Evaluating the User Experience of Multimodal Public Displays

Tuuli Keskinen, Tomi Heimonen, Markku Turunen, Jaakko Hakulinen, Toni Miettinen
TAUCHI, School of Information Sciences, University of Tampere
Kanslerinrinne 1, FI-33014 University of Tampere, Finland
{firstname.lastname}@sis.uta.fi

ABSTRACT
Interactive, multimodal public displays combine two application areas that are challenging to evaluate for their user experience. The public nature of the displays impacts the use and evaluation of the systems, and the multimodality of interaction complicates the assessment of overall user experience. In this paper we report our experiences from a project where an experience-driven multimodal system for browsing a cultural events program was evaluated in a public library.

Author Keywords
Public displays, user experience, multimodal interaction, evaluation.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Experimentation; Human Factors; Measurement.

INTRODUCTION
Interactive displays have become popular in public spaces such as museums, shopping centers and public transportation hubs. These displays serve many roles, from providing advertisements and public service information to enabling sharing information within organizations and communities. Traditionally, these systems enable interaction via traditional means such as a touchscreen displays, button panels or specialized pointing devices. Our work has recently focused on multimodal public displays that can be used with novel interaction methods such as mobile devices, speech and gestures.

One of the key challenges in developing such systems is their evaluation. On one hand, the evaluation deals with the multimodality of the system – or identifying how the different input and output modalities are utilized to interact with the system and how they affect the user experience. On the other hand, a major part of the evaluation is established by the context of the application – how people approach and use public information displays. Various methods for evaluating both multimodal interaction and interactive public displays have been proposed.

One of the major challenges of multimodal evaluation is that the evaluation criteria are less well understood than in unimodal interaction. Möller et al. [7] propose taxonomy of quality factors to clarify the discussion with respect to multimodal evaluation. Their model consists of three layers: influencing factors (e.g., user, context of use and system features), interaction performance (e.g., user workload, appropriateness of input/output modalities), and quality aspects (e.g., usability, acceptability and perceived quality of interaction). Möller et al. suggest that the taxonomy could serve as the basis for systematic evaluation efforts of multimodal interaction. The present authors adopted a similar approach in an evaluation of gesture-based interaction whereby data on the workload, recognition performance and user experience of the system were collected in the same study. In terms of practical evaluation methodology to address different needs, Kühnel [5] proposes a classification that encompasses predictive evaluation (e.g. PROMISE), experimental evaluation methodologies (e.g., the present authors’ work [10]), and expert evaluation. Similar methods have been suggested also for the evaluation of public displays (e.g., [6] provides heuristics for design and evaluation).

In our experience, experimental, in situ evaluations provide the most actionable data although the aforementioned cost and time issues are a limiting factor. Hazlewood, Stolterman and Connelly [3] also advocate the importance of conducting long-term in-situ evaluations of ambient public displays. They also discuss two key evaluation challenges related with such studies. First, there appears to be a need to tradeoff between the naturalness of the setting and the quality of data collection. On one hand interviews and observation increase the awareness of the system and may affect the results, but they also provide rich information about people’s perceptions with respect to the system. Second, they note discrepancies in self-reported behaviors and actual usage data in their own studies. This suggests that an overall understanding of the user experience requires a critical assessment and triangulation of the various data gathered of the system use.
MULTIMODAL PROGRAM GUIDE

The evaluation work reported in this paper took place in the context of a novel event guide application for exploring event information on public displays. The application is targeted for complex events, such as cultural festivals, which can include a large amount of individual events in numerous geographical locations over a long time period. The overall event guide application consists of two graphical interfaces. The first interface, Word Cloud (Figure 1), is used to gather data from a participant to create a personal event plan for him or her. The second interface, Event Route Map (Figure 2), allows the participant to explore events using a metro map metaphor.

To approach and use the system.

The evaluation work took place in numerous locations around a city.

As the goal of the system was to provide a unique experience, we wanted to emphasize the experiential side also when collecting feedback from the users. We used an existing method called SUXES [10] as the underlying method. SUXES is a method for collecting subjective user feedback of multimodal systems. It includes gathering both expectations before the use and experiences after the use, and thus makes it possible to compare these. The user is asked his or her expectations and experiences on a set of nine statements considering qualities of the system or separate modalities. These statements relate to speed, pleasantness or future use, etc. Answers are provided on a seven-step scale ranging from low to high.

In order to integrate experiential aspects in the subjective measurement data, we appropriated a model called Experience Pyramid [9], which is a tool for analyzing and understanding the experience of tourism products, but the authors state it is suited for studying entertainment, culture-based and design products as well. It is meant to help service providers differentiate and develop their products. Tarssanen and Kylänen [9] present six elements of experience: individuality, authenticity, story, multi-sensory perception, contrast and interaction. They suggest that if these elements are present in all product stages from marketing all the way to post-marketing, the experience can ultimately lead to a personal change. Our evaluation did not go this deep, but instead we used the six elements of experience as the basis for defining the statements to be asked from the participants in a similar way as in SUXES. In order to make the questions as easy to understand as possible, we used bipolar differential approach with a seven-step scale, where the lowest level was represented as a negative and the highest level as a positive statement.

Finally, we used a total of eight statements including the elements of experience, such as “The program guide is unique – there are not similar systems elsewhere” to represent the positive end of individuality, and statements representing the pleasantness of both input modalities (gestures and speech). As is characteristic for SUXES, we
gathered both expectations and experiences from users, and the results can be seen in Figure 3. As some participants did not fill all of the questionnaires, the number of responses varied from 18 (expectations) to 23 (experiences).

![Figure 3. Expectations and experiences (interquartile range shown by box width, median with a diamond).](image)

People had rather high expectations about the system, and the experiences reached at least a neutral level on all statements as well. The high hopes on experiencing something new and contrasting from ordinary life were met, unlike the expectations concerning interaction. In total the expectations were met on half of the statements, but not exceeded on any. According to Wilcoxon Signed Ranks Test there were statistically significant differences between expectations and experiences in three statements: authenticity, interaction and pleasantness of gesture control. Reasons behind these disappointments are quite well explained by the feedback and comments we received from the participants, and by our observations.

**Observing the Interaction**

In order to better understand the interaction and opinions of our users, and possible reasons behind these opinions, we also observed the use of the system. We used a semi-structured observation form to gather information about user behavior, whether they seemed relaxed, surprised, confused and so on, and whether their overall reception towards the system was positive, negative, impressed, disappointed and so on. We tallied frequencies of feature usage, such as which modality the participant mainly used, which hand she/he used, how much time it took to comprehend the functionality of different parts of the system, and how the participant navigated through the system (e.g., how many rounds from word cloud selections to navigating in the metro map the user performed).

We observed 38 users in total, who used the system for about five minutes on average. The majority of participants (30) clearly preferred gestures to speech. The majority behaved in a relaxed manner (25) and their overall reception of the system seemed positive (23). However, some were also confused (12) and a few disappointed (6), which seems like a natural response to interaction problems they faced (as per observations).

Although we experienced a few technical glitches, the most severe issues were the problems with gesture recognition, caused by the operating environment, lighting conditions, people’s clothing and so on. Although such problems in recognizing and tracking the user were by no means constant, they usually caused the user to abruptly lose interest in the system, more or less invalidating the “experience”.

**Comments and Subjective Feedback**

To gather subjective user feedback that cannot be quantified in numerical metrics or by observing, we briefly interviewed users to assess the rationale behind their subjective measurements and behavior. The feedback ranged from positive extremes to very negative. Positive feedback stated that the system was something “new and nice” and “exciting”. More neutral comments suggested improvements to the gestures and operating logic. Some participants also mentioned that detailed instructions would have been needed – indeed some respondents stated that the system would be great if it was easier to use or if they had a better understanding of how to use it.

In terms of content, the respondents wished that the event content had been more extensive and interesting. It is clear that some respondents treated the system as a purely information interface, stating that it felt too clumsy for information retrieval and that they cannot see the purpose or the usefulness of the system. Many, on the other hand, were able to envision different purposes for the system (e.g., browsing products at a storefront), keeping with the spirit of experiential interaction. Finally, some participants commented on the fact that the words presented in the word cloud did not correspond to the results in the metro map. These comments together explain why many respondents experienced the program guide to be more “artificial” and “incredible” than they expected.

We also queried 21 respondents whether they felt using the system was “an unforgettable experience”, whether they would like to use it again, and whether they would recommend it to their friend. The responses to the first two questions were evenly split between the different options (yes, no, can’t say). As perhaps the most interesting finding, in terms of the perceived interest of the system, half of the respondents (10) would recommend the system to their friend and only two respondents would not.

**CHALLENGES IN EVALUATING MULTIMODAL INTERACTION WITH PUBLIC DISPLAYS**

Based on the comments received from the participants and our own observations and discussions we identified quite a few challenges in evaluating multimodal interaction in the context of interactive public displays. Some of them are...
more case or context based while some seem to appear in the majority of evaluations (as suggested also by previous work). Next, we briefly describe some challenges.

- **The amount of instructions.** If detailed instructions are provided to users, it has an inevitable effect on the interaction and it is hard to say much about the intuitiveness of the system. On the other hand, people may be unable to use the system without or with too little instructions, which also has an effect on the interaction. There are also major individual differences between users. Thus, providing instructions to the optimal extent is extremely challenging, especially since public displays have a very short time to engage the passers-by [2]. Although several studies point to the value of providing clear affordances and intuitive interaction (e.g., [1, 4]), this is easier said than done in many cases.

- **Subjective differences between evaluators.** Due to resources, it is rarely possible to have the same observer conduct a whole evaluation, not to talk about several different evaluations. This is problematic as observing and judging behavior, problems and so on is highly subjective. Therefore observed data is rarely fully consistent, even within one evaluation case.

- **Combining data from different sources.** One may have a lot of data consisting of observation notes, videos, interviews, interaction logs and measurements, both objective and subjective. One critical challenge is how we could to some extent automate the combination and analysis of these materials to achieve a holistic appraisal of the user experience?

- **Depth of insights.** Due to having to account both for multimodal features and the peculiarities of the application area, which can also interact, often the findings end up being generic as it can be difficult to tease apart factors affecting the participants’ experiences.

- **Getting people to use public displays.** We found that it is not always easy to encourage people use the public display system, echoing the comments by Brignull and Rogers [1]. Multimodality, especially speech and gestural interaction (which can seem odd to bystanders) exacerbate this problem. As Izadi et al. [4] note, we need to find a way to allow use of the system without making people feel self-conscious.

**CONCLUSION**

We evaluated our multimodal public display in a library and gathered both subjective and objective data. We are yet to complete the data analysis, and different approaches have to be considered in order to gain an overall understanding of the user experience. Although the data collection was not designed to fit the layered model by Möller et al. [7], we might be able to achieve better insights by considering our data according to this framework. For example, since the context of use is critical, its effect should be taken into account more. We also have to consider how and which conclusions can be drawn from the existing data, i.e. how to go beyond basic usability issues and really assess the subjective user experience, the experiential aspects, and success of the system.

**ACKNOWLEDGMENTS**

We thank the staff of Turku City Library and the Turku 2011 Foundation for their assistance with the evaluation effort. This work was funded by the Finnish Funding Agency for Technology and Innovation through its Spaces and Places program. We would also like to thank the Centre for Practise as Research in Theatre at the University of Tampere and Lingsoft for their contributions to the project.

**REFERENCES**


