Field evaluation of collaborative mobile applications

[*]

ABSTRACT

This chapter presents a usability evaluation method for context aware mobile applications deployed in semi-public spaces that involve collaboration among groups of users. After reviewing the prominent techniques for collecting data and evaluating mobile applications, we propose a methodology that includes a set of combined techniques for data collection and analysis, suitable for this kind of applications. To demonstrate its applicability, a case study is described where this methodology has been used. It is argued that the method presented here can be of great help both for researchers that study issues of mobile interaction as well as for practitioners and developers of mobile technology and applications.

INTRODUCTION

Mobile devices are part of many peoples' everyday life, enhancing communication, collaboration and information access potential. Their vital characteristics of mobility and anywhere connectivity can create new forms of interaction in particular contexts, new applications that cover new needs that emerge and change the affordances of existing tools/applications.

A case of use of such devices, with particular interest, concerns *public places rich in information* for their visitors, in which mobile technology can provide new services. Examples of such places, are *museums* and other sites of culture (Raptis, Tselios, Avouris, 2005), *public libraries* (Aittola, Parhi, Vieruaho, Ojala, 2004) (Aittola, Ryhänen, Ojala, 2003), *exhibition halls* and *trade fairs* (Fouskas, Pateli, Spinellis, Virola, 2002). In these places mobile devices can be used for information collection and exchange, for ad hoc communication with fellow visitors and for supporting face-to-face interaction.

Usability evaluation of mobile applications is of high importance in order to discover early enough the main problems that users may encounter while they are immersed in these environments. Traditional usability evaluation methods used for desktop software cannot be directly applied in these cases since many new aspects need to be taken in consideration, related to mobility and group interaction. Therefore, there is a need either to adapt the existing methods in order to achieve effective usability evaluation of mobile applications or to create new ones. An important issue, that is discussed here, is the *process* and *media* used for recording user behaviour.

[*] This is an early draft of a chapter to appear in a multiauthored volume on User Interface Design and Evaluation for Mobile Technology. The chapter can be quoted as follows:
A. Stoica, G. Fiotakis, D. Raptis, I. Papadimitriou, V. Komis, N. Avouris (2007), Field evaluation of collaborative mobile applications, in J. Lumsden (ed.), "Handbook of Research on User Interface Design and Evaluation for Mobile Technology", Idea Group Publishers.

Data collection during usability studies is a particularly important issue as many different sources of data may be used. Among them, *video and audio* recordings are invaluable sources for capturing the context of the activity including the users' communication and interaction. It has been reported that in cases of studies that audio and video recordings were lacking, it was not possible to explain why certain behaviour was observed (Jambon, 2006). Recording user behaviour is a delicate process. Video and audio recording must be as unobtrusive as possible in order not to influence the behaviour of the subjects while, on the other hand, the consent of the users for their recording should be always obtained. In addition, questions related to the frame of the recorded scene, viewing angle and movement of the camera are significant. We must stress that there is a trade off between capturing the interaction with a specific device and capturing the overall scene of the activity. For example often crucial details may be missing from a video if recording the scene from a distance. Therefore this video has to be complemented by other sources of related information, like screen captures of the devices used.

In order to conduct a successful usability evaluation, apart from collecting activity data, techniques and tools are needed for analysis of the collected information. In the last years new usability evaluation techniques have emerged, suitable for mobile applications. Many of these methods focus mainly on user interaction with the mobile device, missing interaction between users and user interaction with the surrounding environment.

Taking into consideration these aspects, the aim of this chapter is to discuss techniques and tools used first for collecting data during usability evaluation studies of mobile devices and then for the analysis of these data. In the process, we present a combination of a screen capturing technique and some tools that can be used for analysis of data of usability studies.

BACKGROUND

The usability of a product has been traditionally related with the ease of use and learn to use, as well as with supporting users during their interaction with the product (Dix et al, 2004, Schneiderman 2003). There have been many attempts to decompose further the term and render it operational through attributes and apt metrics. According to ISO 9241-11 standard, usability is defined as the "*extend to which a product can be used with effectiveness, efficiency and satisfaction in a specified context of use*" (ISO 9241). According to this view, a product's usability is directly related to the *user*, the *task* and the *environment*. Consequently, usability cannot be studied without taking in consideration the goals and the characteristics of typical users, the tasks that can be accomplished by using the product and the context in which it is going to be used. Making a step further on defining usability, the same standard suggests three potential ways in which the usability of a software product can be measured:

(a) By analysis of the *features of the product*, required for a particular context of use. Since ISO 9241 gives only partial guidance on the analysis process, in a specific problem there can be many potential design solutions, some more usable than others.

(b) By analysis of the *process of interaction*. Usability can be measured by modelling the interaction with a product for typical tasks. However, current analytic approaches do not produce accurate estimates of usability since interaction is a dynamic process which is directly related to human behaviour that cannot be accurately predicted.

(c) By analyzing the *effectiveness and efficiency*, which results from use of the product in a particular context, i.e. measuring performance as well as the satisfaction of the users regarding the product.

Having in mind the above three perspectives, there is a need for combining methods that capture the specific situation of use in a specific domain. Usability evaluation methods can be grouped in four categories (Nielsen, 1993): *Inspection, user testing, exploratory and analytic methods*. Many techniques have been devised along these lines and have been extensively used in usability evaluation of desktop applications. Therefore the first approach in evaluating mobile applications was to apply these existing techniques. Such an approach can be found in Zhang and Adipat (2005) survey of usability attributes in mobile applications which identified nine attributes that are most often evaluated: learnability, efficiency, memorability, user errors, user satisfaction, effectiveness, simplicity, comprehensibility and learning performance. Such an approach is however limited, given the special characteristics of mobile devices with respect to desktop environments (Kjeldskov, Graham 2003).

The mobile applications introduce new aspects to evaluate. We cannot limit the evaluation only to the device (typical scenario in desktop applications) but we must extend it including aspects of context. The context in which the application is used is highly relevant to usability issues and often bears dynamic and complex characteristics. There is the possibility that a single device is used in more than a single context, in different situations, serving different goals and tasks of a single or a group of users. Also, group interaction, a common characteristic in mobile settings, gives a more dynamic character to the interaction flow of a system and increases the complexity of the required analysis as well as the necessity of observational data.

Along these lines, a new breed of methods for usability evaluation has been proposed (Hagen, Robertson, Sadler, 2005), (Kjeldskov, Graham 2003), (Kjeldskov, Stage, 2004). The process of selecting appropriate usability attributes to evaluate a mobile application depends on the nature of the mobile application and the objective of the study. A variety of specific measures (e.g., task execution time, speed, number of button clicks, group interactions, seeking support, etc.) have been proposed to be used for evaluation of different usability attributes of specific mobile applications. In the next section we discuss problems of data collection during mobile usability studies.

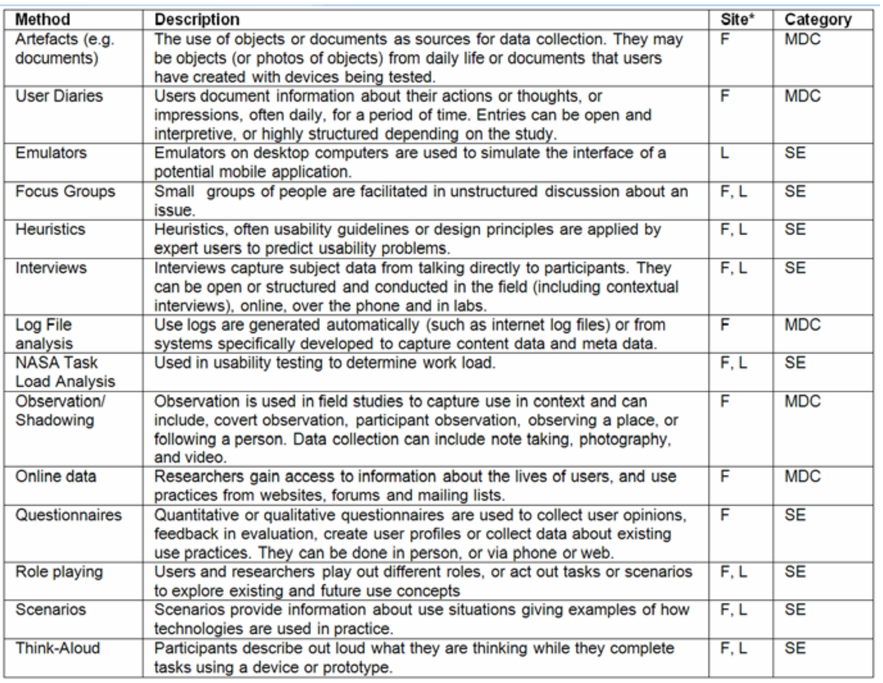
Data Collection Techniques

A significant step during a usability evaluation study is to collect appropriate observational data to be analyzed. Hagen, Robertson, Kan & Sadler (2005) classify the data collection techniques for mobile human-computer interaction in three categories:

(a) *Mediated data collection (MDC)*, access to data through participant and technology, *do it* – the user makes himself the data collection; *use it* – data is collected automatically through logs; *wear it* – user wears recording devices that collect the data. (b) *Simulations and enactments (SE)* where some form of pretending of actual use is involved and (c) *combinations* of the above techniques. A review of different techniques of data collection, according to (Hagen et al, 2005) is shown in Table 1.

The data that are collected by these techniques come either directly from the user (through interviews, questionnaires, focus groups, diaries etc), by the evaluator (i.e. notes gathered during the experiment, observation of videos etc.) or by raw data (log files etc). All types of data need to be analyzed in order to become meaningful. Such data, in most cases, are in the following forms:

- *log files* which contain click streams of user actions. These data can be derived by the application itself or by an external tool that hooks into the operating system message handler list. The latter case for mobile devices requires many system resources and therefore is not technologically feasible today even in the most powerful mobile devices, like PDAs.
- *audio/video recordings* of the users made through various means, like wearable mini cameras and/or audio recorders, static video cameras, operator or remote controlled cameras, from close or far distance.



Method	Description	Site*	Category
Artefacts (e.g. documents)	The use of objects or documents as sources for data collection. They may be objects (or photos of objects) from daily life or documents that users have created with devices being tested.	F	MDC
User Diaries	Users document information about their actions or thoughts, or impressions, often daily, for a period of time. Entries can be open and interpretive, or highly structured depending on the study.	F	MDC
Emulators	Emulators on desktop computers are used to simulate the interface of a potential mobile application.	L	SE
Focus Groups	Small groups of people are facilitated in unstructured discussion about an issue.	F, L	SE
Heuristics	Heuristics, often usability guidelines or design principles are applied by expert users to predict usability problems.	F, L	SE
Interviews	Interviews capture subject data from talking directly to participants. They can be open or structured and conducted in the field (including contextual interviews), online, over the phone and in labs.	F, L	SE
Log File analysis	Use logs are generated automatically (such as internet log files) or from systems specifically developed to capture content data and meta data.	F	MDC
NASA Task Load Analysis	Used in usability testing to determine work load.	F, L	SE
Observation/ Shadowing	Observation is used in field studies to capture use in context and can include, covert observation, participant observation, observing a place, or following a person. Data collection can include note taking, photography, and video.	F	MDC
Online data	Researchers gain access to information about the lives of users, and use practices from websites, forums and mailing lists.	F	MDC
Questionnaires	Quantitative or qualitative questionnaires are used to collect user opinions, feedback in evaluation, create user profiles or collect data about existing use practices. They can be done in person, or via phone or web.	F	SE
Role playing	Users and researchers play out different roles, or act out tasks or scenarios to explore existing and future use concepts	F, L	SE
Scenarios	Scenarios provide information about use situations giving examples of how technologies are used in practice.	F, L	SE
Think-Aloud	Participants describe out loud what they are thinking while they complete tasks using a device or prototype.	F, L	SE

Table 1. Existing techniques for data collection used in studies of mobile technology. (Hagen et al, 2005) **F**=**F**ield, **L**=**L**aboratory, **MDC**=Mediated Data Collection, **SE**=Simulation and Enactments

- *screen recordings* by video cameras or by direct screen capturing through software (running on the device) the interaction flow in form of screen snapshots. This is a sequence of image representations of the user interface at certain instances, that usually are taken at varying frequencies, usually a few snapshots per second. The screen snapshots can be stored either locally on the device (since it is feasible to store a large amount of data in memory cards) or on a central server over a wireless network connection

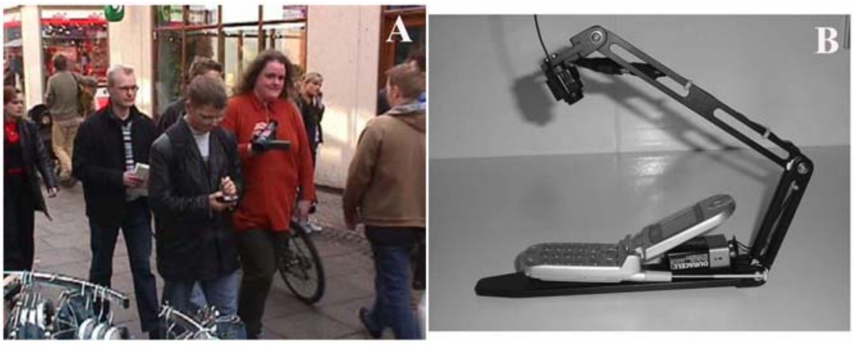


Figure 1. A) Shadowing technique adapted from Kjeldskov, Stage (2004)
B) Recording screen with wireless camera, adapted from Betiol, de Abreu Cybis (2005)

Screen recordings of mobile devices are invaluable resources that can greatly help evaluators identify usability problems. Various techniques can be used for capturing the screen of a mobile device: One is the recording of the screen by using a mini wireless camera (Figure 1B). It can be very helpful in cases of individual users but it is not suitable in the case of an application that involves beaming actions (e.g. Bluetooth, infrared) and/or interaction with the physical space, because it can influence negatively the use of the device and can create obstacles in the infrared beams, sensors or readers attached to the device (i.e. to an RFID reader). The main advantage of this technique is that the camera records, besides the screen, the movements of the users fingers or stylus, capturing valuable data identifying potential interaction problems (for example the user hesitates to click something because the interface or the dialogs are confusing).

An alternative technique is the shadowing technique which can effectively work for individual users (Figure 1A). Again this technique is not suitable for group activities, where often the subjects form groups and move continuously. Even in cases that it is considered possible to record properly, there could be many events missing because of the frequent movements of the subjects or the shielding of the screen by their body and hands.

The direct observation technique has also certain limitations (Cabrera et al. 2005, Stoica et al. 2005) because the observer must distribute his attention to many subjects. In case there are observers available for each user they will restrict the mobility of the users and they will distract their attention when being in so close range. Consequently, all these techniques impose the presence of the observer to the users, thus affecting their behavior.

Another significant issue that directly affects the usability evaluation is related to the location in which the study is conducted. There many arguments in favour of *field usability studies* (Nardi, 1996, Kjeldskov et al, 2004, Zhang and Adipat, 2005, Kaikkonen et al 2005). Comparative studies between laboratory and field evaluation studies have drawn however contradictory conclusions. In a recent survey of evaluation studies of mobile technology (Kjeldskov et al. 2003), 71% of the studies were performed in the laboratory, which revealed a tendency towards building systems based on trial and error and evaluating systems in controlled environments at the expense of studying real use of them. So the question of what is useful and what is perceived problematic from a user perspective often is not adequately addressed.

In summary, in order to conduct a usability evaluation of a mobile application/system, there is a need to take into consideration the attributes that are going to be measured, the data collected for these measurements, the location in which the evaluation will take place and finally the appropriate tools to analyze them, having always in mind the user and the context of interaction.

Data Analysis

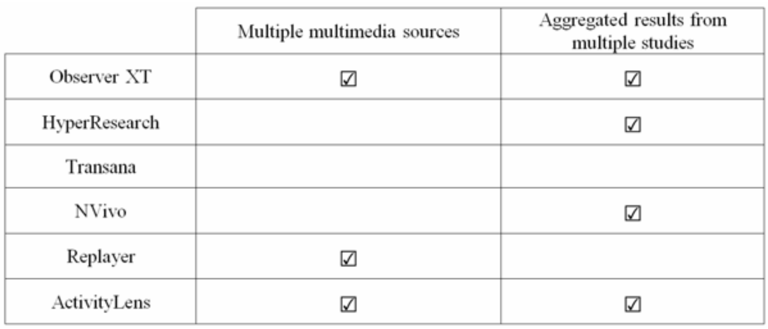
Usability evaluation of mobile applications is more complex than desktop software evaluation, since new characteristics such as group activity and the interaction with the surrounding environment need to be taken in consideration. In order to acquire an understanding of group activity and performance, huge amount of structured and unstructured data of the forms discussed in the previous section need to be collected. These data should capture the activity of subjects, including their movements, facial expressions, gestures, dialogues, interaction with the devices and objects in the environment. Analysis of these data require special attention on details as well as the context of use, thus it can be a tedious process, which can be facilitated by a suitable analysis tool (Benford et al. 2005).

Various tools have been developed to support usability evaluation studies and in general to record and annotate human activity. These tools often handle video and audio recordings and synchronize them with text files, containing hand-taken notes. This combination creates a dataset that is rich in information, which is then annotated through an adequate annotation scheme, which creates quantitative and qualitative measures of the observed user-device interaction. Typical examples of such tools are: the *Observer XT* (Noldus, 2006), *HyperResearch* (Hesse-Biber, Dupuis, Kinder, 1991) (ResearchWare, 2006), *Transana* (Transana, 2006), *NVivo* (QSR, 2006) (Rich, Patashnick, 2002) (Welsh, 2002) and *Replayer* (Tennent, Chalmers, 2005). From them only Replayer and Observer XT, have special provisions for mobile settings. The extra characteristics in evaluation of mobile applications (group activity and interaction with the surrounding space) demand the extended use of multimedia files that thoroughly capture the activity. Thus, there is a need for a tool that combines and interelates all the observational data in a compact dataset and gives to the usability expert the ability to easily navigate them from multiple points of view (access in user – device interaction, access in user – space interaction).

All of the above tools utilize video sources at a different extend, with the exception of *Nvivo* that focuses more in textual sources. Nvivo allows linking of evaluator's notes with video extracts, without permitting more fine grained handling of video content. On the other hand, *HyperResearch* and *Transana* do support flexible handling of video sources but they do not allow the integration and synchronous presentation of multiple video sources in the same study. Thus, Nvivo, HyperResearch and Transana cannot successfully respond to the extra characteristics of mobile applications. On the other hand, *Replayer* is a distributed, cross platform toolkit that allows the integration of multiple video sources and presents analysis data in various forms, such as histograms and time series graphs. Although Replayer efficiently supports usability analysis of mobile applications, its failure to handle and to compare data that come from various studies makes it not suitable for cases of multiple studies in which there is need to aggregate and generalise the findings. On the contrary, *Observer XT* is a powerful commercial tool, widely used in observation studies, that enables the synchronous presentation of multiple video files and also the derivation of overall results about the activity of multiple subjects. Although Observer XT meets the requirements of new characteristics of mobile applications, its use requires a prior lengthy training period.

A tool that has been especially adapted for analysis of data from mobile applications' evaluation studies is the *ActivityLens* which attempts to tackle some of the limitations of existing tools. Its main advantage is its ability to integrate multiple heterogeneous qualitative but also quantitative data. It allows the usability expert to directly access the collected data, thus to simultaneously focus on users' movements on the surrounding environment and user-device interaction. To sum up, ActivityLens supports analysis of collected data and produces results that cover the overall activity concerning all the participants.

Weitzman and Miles (cited in *Berkowitz,* 1997) suggest that a criterion for the selection of an adequate analysis tool is related to the amount, types, and sources of data to be analyzed and the types of analyses that will be performed. In Table 2 a description is provided about how the above tools support the extra characteristics of usability evaluation studies of mobile applications.



	Multiple multimedia sources	Aggregated results from multiple studies
Observer XT	☑	☑
HyperResearch		☑
Transana		
NVivo		☑
Replayer	☑	
ActivityLens	☑	☑

Table 2. Characteristics of usability evaluation tools

Data analysis of Mobile user studies with ActivityLens

ActivityLens is a tool that embodies features especially designed for usability evaluation of mobile applications. ActivityLens is an evolution of the earlier Collaboration Analysis Tool (ColAT) (Avouris, Komis, Fiotakis, Margaritis, 2004), (Avouris, Komis, Fiotakis, Margaritis and Voyiatzaki, 2005), originally designed for video analysis of collaborative learning activities. It was found particularly suitable for the proposed approach which involves multiple perspectives of the activity, based on different multimedia data.

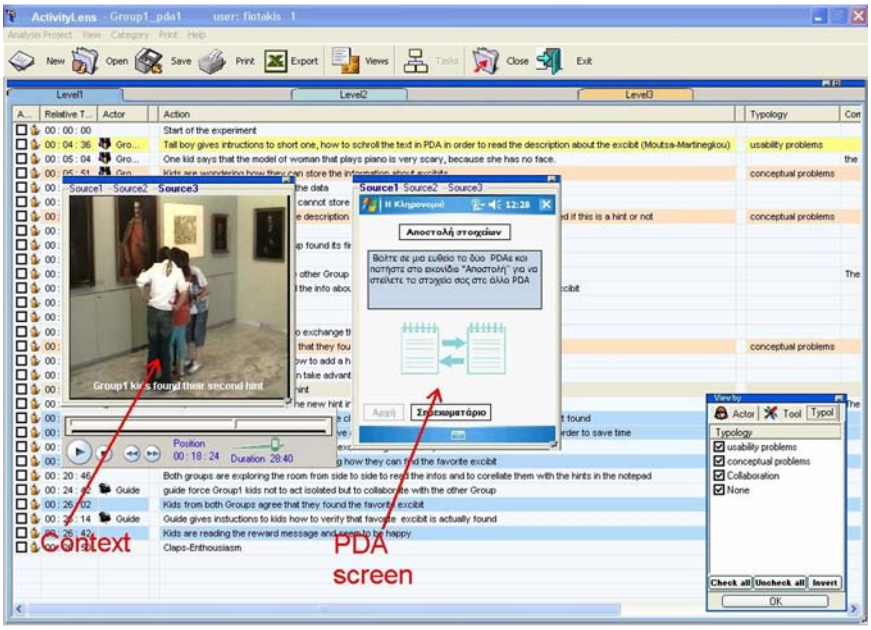
In *ActivityLens* all the collected data are organized into *Studies*. An example of a Study is the usability evaluation that was conducted in a Historical Cultural museum, described in the next sections. The tool allows us to define *Projects* that belong to a specific *Study*. A Project is defined by the evaluator and can have different perspectives depending on the situation. For example, a Project can be defined as the set of data gathered from various groups over a set period of time, or it can be defined as a set of data of a specific group of users.

These data can be video and audio files, logfiles, images and text files including hand-taken notes of the observers. *ActivityLens* supports almost all the common video and audio file formats, including file types that are produced by mobile devices such as .mp4 and .3gp. The observed activity is reported in an XML logfile. This file describes the activity as a set of events, reported in sequential order, following this typical structure:

<event id>, <time-stamp>, <actor>, <tool>, <event-description>, <type of event>, <comments of evaluator>

The logfile events are presented via a simple spreadsheet view in order to be easily accessible for inspection and annotation. In addition, *ActivityLens* permits integration and synchronization of the collected multimedia files.

All the data can be reproduced and annotated on-the-fly in order to highlight interesting events. An example is shown in Figure 2, in which an overview video and a PDA screen are synchronized and annotated. The annotation of the observed events is based on a classification scheme defined by the evaluator. For example, an evaluator is analyzing videos that describe the activity of a group of students that try to solve a problem. During the activity some students propose ways to solve the problem and argue about it. Thus, one representative type of event could be defined as "Proposal". For usability studies an evaluator can define typologies based on usability attributes, concerning for instance user errors, comments expressing subjective view and events marking successful completion of tasks.



Relative T...	Actor	Action	Typology
00 : 00 : 00		Start of the experiment	
00 : 04 : 36	Gro...	Tall boy gives intructions to short one, how to schroll the text in PDA in order to read the description about the excibit (Moutsa-Martinegkou)	usability problems
00 : 05 : 04	Gro...	One kid says that the model of woman that plays piano is very scary, because she has no face.	
00 : 20 : 46		Both groups are exploring the room from side to side to read the infos and to corellate them with the hints in the notepad	
00 : 24 : 42	Guide	guide force Group1 kids not to act isolated but to collaborate with the other Group	
00 : 26 : 02		Kids from both Groups agree that they found the favorite excibit	
00 : 26 : 14	Guide	Guide gives instuctions to kids how to verify that favorite excibit is actually found	
00 : 26 : 42		Kids are reading the reward message and seem to be happy	
		Claps-Enthousiasm	

Βάλτε σε μια ευθεία τα δύο PDAs και πατήστε στο εικονίδιο "Αποστολή" για να στείλετε τα στοιχεία σας στο άλλο PDA

Figure 2. The usability evaluation tool –ActivityLens

ActivityLens provides the evaluator with the ability to reduce the huge amount of collected data through an event filtering mechanism. This feature is of high importance because it helps the evaluator to focus on interesting sequences of events and makes them emerge from the "noise". The evaluator is allowed to define criteria for specific Actors, tools used and types of events or any combination between them. For example, the evaluator can choose to view all occurrences of "Proposals" made by "George" OR "John". The criteria selection tool is shown in figure 3.

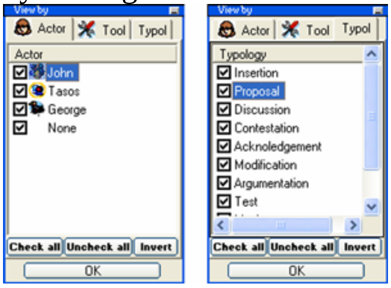


Figure 3. Event Filtering tool in ActivityLens

Proposed methodology

Based on the outlined above data and analysis requirements, in this section we propose a methodology suitable for usability evaluation of mobile applications. This method is proposed for applications deployed in places like museums, libraries etc., in which groups of users interact among themselves and with the environment, in various ways. These semi-public spaces represent 'living organisms' that project, in a visible and tangible form the various facets of information. For example, in a Museum such applications assist the visitors in discovering and acquiring knowledge. We can characterize a museum as an ecology, (Gay and Hebrooke, 2004) that is constituted by two main entities, the exhibits and the visitors, populating the same space. Items of the collection are exhibited to visitors, who react by discovering them in a way that is, at a large extent, influenced by the surrounding space. Also, visitors usually interact with each other, for example because they comment the exhibits independently from the use of technology. This methodology involves initially the preparing study phase, the recording activity phase and then the analysis of the activity.

Preparing the study

Usually, activities that are expected to take place in semi-public places are desirable to be conducted in the field. For example, visitors inside a museum enjoy an experience that cannot be fully reproduced inside a laboratory. Therefore the evaluator needs to conduct a study in a representative place which should be adapted accordingly without disturbing its normal operation. Issues to be tackled are related with technological restrictions (e.g. wireless network infrastructure), recruitment of an adequate number of typical users, the extend of the study, etc. Consequently, it is evident that the preparation phase of the evaluation is a very important one as it builds the foundation for a subsequently successful study.

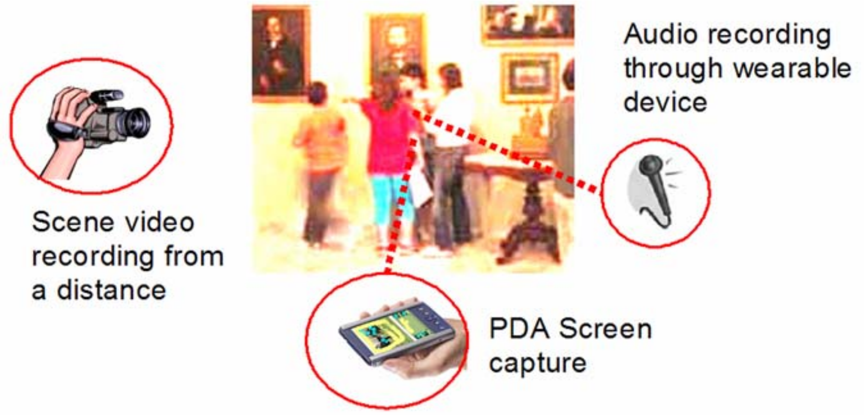

Audio recording through wearable device
Scene video recording from a distance
PDA Screen capture

Figure 4. Sources of observational data

Recording activity phase

A prerequisite in such environments is the low level of activity interference by the observers in order to minimize the behavioral change caused to the participants by the uncomfortable feeling of being observed and thus “disorienting the balance” in the ecology. The proposed recording activity includes an innovative combination of existing data gathering techniques in order to achieve the considered goal. The sources of data (figure 4) include: (a) screen recordings of the mobile devices, (b) audio recordings using wearable recorders, (c) video recordings from the distance, where the camera is operated by an operator or preferably by remote control, and as complementary source (d) interviews and questionnaires to the users. A brief discussion of the process of collecting these data is included next.

(a) Screen grabbing on the mobile device

In order to tackle problems related to the application nature (collaboration, interaction with the environment) we propose the mobile device to be also used as a screen recording device. The collected information can be in the form of screen-shots or aggregated in a low frame-rate video. The main requirement for a mobile device to become a screen recording device is that it must run a multitasking operating system in order to allow a background process to run in parallel with the main application. At the current technological status this is the case for most mobile devices (PDAs and smart phones), as the main operating systems are multitasking: Symbian OS, Windows Mobile, Palm OS (version 6.0 onwards), Java OS etc. Also, the needs of the market drove the mobile devices to handle large amounts of data that have to be consulted, edited, updated by the user while speaking, browsing, watching TV etc. As a result, mobile devices evolved from single process, sequential to multitasking and obtained increased storage capacities which permit the users to store on them a lot of information. Therefore, a mobile device can capture, by a parallel process the screen and either save the pictures on their memory or send them directly to a server via a wireless connection.

We have developed a prototype application that is suitable for the Pocket PC/Windows Mobile environment and runs in parallel with the application which has to be evaluated. It captures screen snapshots and stores them on the device at a predefined time interval. In our tests a compressed quarter VGA (240x320) screen shot was at most 32 KB that at a rate of 4 per second lead to a needed storage of about 450 MB / hour. We must stress that far better compression rates can be achieved by using video encoders.

The decision to grab the screen with a steady frequency and not per number of events, that would make sense in order to stop recording when the device is not used, was imposed by the technical current limitation: the scarce support for global system hooks on the Windows Mobile operating system. The lack of support is due to the fact that such hooks can affect critically the performance of the device.

(b) Audio recording with wearable devices

Audio can capture dialogs between users that express difficulty in interacting with the application and the environment, or disagreement. Audio recordings can often reveal problems that users do not report during interviews or questionnaires.

The audio recordings from the inbuilt microphone of the video camera are sometimes not very useful due to the noise and to the fact that usually the dialogues are in a low voice. Also, the distance between the subject and the camera does not allow recording of good quality sound. The ideal solution would be that the mobile device itself could record both the screen and the audio. Unfortunately, this is not feasible because of several reasons:

- The performance of the device degrades significantly by having two background processes running simultaneously, the one related to screen grabbing, discussed in (a), and the one to audio recording.
- The sounds that are produced by the device itself in most cases cover any other sound in the surrounding environment (i.e. a narration played back covers the dialogue.
- The storage might be a problem. Depending on the audio quality and compression used, 1 hour of recorded sound can take from 50 MB to 700 MB.

For these reasons, it seems that the most suitable solution is to use a wearable audio recorder that can store several hours of sound. These devices are very light; they weight less than 50 grams including the battery. The user can wear it with the help of a neck strap or put it in a pocket and adjust a clip microphone. The wearable audio recorders guarantee that we will not loose rich information concerning the dialogs between the subjects, collaborating and interacting with the application and the environment.

(c) Discrete / unobtrusive video recording

To complement the dialogs and the screen recordings, it is necessary to capture in video the ensemble. From this video, recording the context of the events, the social interactions between the group members (peers) and/or between groups can be depicted. In order to decrease as much as possible the level of obtrusion, the camera must be preferably maneuvered through remote control (allowing zoom and angle changes) or at least by a cameraman that will keep a large enough distance from the activity in order not to disturb the users. Often many video recordings may need to be made, from various angles, distance, or focusing in different aspects. These may be mixed in a single video stream if adequate equipment is used, or, more often, may be kept as separate sources of information. By studying these video recordings the evaluator can obtain a clear idea about the place in which the activity took place.

(d) Interviews and questionnaires

Considering that the above sources constitute the objective information, we also need to obtain from the users their subjective view through interviews and questionnaires. Through these sources, which vary depending on the situation, someone can formulate results regarding *user's satisfaction*, *learning performance* etc. attributes sometimes difficult to obtain simply through observation.

Analyzing activity phase

The purpose of the analysis is to identify instances of use of the devices and the infrastructure, which identify usability problems of the technology used. Analysis of recorded activity of groups in semi-public spaces is not a simple process. Researchers have not only to focus just on the devices but to take into account more complicated issues concerning the interaction between groups, the interaction between peers in a certain group and the interaction with the surrounding space. This analysis has to be meticulously performed in order to cover the above issues. During analysis all the collected sources that describe the group activity have to be combined and iteratively inspected. Initially a quick inspection of recorded activity helps usability experts to isolate the segments that need thorough analysis. Then, detailed inspection of these segments is required to interpret the observed interaction and depict the usability problems. This process can help usability experts to detect certain critical points of interaction that can be further examined in order to measure their frequency and dispersion between groups and to be clear how they affect the use of mobile applications. The proposed methodology concerning the recording and analysis process can be seen in figure 5.

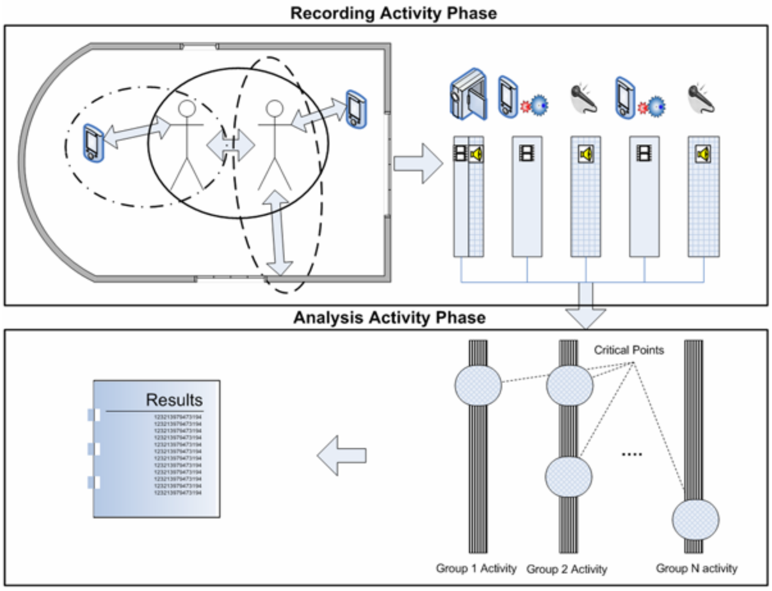


Figure 5. Recording and analysis phases of proposed methodology: interesting incidents are observed in the media files and are cross-checked for better understanding. These incidents are analyzed in terms of device and activity usability issues.

EVALUATING USABILITY OF A COLLABORATIVE CONTEXT AWARE EDUCATIONAL GAME

An example of a study in which we applied the proposed technique was a usability evaluation of a collaborative learning application supported by PDAs in a cultural-historical museum (Tselios et al. 2006). The study involved 17 students of the 5th grade of an elementary school (11 years old) who were invited to visit the museum and use the prototype of an educational application that was temporarily installed there. All the students were familiar with the use of mobile phones but they had not any former experience with PDAs. Furthermore, most of them described themselves in a pre-study questionnaire, as users of desktop computer systems on a daily basis.

The study took place in two of the museums' halls in which portraits and personal objects of important people of the local community were exhibited. First a short introduction to the activity was provided by a member of the research team who undertook the role of the guide. The educational activity was designed in a way that students were motivated to read information about these important people and collaboratively search in order to locate a specific exhibit according to the activity scenario. The children were divided in two groups and each group consisted of two teams of 4 or 5 children each. Each group participated in a different session for approximately one hour.

In order to achieve the scenario's goal each team was provided with a PDA equipped with a RFID tag reader. They used this equipment to locate hints, hidden inside textual description of the exhibits. These were obtained by scanning the exhibit RFID tags. The students could store the hints it in a notepad of the PDA. After collecting all or most of the hints the teams were encouraged to share their hints through beaming to each other.

Then the students using the found information, had to locate a specific-favorite exhibit which matched the description provided by the hints. When two teams agreed that they have found the favorite exhibit, they checked the correctness of their choice by scanning with both PDA's the RFID tag. A correct choice was indicated by the system with a verification message while a wrong one suggested a new search. When the study was over, each student was requested to answer a set of questions related to the group activity in the museum.

Preparation of the evaluation study

During the preparation of the study we contacted the museum and we obtained the permission to run the evaluation study. We examined the space of the museum well in advance (e.g. for determining wireless network setup options) and afterwards we run a small scale pilot in a simulated environment in order to check the suitability of the technological infrastructure. In order to ensure the participation of subjects we contacted a school in the vicinity of the museum and requested participation of a school party in the study.

Collecting Data

In order not to miss important contextual information, three video cameras were used in this study. Two of them were steadily placed in positions overlooking the halls, while the third one was handled by an operator who tenderly followed the students from a convenient distance. One student per team wore a small audio recorder in order to capture the dialogues between them, while interacting with the application and the environment. Furthermore, snapshots of the PDA screens were captured during the collaborative activity and stored in PDA's memory. After the completion of the study the guide, who was member of the evaluator team, had an interview with the students, asking them to provide their opinion and experiences from the activity in the museum, while back at school, a week later their teacher asked them to right an essay describing their experience.

Analyzing Data with ActivityLens

In order to analyse all the collected data according to the proposed methodology, we used ActivityLens that has already been effectively used in similar studies (Cabrera et al. 2005, Stoica et al. 2005).

The main reasons that we used ActivityLens among the discussed tools was its capacity of organizing observations into Studies (collection of projects) and its ability to present multiple perspectives of the whole activity (by integrating multiple media sources). Although Observer XT provides even more capabilities than ActivityLens, the choice of ActivityLens seemed to fit better the specific use case since its use didn’t require long training time. In addition, ActivityLens permits easy access to the activities of the subjects recorded in different data sources.

Three usability experts, with different level of experience, analysed the collected data, in order to increase the reliability of the findings. Initially, we created a new *ActivityLens Study* including 4 projects (each project concerns the observations of a team). We extensively studied the integrated multimedia files and annotated the most interesting situations. We must clarify that we didn't want to study the behaviour of each individual team member but we wished to evaluate the performance of the whole team. The performed analysis through ActivityLens revealed several problems related to the childrens’ interaction with the device and the overall setting, given the surrounding physical space and groups.

Several problems were identified when the students interacted with the handheld devices.

The analysis indicated that almost all the groups could not successfully scan the Exhibits tags in their initial attempts and get information about the exhibits. The RFID tags were located underneath each exhibits label. Since the users had no clear indication of where to place the tag scanner, some of them experienced difficulties interacting with them. Also, there was an unexpected delay in the scanning process between tag and PDA (the PDA needed about two seconds to scan the tag). While from the scene video recording it seemed that the user was scanning repeatedly the same label, combining this with the PDA screen recording gave us the real reason of this behavior – repeated unsuccessful tries to scan the tag. The users learned after a few frustrating attempts that they should target the center of the tags and hold the device for a couple of seconds.

A problem that troubled a specific group was the use of scrollbar in the textual description of the exhibits. The users were not familiar with the procedure of scrolling on a PDA and they repeatedly discussed among them about it. This problem was identified through the combined use of the audio and screen recording and was not visible from the scene video.

An unexpected problem was related to the content of some exhibits descriptions. They contained the word “hints” which confused the children and they were not sure if this was or not a hint that they could add to the notepad. This was spotted from the complementary use of the overview video with the dialogue audio recordings. The problem was overcome by asking the help of the guide.

With the use of ActivityLens we managed to detect many problems that were related to the interaction with the physical space. The most important one was that some of the exhibits tags were placed on the walls in such positions that they were not accessible by short students. In figure 6 an instance of this problem is shown.

Another interesting element that was made clear through the students’ dialogues and the videos was that in a certain area of the room an exhibit inspired fear to some of the children (e.g. a faceless piano player). Particularly one student was clearly afraid to get near the puppet and said to the other members: “I am not going near her. She is very scary!!! Look at her, she has no face!”. This situation made the team to avoid that area which contained exhibits with useful information for the activity.

The children that participated in the study often expressed their concern about being delayed in their play due to the presence of other museum visitors (at a certain point an independent school party crowded the hall). Through the audio it was obvious that the kids expressed their frustration because they were delayed in playing the game and the visitors because they were disturbed by the kids. These problems escape from the traditional usability analysis that focuses only on the device, because they contain the interaction between the user and the surrounding physical space.

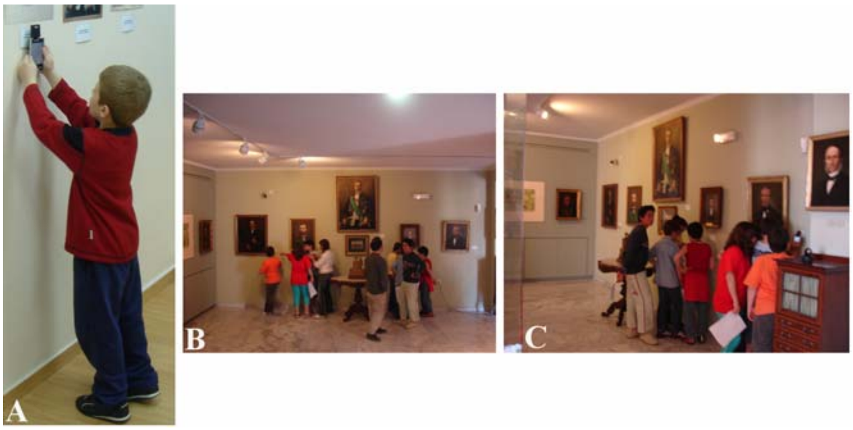


Figure 6. A) Instance of user- RFID tag interaction problem. B,C) Photos from the collaboration activity inside the Museum.

The third dimension of the evaluation concerned investigation of the collaborative nature of the activity and the learning performance. An interesting observation was that by having two teams searching for hints at the same time, and the fact that one of the teams was more successful than the other, constituted a powerful motivation for the second team to search for hints. This was observed from the complementary scene video (pinpointing the event) and the dialog recordings (exclamations etc). Also, we observed that some kids were too excited in using the PDAs and did not allow anyone else to use them. Thus disputes over use of the device influenced negatively the team spirit. From the audio streams we managed to spot the disappointment of the kids that were not allowed to use the device.

Regarding the learning performance through the audio files and the PDA's screen we found out that one team was not reading the descriptions to locate the hints but they were searching for the parentheses that indicated the existence of a hint. We must say that we adopted the solution with the parentheses and not colored text because we wanted to avoid those specific situations, but this didn't actually work in all teams. In the future version the hints will be visible only when the users click on them inside the description of the exhibits.

Our results are also based on study of questionnaires, independently of the ActivityLens analysis. In this point we have to underline the limitation of ActivityLens in analyzing user questionnaires. This weakness is a matter of further development and research.

In order to have a general view about the educational value of the activity when the children returned to their classrooms, they wrote an essay in which they reported on the Museum experience. The teacher’s view after going through these texts was that almost all the kids that participated in the activity learned something meaningful in a funny and enjoyable way. However a more systematic study on these issues should involve a more quantitative experimental approach through a pre and post-test questionnaire and a control group.

CONCLUSIONS

This chapter has presented a brief overview of usability evaluation techniques for mobile applications, including collection of multiple observational data and their analysis. Due to the growing use of mobile devices it is evident that there is a need for established techniques that support the collection and analysis of data while conducting usability evaluations. Since there are considerable differences between desktop and mobile environments, researchers are obliged to develop and fine tune these new techniques. Through this chapter we proposed a methodology for evaluating mobile applications focusing on collection and use of observational data. The proposed methodology was demonstrated through a usability study of an educational game in a Historical Museum.

The proposed recording activity technique can be characterized as unobtrusive regarding the users and allows evaluators to study the activity in conditions as close as possible to the typical conditions of use of the application, through various perspectives. The ActivityLens tool was used for analysis of the collected data which facilitates interrelation and synchronization of various data sources and was found particularly useful, since the collected data were of particularly high volume and often a finding was based on a combination of data sources. The methodology revealed usability problems of the application as well as issues about collaboration and interaction with the environment that would not be easy to discover in the laboratory and without the combined use of the multiple media data.

Studies that take place in semi-public spaces and involve groups of people have to tackle various problems. In most cases the willingness of people but also the availability of spaces is difficult to be guarantied for the long periods of time. Researchers that conduct such studies have to be as unobtrusive as possible to the users and to pay special attention in order to minimize interference with the environment.

A limitation of the proposed approach is that it requires the users to carry light equipment (audio recorders) and also that a screen capturing software had to be installed in the mobile devices. However these limitations did not inhibit the users to act naturally and recreate a realistic but controlled context of use. The typical studies of the proposed approach lasted a short time and thus it is difficult to measure long term usability aspects, like memorability and long term learning attitudes. It is still under investigation how to extend this technique to long term mobile usability studies involving different contexts of use.

What is however missing from our story is an analysis scheme that can describe user interaction with the surrounding physical and information space and metrics that map usability attributes. Such scheme would describe usability as a set of attributes that refer to interaction with the device, interaction with the space and group interactions. This scheme could be supported by a tool like ActivityLens, which facilitates easy navigation of the collected media data, allowing creation of pointers to incidents in the data, justifying the calculated values of the usability attributes. Definition of such scheme should however be the result of a wider research community process.

REFERENCES

Aittola M, Parhi P, Vieruaho M & Ojala T (2004). *Comparison of mobile and fixed use of SmartLibrary*. Proc. 6th International Conference on Human Computer Interaction with Mobile Devices and Services, Glasgow, Scotland, 383-387.

Aittola M, Ryhänen T & Ojala T (2003). *SmartLibrary - Location-aware mobile library service*. Proc. Fifth International Symposium on Human Computer Interaction with Mobile Devices and Services, Udine, Italy, 411 - 416.

Avouris N. , Komis V. , Margaritis M. , Fiotakis G., (2004), *An environment for studying collaborative learning activities*, Journal of International Forum of Educational Technology & Society, 7 (2), pp. 34-41, April 2004 .

Avouris N., Komis V., Fiotakis G., Margaritis M., Voyiatzaki E., (2005). *Logging of fingertip actions is not enough for analysis of learning activities*, Proc. Workshop Usage Analysis in learning systems, AIED 2005, Amsterdam, July 2005, available from http://hcs.science.uva.nl/AIED2005.

Benford S., Rowland D., Flintham M., Drozd A., Hull R., Reid J., Morrison J. and Facer K., (2005). *Life on the edge: supporting collaboration in location-based experiences*, Proc. CHI 2005, Portland, Oregon, USA, 2005

Berkowitz S. (1997), *Analyzing Qualitative Data*, Chapter 4 in User-Friendly Handbook for Mixed Method Evaluations, J. Frechtling, L. Sharp , Westat (ed), NSF Publ. August 1997.

Betiol H. A., de Abreu Cybis W., (2005), *Usability Testing of Mobile Devices: A Comparison of Three Approaches*, Proc INTERACT 2005, pp. 470-481.

Cabrera, J. S., Frutos, H. M., Stoica, A. G., Avouris, N., Dimitriadis, Y., Fiotakis, G., and Liveri, K. D., (2005). *Mystery in the museum: collaborative learning activities using handheld devices*, Proc MobileHCI 2005, Salzburg, Austria, pp. 315-318., 2005

Dey, A., (2001). *Understanding and Using Context*, Personal and Ubiquitous Computing Journal, Vol. 5(1), pp.4-7.

Dix A., Finley J., Abowd G., Beale A., (2004). *Human-Computer Interaction*, Prentice Hall, 2004

Fouskas K., Pateli A., Spinellis D., Virola H. (2002). *Applying Contextual Inquiry for Capturing End-Users Behaviour Requirements for Mobile Exhibition Services*, 1st International Conference on Mobile Business (8-9 July 2002, Athens)

Gay, G. and Hebrooke, H., (2004). *Activity-Centered Design. An Ecological Approach to Designing Smart Tools and Usable Systems*, MIT Press, 2004.

Hagen P., Robertson T and Kan M, (2005). *Methods for Understanding Use of Mobile Technologies*, Technical Report, June 2005. available at http://research.it.uts.edu.au last accessed September 2006

Hagen P., Robertson T., Kan M. and Sadler K., (2005). *Emerging research methods for understanding mobile technology use*, ACM Int. Conference Proceedings, Vol. 122, pp.1-10, 2005

Hesse-Biber, S., Dupuis, P., and Kinder, T.S., (1991). *HyperRESEARCH, a Computer Program for the Analysis of Qualitative Data with an Emphasis on Hypothesis Testing and Multimedia Analysis*, Qualitative Sociology, 14, 289-306.

Jambon F., (2006). *Reality Testing of Mobile Devices: How to Ensure Analysis Validity?*, Proc. CHI 2006, Montreal, Canada.

Kaikkonen A., Kallio T., Kekalaien A., Kankainen A. & Cankar M., (2005). *Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing*, Journal of Usability Studies, Issue 1, Vol. 1, November 2005, pp. 4-16

Kjeldskov J., & Graham C., (2003). *A Review of Mobile HCI Research Methods*, In Mobile HCI 2003, Udine, Italy.

Kjeldskov J., & Stage J., (2004). *New Techniques for Usability Evaluation of Mobile Systems*, International Journal of Human-Computer Studies(60), 599—620.

Kjeldskov, J., Skov, M. B., Als, B. S., & Hoegh, R. T. (2004). *Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field*, Mobile HCI, Glasgow, UK, pp. 61-73.

Nardi, B. (1996) *Studying context: a comparison of activity theory, situated action models, and distributed cognition*. In: B. Nardi (ed.) Context and Consciousness: Activity Theory and Human-Computer Interaction. MIT Press, Cambridge, Mass.

Nielsen J., (1993). *Usability Engineering*, Academic Press, London.

NVivo 7, QSR International, available at: http://www.qsrinternational.com/, accessed on September 2006.

Observer XT. [6.0], Noldus Information Technology, available at: http://www.noldus.com accessed on Septemper 2006.

Raptis, D., Tselios, N., and Avouris, N. (2005). *Context-based design of mobile applications for museums: a survey of existing practices*, In Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services (Salzburg, Austria, September 19 - 22, 2005). MobileHCI '05, vol. 111, ACM Press, New York, NY, 153-160.

ResearchWare, Inc website available at: http://www.researchware.com/hr/index.html, accessed on September 2006

Rich M, Patashnick J. (2002). *Narrative research with audiovisual data: Video Intervention/Prevention Assessment (VIA) and NVivo*. Int. Journal of Social Research Methodology 5(3), pp. 245-261.

Schneiderman B. (2003), *Designing the User Interface*, Addison Wesley

Stoica, A., Fiotakis, G., Cabrera, J. S., Frutos, H. M., Avouris, N. and Dimitriadis, Y., (2005). *Usability evaluation of handheld devices: A case study for a museum application.* Proc PCI 2005, Volos, Greece.

Tselios N., Papadimitriou I., Raptis D., Yiannoutsou N., Komis V., Avouris N., (2006). *Design for Mobile Learning in Museums*, in this volume.

Tennent, P, Chalmers, M., (2005). *Recording and Understanding Mobile People and Mobile Technology*, Proceedings of 1st. Int. Conference on E-social science, First International Conference on e-Social Science, 22-24 June 2005, Manchester UK, http://www.ncess.ac.uk/events/conference/2005/

Transana, available at http://www.transana.org/, accessed on September 2006

Welsh, E. (2002). *Dealing with Data: Using NVivo in the Qualitative Data Analysis Process*. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal], 3(2). Available at: http://www.qualitative-research.net/fqs/fqseng.htm, accessed on September 2006

Zhang D. and Adipat B. (2005). *Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications*, Int. Journal of Human-Computer Interaction, 2005, vol.18 (3), pp. 293-308.