automatically when necessary [26].

To facilitate and enhance these two techniques, the framework supports active data gathering by offering means for designers to create and include, within their prototypes, questionnaires that can be completed by users during in-situ evaluation sessions. Using the same mechanism and interactive elements that are used to build the high-fidelity prototypes, questionnaires can be easily configured to focus the details and goals of the evaluation or target users and include whichever questions designers find relevant.

Tailoring the Experience Sampling Method. Taking advantage of a behavior engine and respective conditions and actions, the framework provides means for designers to define specific conditions or settings in which these questionnaires can or should be presented to users. This technique, if well used, provides support for intelligent ESM and pro-active Diary Studies since usability questionnaires can be prompted according not only to time, but also location or behavior triggers. For instance, if the user misses a specific screen location or button several times, tapping a nearby location, or is taking more than 1 minute to respond to a question, a questionnaire can be automatically popped up. Here the end-user can be requested to explain the reason behind the low accuracy or describe the setting in which s/he was working (e.g., while jogging, seating on the sofa at home). Also, if targeting navigational issues or the application/user interface's structure, questionnaires can be

configured to appear when certain cards/screens are reached (e.g., after the "send SMS" interface). Rule/behavior definition is wizard-based and does not require programming knowledge.

Enhanced Data Gathering. The various input and output modalities that the framework offers allow designers to gather data in diverse formats. For instance, when a questionnaire is popped, and in order to allow the user to continue with his/her activity while completing it, instead of typing the answers on the device, these can be given by voice, depending on the user's activity or preference. Besides increasing the flexibility and ease of responding to usability questionnaires it also provides richer data allowing designers to have an idea of the environment in which the user is interacting with the prototype (e.g., quiet/noisy, alone or accompanied by other users). For instance, if the user completes the questionnaire by recording his/her answers with an audio recording element (see figure 1, screen on the right bottom), the surrounding noises can or might also be recorded. Another positive consequence from using several modalities to collect data is the resulting accessibility it provides for users with impairments. For example, if testing a tool for blind users, they can be requested to easily collect usability data by recording thoughts instead of writing.

Furthermore, if needed, users are also able to film or take pictures (whenever the device includes a camera) of the environment in which the activity is taking place or where the usability issue was detected.This contextual information is extremely valuable, providing information on the user's location and the environment and can be used in a posterior analysis relating it with the data that is gathered through logging.

Screen/Card Annotation and Tagging. Text, audio or video (if the device has a microphone or camera) annotations can be added to every screen while using the prototypes. Annotations can store thoughts, opinions or users' impressions, pictures or the usage context, or any other information that the user whishes to collect (e.g., diary studies). The aim is to compensate for the absence of ESM questionnaires that focused a detected issue while the user interacted with the prototype and to allow users to freely provide evaluation data about the usage experience, even if not requested to do so. Screens can be tagged with sequence numbers in-situ and their sequence can be easily modified. This supports contextual card-sorting, adjusting the navigation sequence or desired workflow to the requirements of particular situations.

5   Analysis Tools

The framework includes two different approaches to the analysis of data which are directly correlated with the supported data-gathering techniques.

5.1 ESM/Diary Studies Results

All the data that is gathered by end-users, through ESM, diary studies, annotations and questionnaires can be directly reviewed and browsed on the runtime environment. As seen in figure 2 (middle and right), results of completed questionnaires are stored on the mobile device and can be loaded together with the questionnaires. Interaction

logs that are automatically stored can also be consulted directly on the mobile device. All the results are stored in XML files and can also be reviewed in any text editor.

Fig. 3. Log Player – Interaction heat maps on two different approaches for the same tool.

5.2 Log Player

In order to evaluate the users' behavior towards the user interface on real scenarios, we intended to replace, as far as possible, direct observation with a similar mechanism. The log player resembles a "movie player" which re-enacts every action that took place while the user was interacting with the prototype (figure 3). Adjusting the speed in which events are (re)played is also possible (e.g., fast-forward; double speed). Events can be played sequentially and according to the time-stamps that were recorded or they can be aggregated and searched by type (e.g., heat maps that show all the taps in a screen or browsing every "next screen" event). Although these logs are limited to direct interaction with the device, they still present enough detail to indicate whether a button needs to be enlarged, if the screen arrangement should be changed or what type of element or modality is preferable in certain situations [26].

The log-player tool also includes a communication module that allows the player to be connected to another mobile device while a user is interacting with a prototype. Here, the logging mechanism forwards every event to the monitoring device, allowing the designer to remotely review, in real-time, the user's interaction with the prototype. For instance, if the prototype is running on a Smart Phone with a GPRS connection or within the range of a Wi-Fi network, the designer is able to monitor and gather data on real-time on the user's interaction with the prototype directly on his/her desktop computer or even another mobile device.

6 Case Studies

The framework has been used to design and evaluate several prototypes on various domains (e.g., psychotherapy, physiotherapy, personal training, education, rich digital books). Throughout this process, both designers and end-users created and used their own prototypes on various locations, gathering data through the various techniques

that are supported by the framework. Experiments occurred during the initial stages of design.

In order to assess the benefits of using our framework, and to further improve it, end-users participated in the evaluation sessions in two different stages. In the first stage, traditional [2] and updated techniques [24], focusing on mobile concerns, were used. In the second set of sessions, the framework was used. End-users and designers were interviewed after each session. When using the framework, end-users were also requested to use the included annotation features (and some questionnaires that were included in each prototype) to provide their opinion of the overall experience and suggestions. Designers responded to questionnaires after analyzing the resulting data from the tests that were conducted.

Fig. 4. Users interacting with mobile prototypes on different devices and different settings.

A total of 5 designers and 50 end-users with diverse ages and backgrounds were involved in these experiences. From these 50 end-users, around 15 participated in the psychotherapy case study while the remaining were divided by the other experiences. Evaluation sessions took place in several settings (e.g., psychotherapy - therapist's office – fig. 4(d); personal training – soccer field – fig. 4 (a); education – university and gardens fig.4(e)). Different mobile devices were also used (e.g., TabletPCs, PDAs, SmartPhones with and without keyboards).

Throughout the tests that took place with evaluating the framework, the main concern was to understand the differences and advantages, for designers and end-users, between the use of traditional low-fidelity (figure 4(c)) and software prototypes with gathering techniques on the evaluation framework. When traditional techniques were used, paper cards/sketches were changed by a designer, using the WOz technique or by the user, when trying out the prototypes by him/herself. With these low-fidelity prototypes, ESM, end-users' and designers' questionnaires and registries were supported by paper and pen. Some experiments were filmed for posterior analysis and to detect issues that could be improved on the framework.

7   Results - Usability, Accessibility and Multi-modalities

Globally, designers' reactions to the evaluation framework were very positive. The ability to run their prototypes on actual devices (sometimes on the users' devices) was particularly appreciated. As stated by one of the designers:

"Mobile prototypes, especially phones, are very personal devices and users interact with their own differently than when using an unknown device, so seeing and getting usage information of users interacting with their devices was very positive."

ESM questionnaires that could be triggered and configured to request users for information at specific times were also very useful. However, the definition of the triggers and conditions took some time to be easily manipulated by the involved designers. Moreover, the construction of the questionnaires, especially when using audio and video recorders was somewhat difficult to understand and use. When asked about the active data gathering techniques, one of the designers said:

"It's a bit difficult to test sounds and videos on top of the sketches. Initially I also had trouble when creating questionnaires that didn't use text-boxes. The best feature was being able to add questionnaires to each screen and define when and if they should pop up."

Nevertheless, the analysis of the resulting data was stated to be very easy and natural since it followed the same mechanisms used to interact with the prototypes. In particular, the captured videos and photos (albeit very few) gave designers a very clear idea of the context in which users were while using the prototypes.

Experiences with the log-player also led to interesting findings. Usage patterns and the adequacy of some modalities were noticeable while reviewing the logs. For instance, track-bars, although not requiring text input, raised some difficulties mainly given the small size of the interactive counter. Moreover, when completing a task, if users were seated, they usually used the device's QWERTY keyboard (when available). However, once walking they preferred to use the virtual keyboard, using one hand to hold the device and the other to tap on the virtual keyboard, alternating with any other activity that required their hand. Curiously, once seated again, they would not return to the physical keyboard. Also, while walking, accuracy towards buttons was much lower. Audio input was also avoided when users were accompanied. However, audio output was preferred in most situations [26].

Figure 3 shows screenshots of the log player being used to evaluate a low-fidelity prototype for a movie player designed by a teacher during a brainstorming/evaluation session that took place at a public university. The images show the two prototypes being analyzed on the log player. Since all the logs have time-stamps and are cataloged by date, it was simple to correlate the logs and the locations/settings from which they resulted. Moreover, even specific portions of each evaluation session could be identified (e.g., at the beginning of the test, the user was seated; at the end of the evaluation test, the user was walking to another class). These situations were mapped to parts of the log where we noticed different accuracies regarding button selection and interaction, which allowed us to see that most of the missed taps on the screen referred to the situations where users were walking. As expected, while they were seated, accuracy was much higher. However, the log analysis provided a fairly precise idea of the necessary size and location for each button. With the first prototype, the log shows that users had difficulties while using the video controls. With the second version of the same prototype, with larger buttons, accuracy was much higher. Results showed that components placed near screen edges raised usage difficulties, especially when users interacted with their fingers instead of the stylus.

So far, all the involved designers considered the revision of users' behavior, without the need for direct observation, extremely useful. In fact, when combined

with the questionnaire results this allowed the detection of several issues which translated directly into UI improvements. Results were particularly interesting since they focused not only on a wide variety of contexts but also allowed the detection of problems that emerged while transiting between contexts. However, interestingly, the most noteworthy and positive results were found when interviewing users.

The use of actual devices and prototypes that could be interacted with and used was very appealing to users and provided a much more tangible and realistic experience, especially when compared to the traditional low-fidelity prototypes.

"I only really understood what they (the designers) were showing us when I was able to use the PDA and the buttons started to actually do something."

Moreover the ability to participate on the prototyping process, even during field tests, was one of the favorite features.

"It was really nice to be able to propose ideas to them and show what would be a better fit for me by drawing it on the mobile device and bringing it to life right there. It made me want to become a designer."

Finally, when designers were able to create questionnaires with different modalities (e.g., video and audio recorders), users provided more and richer data. As commented by most users, this was a reflection of the easiness to speak or photograph when using the prototype instead of stopping and writing thoughts or detected issues.

"Compared with the paper version, it was much easier to record what I had in mind or photograph what was going around than writing it down."

Moreover, as previously mentioned, it allowed users with disabilities (on the rich digital book player case) to test the prototypes since the beginning and to provide usability data as well, regardless of the used device.

Globally, the results of the experiments and on users' opinions validated the positive influence of the prototyping and evaluation framework on the design process. Some of the findings that emerged when using the evaluation and analysis features resulted in modifications that were specific to the domains of each case study while others confirmed generic guidelines that can apply to most mobile devices when used ubiquitously (e.g., element location, content per screen, preferred modalities).

7.1 Design Perspectives and Guidelines

From the design process perspective, we believe the use of the evaluation framework can support and enhance the design of mobile applications and user interfaces. The main reasons and discussion topics that resulted from the case studies and evaluation sessions indicate that:

• Evaluation in-situ provides shorter design cycles. By gathering data, analyzing and adjusting prototypes on the field, designers are able to redefine, adjust concepts or approaches and test them very quickly and directly on the used device.

• Users grasp ideas and concepts better. The use of actual devices and interactive prototypes, that do not depend on the designer to work or change cards, provide a better usage experience and end-users and allows them to understand what designers are trying to convey more quickly.

• Data gathering works better when using multimodalities. Video and voice recording allowed users to quickly register their thoughts, opinions or answer ESM questionnaires without requiring them to stop what they were doing.

• Since data gathering is easier, users provide richer data. When compared with initial experiences using traditional techniques, users collected much more data. The ability to film/photograph brings great benefits. Capturing a video or taking a photo of the surrounding environment is much easier than writing down a description of the user's context and provides much more detail.

• Intelligent ESM provides means to focus different or specific details. The ability to detect the user's activities and define questionnaires that can be displayed according to the user's behavior allows designers to focus key issues (e.g., user takes too long when viewing a screen or frequently misses a button).

• Automatic Wizard-of-Oz facilitates designer's work and reduces the need to follow or ask users to change the screens/cards by themselves, facilitating the process and, once again, providing better and more realistic usage experiences.

• Logging provides information on various details for low and high-fidelity prototypes. Furthermore, the ability to review how users interact with a prototype provides paramount information regarding usability issues.

• Overall, since it requires no programming knowledge the framework allowed users to actively participate on the evaluation, creating and adjusting their own prototypes, providing a softer and sounder transition between design fidelities.

8 Conclusions and Further Research Directions

This paper presented extensions to a framework which aim at taking advantage of some emerging techniques by offering support for in-situ mobile evaluation. The framework includes several methods and approaches into tools that support and integrate prototyping and evaluation, providing a tighter fit between these two stages and resulting in shorter design cycles on the initial design stages.

The experiences and case studies in which the tools were used confirmed the benefits that it provides for the design of mobile applications and user interfaces, especially during the initial stages of design. Besides facilitating the gathering of data on ubiquitous settings, it promoted user involvement and provided better and more realistic usage experiences which resulted in richer evaluation and field sessions.

Although we do not aim at replacing direct user observation or other contextual techniques, we believe that the presented framework can play an important part on the quick design and testing of design concepts and user interfaces. The encountered limitations, especially the development of ESM questionnaires and their integration within the prototypes, will be the focus of further evaluation and adjustments.

Finally, we are integrating the framework into a new group version which includes a large screen display module where several logs can be seen simultaneously. It enables teams to review logs simultaneously, comparing a user or a prototype's performance in various settings. The group log player also allows designers to monitor several users, even detecting interactions between each other. These will allow for the evaluation of team work and collaboration within mobile settings.

References

1. Barnard, et al. Capturing the effects of context on human performance in mobile computing systems. Personal and Ubiquitous Computing, Vol. 11, No.46.
2. Beyer, H., Holtzblatt, K., 1998. Contextual Design: Customer Centered Approach to Systems Design, Academic Press, San Francisco, CA, USA.
3. Carter,S.,et al.Momento:Support for Situated Ubicomp Experimentation. CHI07, ACM.
4. Consolvo, S. and M. Walker, Using the Experience Sampling Method to Evaluate Ubicomp Applications. IEEE Pervasive Computing, 2003. 2(2): pp. 24-31,IEEE.
5. Costabile, M. F., D. Fogli, et al. (2003). Building Environments for End-User Development and Tailoring. IEEE Symposia on Human Centric Computing Languages and Environmnets
6. Davis, R., et al. SketchWizard: Wizard of Oz Prototyping of Pen-Based User Interfaces. In Procs. UIST'07, ACM.
7. Duh, H. et al. Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests. Mobile HCI'06. ACM.
7. Froehlich, J., et al. MyExperience: A System for In Situ Tracing and Capturing of User Feedback on Mobile Phones. MobiSys'07, pp. 57-70, ACM.
9. Holmquist,L.,Prototyping:generating ideas or cargo cult designs?Interactions, 2005. ACM.
10. Jones, M. and G. Marsden, Mobile Interaction Design. 2006, John Wiley & Sons, England.
11. Kjeldskov, J. and J. Stage, New Techniques for Usability Evaluation of Mobile Systems. International Journal of Human Computer Studies, Elsevier, 2003.
12. Kjeldskov, J. and C. Graham, A Review of Mobile HCI Research Methods. Mobile HCI'03. Vol. 2795/2003. 2003: Springer Berlin / Heidelberg. 317-335.
13. Klemmer, S. R., et al, SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In Proc. of UIST'00, San Diego, California, USA, pp. 1-10, ACM.
14. Landay, J. SILK: Sketching Interfaces Like Krazy. CHI'96, pp.398-399, ACM.
15. Li, Yang, Hong, J., Landay, J. Topiary - Design Challenges and Principles for Wizard of Oz Testing of Location-Enhanced Applications. Pervasive Computing,70-75, 2007, IEEE Press.
16. Lin, J. et al. Denim: Finding a Tighter Fit Between Tools and Practice for Web Site Design. CHI'00, ACM.
17. Nielsen, C.M., et al. It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field. NordiCHI'06. ACM.
18. Oulasvirta, A., & Nyyssönen, T. (in press). Flexible hardware configurations for studying mobile usability. Journal of Usability Studies.
19. Sá, M. Carriço, L., Low-fi Prototyping for Mobile Devices. CHI'06, pp. 694-699, ACM.
20. Svanaes, D. and G. Seland. Putting the users center stage: role playing and low-fi prototyping enable end users to design mobile systems. In Procs. CHI'04, Austria ACM.
21. Scholtz, J. & Consolvo, S. (2004). Toward a framework for evaluating ubiquitous computing applications. IEEE Pervasive Computing, 2, 82–86.
22. Sohn, T., Li, K. A., Griswold, W. G., and Hollan, J. D. 2008. A diary study of mobile information needs. In Procs. CHI'8, Florence, Italy, 2008. ACM.
23. Thompson, K.E., Rozanski, E.P. & Haake, A.R. (2004). Here, there, anywhere: Remote usability testing that works. In Proceedings of SIGITE'04, ACM, Salt Lake City, USA.
24. Sá, M., Carriço, L. Lessons from Early Stages Design of Mobile Applications. MobileHCI'08, pp. 127-136. ACM.
25. Sá, M,et al.A Mixed-Fidelity Prototyping Tool for Mobile Devices.AVI'08,.225-232,ACM.
26. Sá, M, et al. A Framework for Mobile Evaluation. CHI'08,(extended abstracts), 2673-2678, ACM.