Gesture and Speech-based Public Display for Cultural Event Exploration

Jaakko Hakulinen (jaakko.hakulinen@sis.uta.fi)
Tomi Heimonen (tomi.heimonen@sis.uta.fi)
Markku Turunen (markku.turunen@sis.uta.fi)
Tuuli Keskinen (tuuli.keskinen@sis.uta.fi)
Toni Miettinen (toni.miettinen@sis.uta.fi)
School of Information Sciences, University of Tampere
Kanslerinrinne 1, FI-33014 University of Tampere, FINLAND

Abstract

We introduce a novel, experiential event guide application for serendipitous exploration of event information on public displays. The application is targeted for complex events, such as cultural festivals, which include a large amount of individual events in numerous geographical locations. The application consists of two interfaces, both used in a multimodal manner with hand gestures and spoken interaction: a three dimensional word cloud is used to select events, which can then be explored using event visualization utilizing “metro map” metaphor. A one-week field study of the application in a public location showed strong bias towards the use of gestures against speech.

Keywords: Multimodal interaction; gestural and spoken interaction; public displays.

Introduction

Interactive public displays provide access to information in public and semi-public areas such as offices, libraries, and shopping centers. What differentiates these displays from traditional information broadcast is that they can be used to provide specific information to the users (Vogel & Balakhrisnan, 2004). However, it is rare to find research that studies these systems in the context of real world installations with focus on the interaction (Hardy, Rukzio, & Davies, 2011). In fact, it has been argued that many interactive public displays utilize an interaction model based on touchscreen use and physical keys, despite several works that propose other modalities (Müller, Alt, Michelis, & Schmidt, 2010).

To address this gap, we introduce a novel spoken and gestural application for public displays. In particular, our aim is to provide a real world application that can deal with large data sets, both in terms of quantity and duration. More specifically, our research questions focus on (1) how gestures and speech are used in this context, and (2) what are the barriers to the use of the different interaction techniques.

The rest of the paper is organized as follows. We first cover related research on the design of interactive public displays. Then, we introduce our novel application and report the key findings of its field study. We conclude by discussing the implications of our research to the design of multimodal public displays.

Background

It is critical how usage possibilities of a public displays are communicated to potential users (Tang et al., 2008). This can be done, for example, by using natural metaphors to illustrate the interaction (Hardy et al., 2011), or visual self-revealing help mechanisms (Vogel & Balakhrisnan, 2004). Because of the short time during which public displays have to engage users, it is challenging to decide how to communicate the interaction techniques (Hardy et al., 2011). In addition, the system should effectively communicate what it has to offer to the user, and establish the participation threshold, i.e., the expected effort and engagement level (Brignull & Rogers, 2003). Ideally, potential users can decide in a few seconds whether they are interested in interacting with the system (Huang, Koster, & Borchers, 2009). In terms of interactivity, it is thus important to visually indicate availability of gesture interaction, and that the system is responsive (Vogel & Balakhrisnan, 2004).

Comprehension of display content can be supported by allowing the information source to affect the way in which the visualization is encoded, for example by using a geographical metaphor as the basis for the spatial layout (Skog, Ljungblad, & Holmquist, 2003). It has also been argued that it is important to maintain a balance between the aesthetics and usefulness of the display (Skog et al., 2003; Mankoff et al., 2003). However, interactions and content that are too predictable may not be appealing. Müller et al. (2010) propose that interactive public displays should contain some sense of uncertainty, which may evoke feelings of curiosity and exploration. Ambiguity, like imprecise representations that emphasize the uncertainty of use (Gaver, Beaver, & Benford, 2003) can encourage personal engagement. This approach was leveraged in our system to create a sense of serendipitous discovery, rather than task-oriented information retrieval.

People’s resistance to participate in the public use of the display can become a significant issue in real world deployments, such as the one described in this paper. Izadi et al. (2005) note that in social settings users should be able to interact with the display without requiring aid or feeling self-conscious. This underlines the affective aspects of public display experience (Brignull & Rogers, 2003). For example, if the system fails to appropriately account for the public nature of interaction, potential users may avoid interaction to maintain a social role (Müller et al., 2010). The social context of use may also aid, as observing use by others may increase interest in the display (Hardy et al., 2011) and provide means of learning the interaction vicariously (Brignull & Rogers, 2003).

Usually public displays are offering information related to the physical environment where they are located. Therefore, all local information helps people spending time at the place to become potential users (Skog et al., 2003), increasing the chances for adoption.

Multimodal Event Guide

Instead of aiming for an effective browsing of event data, multimodal event guide aims at providing an experiential way to browse a collection of cultural events. It consists of two parts; a word cloud interface (Figure 1) for creating an unexpected set of events, and a metro map interface for browsing the event set (Figure 2). Both parts can be operated using either gestures or speech input or a combination of the two. In evaluation setup, the graphical content was presented on a Full-HD TV. User operates the system by standing in front of the TV, where the OpenNI natural interaction toolkit -based tracking system detects user gestures utilizing the Microsoft Kinect sensor. Speech recognition was implemented with a dedicated microphone and Finnish automatic speech recognition (ASR) software.

Word Cloud

Interaction with the public display starts with the Word Cloud, which presents a set of keywords related to the event descriptions. The words are not descriptive keywords, but words associated from the themes of the events. Since the events are dynamically filtered to include only the events in the future, the number of words actually visible in the clouds grows increasingly smaller. In our case, the maximum number of keywords was 44. To form a cloud, the words are positioned randomly around an invisible sphere. Word graphics move, but do not rotate, when the cloud is moved. However, text size reflects the distance of the word from the screen.

The user selects a word from the word cloud, which then shrinks and disappears and a new word cloud appears onto the screen. The user selects altogether three words, which form a sentence (adjective – noun – verb). The selected words appear on the bottom of the screen. Each word is mapped to at least one event, meaning that the three-word sentence results multiple events.

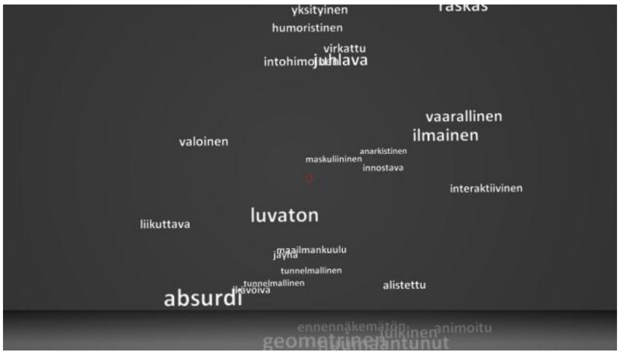
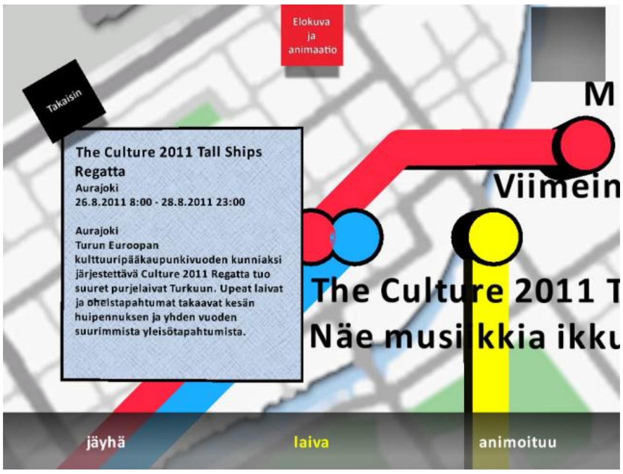


Figure 1: The Word Cloud interface containing keywords related to the event guide content.

Metro Map

A metro map type visualization is generated based on the selected events. Each "stop" corresponds to a set of co-located events. The locations are quantized to a grid so that the lines can run between stops by using cardinal and inter-cardinal directions. If multiple events share the same location, they are considered the same stop. The stops are mapped to distinctly colored metro lines based on event categories (e.g., plays, exhibitions, and musical performances). The actual routes are laid out by starting from a fixed location at the center of the city using an algorithm that aims at minimizing the route length and intersections within the route. Overlapping line segments are displaced vertically and horizontally to create a more readable display. If a stop belongs to multiple lines, it is extended to overlap each line. The resulting visualization is rendered using a 3D graphics engine (jMonkey) so that stylized city map is below the lines and stops, which are given some depth.



The Culture 2011 Tall Ships Regatta
Aurajoki
26.8.2011 8:00 - 28.8.2011 23:00

Aurajoki
Turun Euroopan kulttuuripääkaupunkivuoden kunniaksi järjestettävä Culture 2011 Regatta tuo suuret purjelaivat Turkuun. Upeat laivat ja oheistapahtumat takaavat kesän huipennuksen ja yhden vuoden suurimmista yleisötapahtumista.

Figure 2: The Metro Map interface arranges events by type onto geographically linked "lines" according to the words selected with the Word Cloud interface.

On top of the screen, there is a menu whose elements represent event categories, i.e., the metro lines. When a line is selected the view to the metro map flies into the first stop of the line. Once the view arrives at a stop, details of the events corresponding to the stop appear as a card to the left of the stop and the word(s) corresponding to the events on the bottom of the screen are highlighted. The top menu is updated to include only the lines that travel through the current stop. Further selections move to the next stop in the line. There is always "back" item as the first menu item. It moves to the map overview if the view is at a stop, and goes back to the Word Cloud interface from the overview.

Speech and Gesture Interface

Both the Word Cloud and the Metro Map interfaces can be interacted with concurrent gestural and spoken interaction techniques. Thus both interaction modalities can be used for all functionality of the application, and users are free to choose when they use gestures and when speech. Furthermore, the Word Cloud interface is designed for combined use of these modalities.

Gesturing. In the Word Cloud interface, the spatial location of user's hand is used to rotate the sphere. The sphere is always attached to the hand so that all hand movements rotate it. A single hand movement can rotate the sphere around so that each word can be moved to the front. Individual words are selected by moving the word into the middle of the view, on top of a selection activation area indicated by a small circle. The word must stay in the area for two seconds in order to activate the selection. When a word overlaps the circle, it starts growing in size to indicate the progress of the dwell time.

In the Metro Map interface the gestural interaction is based on menu selections. Menu items are laid out in an arc on top of the screen, and they can be selected by pointing an item for two seconds. This interaction technique is well known from Kinect games. Similar to the Word Cloud, the menu items are animated to indicate the dwell time progress. Visual feedback of the pointing is provided by a shadow image of the hand overlaid on the interface. The shape of the pointer matches the used hand.

The pointing is based on the locations of users' elbow and wrist as detected by OpenNI. Angle of the vector between elbow and wrist is calculated to identify gesturing activity. If the vector points upwards, the hand is considered active and a pointer is shown. The user can use either hand to point, and the one in more upright position is considered to be the active one. The on-screen pointing location of the active hand is calculated from the upright (vertical position) and sideways (horizontal position) rotation of the user's arm. This way the menu is linked to the way menu selections are done, i.e., the users can select different menu items by moving their hand in arc while holding the elbow in place. This proved to be a natural mapping to users during development.

In addition to the menu selections, the user can flip between pages in the event detail area when the selected stop consists of multiple events. This is done by moving the pointer over the card, which then rotates according to the user's horizontal hand movement over the card.

Speech interface. Spoken interaction in both interfaces is based on the "What You See is What You Say" design principle, i.e., the graphical content always includes spoken affordances. In the Word Cloud interface the user can speak any of the keywords in a cloud at any time while manipulating the cloud with their hand gestures. The user can speak a word without the need to bring it to the center of the screen for selection. In this way, the overall interaction metaphor supports parallel, combined use of these modalities. In the Metro Map interface users can do menu selections by speaking an item name, e.g., event category such as *"Exhibitions and visual arts"* as an alternative for using the gestural menu interface.

Evaluation

We evaluated the event guide in Turku, Finland, during its European Capital of Culture year in 2011. The application included information on 4311 cultural events taking place during the year. However, at the time of evaluation in October, only a small amount of events remained active.

During the study, interested passers-by were opportunistically recruited to use the system in the lobby of the city main library. A researcher was present during system use, but initially only informed the participant that the system could be used either by moving hands or by saying out loud the words visible in the Word Cloud. As we wanted to see how intuitive the system is to use, and how people prefer to use it, more instructions were given only after a while if it was clear the user had trouble interacting with the display or they explicitly asked for help.

Observation was done using a specifically prepared data collection form to enable comparison of notes between sessions and researchers. The form included information such as gender, age group, spontaneous comments by the participant, overall perceived state of mind, used modalities, preferred modality, handedness, time taken to understand the logic of different views and modalities, progress through the different interfaces in the system, and approximate duration of use. All of the reported data here is based on the interpretations of the researchers as recorded in the forms. There were no tasks, but instead the participants could use the system as they wished, and thus the observation data is not exhaustive. We also gathered subjective data, but covering that is out of the scope here, and will be done in future publications.

Results

Altogether, we observed 34 people (20 male, 14 female) using the system. The majority of users belonged to the 20–35 year-old (38 %) and 35–50 year-old (26 %) age groups. Both speech and gesture input seem unfamiliar to the users: only one user (of 21) reported to use speech recognition based systems monthly, even 52 % less frequently and 43 % not at all, while gesture-based systems were not used even on a monthly basis by 56 % (of 18 persons) and not at all by 44 %.

System Usage

Usage sessions lasted on average five minutes (range: 2–10 min). Similarly, there was variation in how far the users proceeded in the system. Most of the users (87 %) proceeded to the Metro Map and interacted with it, while 32 % even returned to the Word Cloud and again to the Metro Map at least once more. A clear majority of the users (81 %) used primarily gestures for interacting with the system: only one user used speech primarily, and two users used both modalities alike. For the gestures, 69 % used the right hand exclusively, 24 % both hands, and only 7 % the left hand. The times taken to understand the interaction logic in the Word Cloud (by gesture and speech input) and Metro Map can be seen in Table 1.

Table 1: User percentages on time taken to understand the interaction logic of the application.

Adoption Time Interface	Immediate	< 30 sec	30-60 sec	> 60 sec	Did not use
Word Cloud					
Gestures	3%	42%	35%	16%	3%
Speech	10%	10%	13%	0%	68%
Metro Map	6%	35%	29%	10%	19%

This data suggests that although we attempted to lower the barriers to use during design by leveraging familiar, real life metaphors for the gestures, the interaction affordances were not immediately evident to many users.

Technical challenges. There were some technical problems with the system related to the public setting. The most severe issue was the poor recognition and tracking of the user, which was reported five times. The likely causes were the visual environment, lighting conditions, people's clothing and so on, which affected the tracking accuracy. Unfortunately, these problems usually led to a sudden death of interest. Other issues included unintentional selections by speech or gestures, e.g., selecting a word while talking to someone, or returning back to the Word Cloud from the Metro Map because the user inadvertently held his or her hand in a specific spot. Issues related to the unintentional selections could be solved to some extent by utilizing contextual cues, e.g., in order to identify if the user's attention is on the display when speaking or activating a gesture command.

Discussion and Conclusion

Our research was concerned with establishing how the different modalities are used in the public context of use, and what barriers users encounter during the use. Considering that gestures and speech could both be used to operate the system, the bias towards the use of gestures in the data is very strong. People started using the system with gestures and did not want to either switch to speech, or use it together with gestures. This can be considered somewhat surprising, given that several users had problems in understanding the principles of how hand movements manipulated the word cloud. The microphone was also situated just next to the user, and thus provided perhaps even a stronger affordance towards speech input than the Kinect below the TV for gesture input.

One reason for the prevalence of gestures may be that the participation threshold for gesture use was removed almost completely; when users stepped in front of the system, they were immediately manipulating the word cloud without having to activate any commands. In this respect, it was the "default" interaction technique for the system. In addition, the public location may have made the social situation unfavorable to speech, emphasized by the established norm of silence in libraries, although such was not expected of patrons in the busy lobby area.

Although it is clear our users had some difficulty grasping the principles of the multimodal interactions, it can be argued that in the context of experiential applications some ambiguity is needed to foster exploration and discovery (as also suggested by Müller et al., 2010), unlike with "serious" applications wherein the interaction should not distract from the productive use of the system. The critical issue in design is thus to balance the intuitiveness of interaction and potential for engagement with the application. The potential benefit from the application needs to outweigh the cost of learning the gesturing fundamentals, and in the case of speech, the potential contextual barriers to its use in a public setting. When considering a stand-alone public installation, one way of accomplishing this would be to provide a contextually aware help mechanism that progressively exposes the system functionality, e.g., from demonstrating system use via video to detailed explanation of the interaction techniques.

Acknowledgments

This work was part of the "Space, Theatre & Experience – Novel Forms of Evental Space" project ("DREX"), funded by the Finnish Funding Agency for Technology and Innovation. We gratefully acknowledge Turku City Library, Turku 2011 Foundation, Centre for Practise as Research in Theatre at the University of Tampere, and Lingsoft for their contributions and assistance.

References

Brignull, H., & Rogers, Y. (2003). Enticing people to interact with large public displays in public spaces. *Proceedings of INTERACT'03* (pp. 17–24). IFIP.

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 233–240). New York, NY: ACM.

Hardy, J., Rukzio, E., & Davies, N. (2011). Real world responses to interactive gesture based public displays. *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia* (pp. 33–39). New York, NY: ACM.

Huang, E. M., Koster, A., & Borchers, J. (2009). Overcoming assumptions and uncovering practices: When does the public really look at public displays? *Proceedings of the 6th International Conference on Pervasive Computing* (pp. 228–243). Heidelberg: Springer-Verlag.

Izadi, S., Fitzpatrick, G., Rodden, T., Brignull, H., Rogers, Y., & Lindley, S. (2005). The iterative design and study of a large display for shared and sociable spaces. *Proceedings of the 2005 Conference on Designing for User eXperience* (Article 59). New York, NY: American Institute of Graphic Arts.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003). Heuristic evaluation of ambient displays. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 169–176). New York, NY: ACM.

Müller, J., Alt, F., Michelis, D., & Schmidt, A. (2010). Requirements and design space for interactive public displays. *Proceedings of the International Conference on Multimedia* (pp. 1285–1294). New York, NY: ACM.

Skog, T., Ljungblad, S., & Holmquist, L. E. (2003). Between aesthetics and utility: Designing ambient information visualizations. *Proceedings of the Ninth Annual IEEE Conference on Information Visualization* (pp. 233–240). Washington, DC: IEEE Computer Society.

Tang, A., Finke, M., Blackstock, M., Leung, R., Deutscher, M., & Lea, R. (2008). Designing for bystanders: Reflections on building a public digital forum. *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (pp. 879–882). New York, NY: ACM.

Vogel, D., & Balakrishnan, R. (2004). Interactive public ambient displays: Transitioning from implicit to explicit, public to personal, interaction with multiple users. *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology* (pp. 137–146). New York, NY: ACM.