

Interaction by Movement - One Giant Leap for Natural Interaction in Mobile Guides

Markus Eisenhauer, Andreas Lorenz,
Andreas Zimmermann
Fraunhofer FIT
Institute for Applied Information Technology
Sankt Augustin, Germany
+49-2241-14{2859, 2971, 2561}
{markus.eisenhauer, andreas.lorenz,
andreas.zimmermann}@fit.fhg.de

Townsend Duong, Frankie James
SAP Labs, LLC
Palo Alto, CA (USA)
+1-650-{849-2690, 320-3065}
{to.duong, frankie.james}@sap.com

ABSTRACT

This paper introduces the research issues in multimodal interaction in mobile guides. The goal is to make a major step towards natural human interaction with multimodal systems in mobile systems: we aim to do this by combining explicit and implicit interaction to blended multimodal interaction.

Natural interaction and the requirement to anticipate a mix of purposive behavior and uncertainty requires rendering information in a way that is compatible with the characteristics of the device, the environment, the cognitive load, and the user's personal preferences. Application systems thus need to be aware of the situation and context.

In the description of two recent applications for guiding systems we describe our past research work and experience gained from these applications: SAiMotion is a guide system for professional trade fair visitors, and LISTEN is a museums guide. Motivated by the results of the projects, we illustrate our approach in MICA, which is a hands free support system for blue collar workers.

Categories and Subject Descriptors

H.5.2 User Interfaces: Interaction styles, User-centered design.

General Terms

Design, Human Factors.

Keywords

Multi-modal interaction, natural interaction, context modeling, pro-active support

1. INTRODUCTION

Technology available today allows realizing a wealth of services for mobile people. Such services may support mobility as such, e.g. supply information about transportation or other facilities needed by traveling or commuting people. Or the services simply supply any sort of information and service that may be needed while moving, e.g. shopping, tourism or task-related content. In this paper, we refer to such services as mobile guides.

The success of future mobile systems and applications will be highly affected by user-friendly and intuitive humanlike interaction procedures. The raising concern on accessibility and

usability are pushing research into exploring and employing different modalities of information input and output. For example, if mobile users are moving around, they cannot easily devote all of their visual attention for interacting with a graphical interface [1]. Conversational multimodal interaction will be a paramount requirement to enable ubiquitous computing and therefore for the acceptance of third generation mobile services by the market. Most of the research and design efforts were invested so far in Graphical User Interfaces; nevertheless, experiments and research in multimodal environments have proved the potential efficiency of combining different communication modes.

According to Oviatt's myths of multimodal interaction [2], speech and pointing as the dominant multimodal integration pattern has been widely overrated and other modalities have been culpably neglected so far. Speech as the primary input mode doesn't work (e.g. in noisy environments) and other modalities might be preferred under such circumstances. In our work we will take into account changing environments and address the benefits of other modalities with the goal to achieve natural interaction. We introduce a new approach to interaction that exploits the natural spatial motion of actors and objects to each other.

In our work we employ a user-centered approach¹ in the design of systems supporting blended multimodal interaction. The goal is to make a major step towards natural human interaction with multimodal systems, by endowing these systems with humanlike cognitive capabilities that are necessary to interact with real users, especially in applications for which extensive user training is not feasible.

2. TOWARDS NATURAL INTERACTION

We aim at natural interaction with multimodal systems. Despite the fact that multimodal interaction is a relative young research area, it is already quite diverse. This is because of the large range of interaction modes, communication channels and

¹ We forego to explain our UCD-approach in this paper, however in all mentioned projects we run through several cycles of requirement gathering (scenarios, use cases, on site observations, user questionnaires, and focus groups), implementation and validation.

applications that are under investigation. We are focusing on the type of interaction where nomadic users access information, transaction and support systems at the right time and place, and through a range of input and output devices.

Recent multimodal applications (e.g. [3], [4]) have especially achieved success in support for sensing the multimodal user interaction. This research results in high-end abilities for recognition and synthesis in common modalities such as speech and handwriting. Additionally, current architectures focus on synchronizing events coming from different devices, such as keyboard, mouse, microphone, etc., allowing flexible handling multimodal interactions. In our work, we go beyond observing input from the different modalities to also integrate the recognition and interpretation of meaningful user-related and environmental parameters. For example, the interaction of a user in a museum or a mobile worker with her device will particularly be determined by her current tasks, goals and situation.

So far, all multimodal services and demonstrators have been designed and built by trial and error. This was inevitable, because of the general lack of understanding of the basics of multimodal interaction. Components that relate to semantic interpretation are far from being standardized. Examples include the fusion of input signals into symbolic representations, representations of the system's interpretation of the overt and covert intentions of the user, dialogue management strategies, and representation of the output information in a way that is contextually appropriate. As a result, services tend to come with idiosyncratic interaction styles, which must be learned from scratch by new users. This confronts prospective customers with long and steep learning curves, and it severely reduces the sustainability of systems and services.

Research must ameliorate this state-of-affairs significantly, by focusing on smart sensor combinations for flexible natural interaction with human users. The goal is to exploit natural human-human communication and interaction patterns for their use in human computer interaction. Our approach will be put forward in the next paragraph and will be illustrated by three application examples.

2.1 Implicit and Explicit Interaction

If we speak about implicit interaction, we rely on the understanding of the term implicit as used in the paradigms of implicit Learning [5][6], as a process occurring inescapably and automatically in the absence of conscious, reflected learning strategies and leading to an abstract representation of knowledge. Implicit process are involved whenever peoples knowledge exceeds what could be concluded from their primary task (plausibility criteria) or from what they could report (operationalisation criteria).

Nielsen [7] already voted in 1993 for what he called non-command user interfaces, and Sharon Oviatt called passive modes of interaction [8]. We prefer to call this "*non-verbal implicit user interaction.*" The idea is simply to use implicit natural behaviors that don't require an explicit user command (e.g., gestures and facial expressions) and to extract that information for interaction. However, as these methods tend to be non-obtrusive, they also have the disadvantage of being less reliable than explicit commands.

Users tend to intermix unimodal and multimodal interaction; therefore, a combination of several modalities with explicit and implicit interaction can be used for disambiguation. The interpretation and disambiguation of implicit, as well as explicit, interaction must be resolved through context modeling. In contrast to previous work from Albrecht Schmidt [9] implicit and explicit parameters are both parts of situational parameters for context modeling.

In our approach we seize the suggestion of Oviatt to develop a *blended interface style* that combines explicit and implicit interaction methods. Unlike the IBM MAGIC system, which uses gaze tracking for the prediction of cursor movement [10], we favor the combination of speech and pen input with user movement in his physical environment.

2.2 Interaction by Approaching Objects

The disadvantage of most approaches to multimodal interaction, especially when gesture input is concerned, is that the action to be performed is often circuitous and hard to learn. But more important gesture interaction is mostly socially obtrusive. Excessive gestures performed without a human counterpart look strange to a casual bystander. Our proposal is to take advantage of natural behavior. It's a very natural thing, that if you are interested in something you approach the object to get a closer look and to investigate it in more depth as well that you ignore things that you are not interested in. Exactly that behavior (the attitude of curiosity) will be exploited by our approach for explicit as well as implicit interaction with the user to learn her interests and preferences.

The consequence is that the interaction in a physical space can be matched to the user's interests. The idea is simple and depends on matching the physical and informational space to derive current user preferences and interests. In order to do that the relation between user and his spatial interaction must be defined.

The aim to use a spatial model as a main source to update and refine the user history requires that the relation between user and spatial model is mapped on a domain model. The domain model describes and classifies objects in the domain and which information they contain, as well as the relation to the user model. The spatial model describes the physical environment of the user and the location of the domain objects in the physical space. The user model describes the knowledge, the interests, and the personal preferences of the user. The domain model and the spatial model are assumed to be static, i.e. the domain objects are described and their locations are identified before the system can be used. If changes occur in the environment, the domain model or the space model has to be updated explicitly. The user model is dynamic, i.e. the users' interactions with the information system and their movements in physical space are evaluated to update the user model automatically.

3. APPLICATION EXAMPLES

The rather abstract description of how motion can be interpreted by the system as an interaction is now concretized through three specific examples. The following paragraphs describe completed or ongoing projects of the Fraunhofer Institute FIT making use of the user's movement in the spatial environment as an interaction means.

3.1 A Trade-Fair Information System

SAiMotion [2001-2003] was a national funded project together with five other Fraunhofer Institutes and developed a nomadic information system to support the preparation, the visit and the evaluation of big trade-fairs. SAiMotion benefited to a large extend from the experiences that have been made in former projects.

Beyond known context adaptive approaches, SAiMotion aimed to provide an exhaustive situation model identifying and using all relevant situative parameters. From an HCI perspective the aim was to realize a simplified user interaction and proactive information supply adapted to situation, location, task and user.

The short interaction phases that are typical for a trade fair may not be sufficient to automatically learn an adequate user model. Therefore, SAiMotion used stereotypes to initialize a user model (e.g. business visitor with tight time schedule, journalist or leisure visitor) and refined the model through interaction of the user with the system. Information of the exhibitors are matched and presented according to the interest profile of the visitor. The system supported scheduling of appointments, suggested personalized tours to the exhibition, and gave directions to points of interest. In situations without time pressure, the system was able to suggest exhibitors and exhibits in the vicinity of the user that matched its interests. The system called the user's attention to changes in the schedule and was able to dynamically reschedule.

Location and tracking played a prominent role in achieving above mentioned interaction goals. The project was based upon a sophisticated spatial model that was used as a main source to update and refine the user history. As already explained above the idea is simple and depends on matching the physical and informational space to derive current user preferences and interests.

In SAiMotion we used a WLAN-based tracking system with a precision of 3 to 5 meters. For a large trade fair scenario like the CeBIT 2003 or the MEDICA in 2003 the resolution is sufficient to identify the booth being visited and to use this information to refine the user interest and preferences. The use of location thus is limited to implicit interaction or what Oviatt called passive interaction modes. The precision of the tracking system and the heterogeneity and complexity of a trade fair didn't allow explicit spatial interaction. Nevertheless the analysis of spatial information and current tasks allowed proactive information to points of personal interests. Especially this feature of the system was evaluated very positively by 12 users on the CeBIT 2003 and further 42 users of the system during the MEDICA trade fair in November 2003.

3.2 LISTEN

In October 2003 the LISTEN system was trialed with visitors of the August Macke art exhibition (see Figure 1) at the Kunstmuseum in Bonn, in the context of the "Macke Labor" [11]. Combining high-definition spatial audio rendering technology with advanced user modeling methods creates audio-augmented environments [12]. Visitors are immersed in a dynamic virtual auditory scene that consistently augments the real space they are exploring. The physical environment is augmented through a dynamic sound-scape, which users

experience over motion-tracked wireless headphones for 3-dimensional spatial reproduction of the virtual auditory scene.

While using the LISTEN system, visitors automatically navigate an acoustic information space designed as a complement or extension of the real space [13]. A sophisticated auditory rendering process takes into account the current position and orientation of the visitor's head in order to seamlessly integrate the virtual scene with the real one. Speech, music and sound effects are dynamically arranged to form an individualized and situated sound-scape offering exhibit-related information as well as creating context-specific atmospheres. In addition to the adaptation of the sound scene according to the position and orientation of the user, the audio stream is controlled in two ways: Events (mediated interaction), that are used to start and stop the playback of information items in form of audio recordings, and continuous control (immediate interaction) changing parameters in the audio-generation of the presentation (e.g. a sound that gets continuously louder as you approach a certain position within the space).



Figure 1 The LISTEN System Applied at the Kunstmuseum in Bonn

3.2.1 Evaluation of the preliminary Prototype

In the course of UCD-cycles, we performed evaluation tasks on prototypes of the user modeling component at two LISTEN Expert Workshops with museum curators, artists, and composers. The scenario build up for the workshops was composed of a set of selectable stereotypes and tour recommendations. In order to receive an explicit statement from the user without bothering her during the presentation, we chose to have the selection to be done in advance. To create expressional ones, we have defined three stereotypes: Fact-oriented (with high weight on spoken text), emotional (music pieces and sound effects) and overview (short sound items). The stereotype influenced the character of the sound and the deepness of information the user would hear; whereas the selection of a tour recommendation results in guiding the visitor on a tour being directed by an attractor sound.

Some critical points were especially noticed in the lack of flexibility of the space model: The zones surrounding each artwork were sometimes too small, thus forcing the visitor to

approach the artwork very closely. Particularly for overlapping zones the users could hardly localize the boundaries.

Further effort needed to be put into the recognition of the user's real focus as well: the tracking system senses the visitor's spatial position, but her focus can be on a visual object belonging to another zone. Besides, some visitors could not realize whether the changes in the audio virtual environment were due to their movements in the space or were part of the audio sequence. A minority group did not even experience the personalized character of the audio presentation.

3.2.2 Application of interaction by movement

In order to overcome these problems, we aimed to attach the sound more to the user's behavior in observing visual objects. In this sense, auditory icons providing landmarks in the virtual environment navigation were inserted in the audio presentation to make the user aware of the interaction with the environment. Due to the fact that the visitors did not like to be clustered, and due to the decision of not providing any input devices, we intended to use stereotypes as an internal model for refining the system and to gather more significant information about the user. Providing stereotypes that are meaningful, easy to detect and to revise without disturbing the visitor was the main challenge for user model development in the next project phase. For the final scenario we developed internal stereotypes representing the visitor's style of moving.

For mobile users, *traveling*, *wandering*, and *visiting* were seen as three ways to qualify the essence of mobility [14]. Traveling is defined as "the process of going from one place to another in a vehicle". Wandering, on the other hand, refers to a form of "extensive local mobility" where an individual may spend considerable time walking around. Finally, visiting refers to stopping by at some location and spending time there, before moving on to another location.

Several kinds of common behavior can be identified with people walking through museums (e.g. clockwise [15]). At runtime, the LISTEN system automatically categorizes the visitor into stereotypes. In comparison to [14], we see traveling more as goal-driven relatively fast movement towards a certain destination, wandering as slow sauntering around from one artifact to the other, and finally we split up visiting into two sub-categories depending on whether the visitor is focusing one specific artifact or not. In order to form the basis of adaptive system behavior, the system accomplishes therefore different audio presentations depending on four *motion styles* of the user: "Goal-Driven" "Sauntering around", "Standing, focused", and "Standing, unfocused".

The interpretation of the user's motion style in combination with the location determinates the presentation style and facilitates the pre-filtering process of relevant sound pieces. If the user stands still focusing on the object, object-dependent information is presented. If the visitor moves slowly not being focused on one specific object, a zone-dependent, more general presentation starts. Finally, the selection of one specific sound piece to be played depends on the user's history of already known sounds and personal interests.

The final evaluation touched upon aspects of interaction as well. The compositional structure of the "Macke Labor" allowed a

more distinctive analysis of visitor's acceptance of the introduced forms of interaction, because their effectiveness and functionality could be interpreted more directly. We asked the evaluated persons how they personally experienced the activation of auditory information (technological functionality), emphasizing that they could give several answers: 68% of the 699 evaluated persons acknowledged that the "auditory information seemed to have been activated depending on ones physical movement/position"; only 4% marked "there was no comprehensible connection between ones movement/position and the activation of auditory information." There actually were more than 100 personal remarks on this question. Some people described their experience in terms of having to get used to the system first. We collected statements of the following kinds: "I realized very late how I should behave" or "the longer I moved in the system, the better I could cope with it" as well as positive self-observations such as "one can learn quite easily how to navigate the textual segments." [12]

3.3 MICA

MICA (Multi-Modal Interaction with Context-Aware systems) is a project on behalf of and in co-operation with SAP. The project started in December 2004 running until November 2006. The goal of MICA is to support humans in their working environment in a natural and unobtrusive way and to proactively help them completing their daily tasks.

In a first phase MICA will be implemented in a warehouse setting to support warehouse men in the picking process. The warehouse setting poses a real challenge for multi-modal interaction. In a warehouse the workers are working with their hands thus it requires a hands-free support. The environment is often very noisy and the light conditions might change as well so the interaction might use different modalities according to the needs of the current situation and task. Warehouse men often have to work under time pressure requiring a very responsive system.

The goal regarding interaction in MICA is to provide a combination of explicit and implicit interaction methods in blue collar environments. Beyond the Listen project we face situations in which the spatial relations of objects are changing dynamically and need to be monitored and interpreted in real time. Intelligent fusion of fine grained tracking with RFID technology, in combination with pen or speech input and adaptable audio and graphical output will lead to natural blended interaction with the MICA-system.

In MICA we will combine a low precision WLAN tracking with fine grained Ultra-Wide-Band Tracking, RFID and Camera based motion tracking to be able to determine the spatial and situational parameters determining the interaction. On the one hand the system will be able to identify situations in which help might be needed and react on implicit clues in the interaction like stumbling or search behavior. On the other hand the worker will be able to interact explicitly with the system by approaching objects in the shelf. In particular the combination of implicit and explicit interaction on various modalities will lead to natural blended interaction. It is essential though to improve our understanding of the interactive capabilities that are most important for an automated system to conduct a natural multimodal dialogue.

3.4 Summary

The projects just described show different aspects of the interaction by movement regarding the way user movements are interpreted, and how they are technologically realized. The introduced three projects represent an evolution from implicit to explicit to natural blended interaction combining more and more the spatial interrelations of objects and users (see Table 1).

Table 1: interaction and spatial relations in the projects

	interaction		Spatial relation
	implicit	explicit	
SaiMotion	preferences interests	-	person in space
LISTEN	preferences interests	info on details	single person to objects
MICA	preferences interests	Info on objects	various persons and objects

In SaiMotion only the spatial relation of a user on the trade fair site was used to derive and refine its user profile of preferences and interests. The substantially improved precision of the fine grained tracking system in LISTEN allowed explicit interaction in the exhibition and the interpretation of the spatial relation of the user to objects in the space. Finally the smart combination of different tracking technologies in MICA shows a first example of this interaction paradigm in a blue collar environment. The interaction of warehouse workers in the warehouse will allow for dynamically interpreting their spatial interrelation to each other.

4. CONCLUSIONS AND FUTURE WORK

In this paper we presented our approach to enhance current multimodal applications to consider more than speech and gestures. From our research in context-aware system development we concluded to integrate other non-intrusive modalities to support natural blended interaction. In particular, the movements of the user observed by a tracking system provide valuable implicit and explicit feedback to the system.

Users positively comment their ability to interact with the system only through their movements. By using the body as an interface the user recognizes the interaction means very fast and is able to intuitively use these mechanisms. The presented projects come closer to the goal of intuitive interaction. The handling should need neither a complicated instruction manual nor an introduction and should meet all communicational constraints in a social environment: the offered interaction has to take into account the context avoiding the need to navigate through extensive dialogue steps or menus to get to the currently needed information. Appropriate context modeling though is needed to make a major step in intuitive user interaction and proactive information supply adapted to situation, location, task and user.

The MICA project will make an extensive use of user movements. This project is just finishing the requirement acquisition phase with structured user interviews, scenario and mock-up validation and observations in the field and is currently setting up the hardware and starting the specification and implementation phase. A first prototype will be available at the end of this year, which will be evaluated in extensive user tests.

5. ACKNOWLEDGMENTS

This work was supported by SAP Research, Contract Nr. DE-2004-044

6. REFERENCES

- [1] Brewster, S.A. Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, 6, 3, 2002, 188-205.
- [2] Oviatt, S. Ten myths of multimodal interaction. *Communications of the ACM*, 42,11, 1999, 74 – 81.
- [3] Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. Quickste: Multimodal interaction for distributed applications. *Proceedings of the Fifth ACM International Multimedia Conference*, New York, NY: ACM Press, 1997, 31-40.
- [4] Eccher, C., Eccer, L., Falavigna, D., Nardelli, L., Orlandi, M., and Sboner, A. On the usage of automatic voice recognition in a web based medical application. *Proceedings of ICASSP 2003*, Hong Kong, China, 2003.
- [5] Reber, A.S. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 1998, 219-235.
- [6] Buchner, A. *Implizites Lernen: Probleme und Perspektiven*. München: Psychologie Verlags Union, 1992.
- [7] Nielsen, J. Noncommand User Interfaces. *Communications of the ACM*, 36, 4, 1993, 83 – 99.
- [8] Oviatt, S. Multimodal interface research: A science without borders. *Proceedings of the 6. International Conference on Spoken Language Processing*, 2000.
- [9] Schmidt, A. Implicit Human Computer Interaction Through Context. *Personal Technologies*, 4, 2&3, 2000, 191-199.
- [10] Paterno F. *Mobile HCI 2002*, LNCS 2411. Springer-Verlag Berlin Heidelberg, 2002.
- [11] Unnützer, P. LISTEN im Kunstmuseum Bonn, *KUNSTFORUM International*, 155, 2001, 469-470.
- [12] Zimmermann, A. and Lorenz, A. Creating Audio-Augmented Environments. *Journal of Pervasive Computing and Communication*, 1, 1, 2005, 31-42.
- [13] Eckel, G. LISTEN – Augmenting Everyday Environments with Interactive Sound-scapes. *Proceedings of the 13 Spring Days Workshop "Moving between the physical and the digital: exploring and developing new forms of mixed reality user experience"*, Porto, Portugal, 2001.
- [14] Kristoffersen, S. and Ljungberg, F. Mobility: From stationary to mobile work. In K. Braa, C. Sorensen, and B. Dahlbom, Eds., *Planet Internet*, Studentlitteratur, Lund, Sweden, 2000, 137–156.
- [15] Oppermann, R. and Specht, M. A Context-sensitive Nomadic Information System as an Exhibition Guide. *Proceedings of the Handheld and Ubiquitous Computing Second International Symposium*, Bristol, 2001, 127–142.