Multimodal, Touchless Interaction in Spatial Augmented Reality Environments

Monika Elepfandt and Marcelina Sünderhauf

Technische Universität Berlin, Zentrum Mensch-Maschine-Systeme,
GRK prometei, Franklinstr. 28/29, 10587 Berlin, Germany
{monika.elepfandt,marcelina.suenderhauf}@mms.tu-berlin.de

Abstract. Spatial augmented reality environments are increasing in many domains. However, from a user-centred perspective we still do not know how to interact with them properly. Thus, there is a need for new interaction concepts. In this paper we discuss certain aspects which are special for the interaction in spatial augmented reality environments. We conclude that the most promising form of interaction will be multimodal and touchless. Additionally it should be taken into account that the user is interacting in a 3D space. Therefore different ways of interaction might be more appropriate for different spaces around the user instead of only one. We give a short review of advantages and disadvantages of touchless and multimodal input devices and then outline spatial aspects of perception and interaction in 3D space.

Keywords: spatial augmented reality, touchless interaction, multimodal interaction, multimodality, interaction in 3D space.

1 Introduction

With technological advances, our everyday life is getting more and more augmented with additional information. Large displays are becoming ubiquitious and it will soon be possible to use every physical surface as an interactive display. A commonly accepted definition of such augmented reality (AR) systems is provided by Azuma [1]: These systems combine the real and virtual world, are interactive in real time, and are registered in 3D.

That means that the real world is overlaid with computer-generated virtual information to support the user. The real and the virtual world are combined such that the user interprets the new created three-dimensional environment as a unit and can interact with it in real time. The most commonly known example for AR is the use of head-up displays in aircrafts to visualise airspeed, altitude, the horizon line etc.

Despite the fact that head-mounted displays are still mainly used for AR applications if the user is moving around, spatial AR is a fast-growing field. In these systems the displays are placed and integrated in the environment instead of being attached to the user as with head-mounted or hand-held displays. Normally this is done by using projectors and screens, and the information is directly projected on real objects or even on daily surfaces. Olwal et al. [2], for example, used spatial AR for industrial CNC-machines. The operator sees the real machinery in the workspace

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2011, LNCS 6777, pp. 263–271, 2011.
© Springer-Verlag Berlin Heidelberg 2011

through a display. Process data is presented on the display in such a way that tools and workpieces are overlaid with the visualisations to augment the user's understanding and to simplify machine operation.

The advantages for users are obvious: with spatial displays the users do not have to wear heavy equipment. Hence, these systems are less cumbersome and more ergonomic. On top of that, the field-of-view is less limited compared to head-mounted displays and spatial AR is more convenient for multi-user settings [3], [4].

Application domains for spatial AR are wide-ranging and include industrial settings like control rooms, product design, manufacturing, medical settings like surgery or scientific simulations. Spatial AR applications can also be found in everyday life settings such as office environments, museums or smart homes. For a more detailed description see also Bimber and Raskar [3] or Streitz, Kameas and Myrommati [5].

However, as the development of such systems is mainly technology-driven, we still do not know how to interact with them properly from a user-centred perspective. Though there is a lot of research on interaction techniques, it mostly concentrates on a specific implementation which is often not tested with users or in terms of usability criteria. And even if interaction techniques are tested with usability criteria they are seldom compared with each other. Thus many challenges remain for interaction in spatial AR environments and there is a considerable need for new interaction concepts and design guidelines. As a first step this paper discusses key aspects for interacting in spatial AR, focusing on input devices.

The first aspect is the question how the input should be realised. In most human-machine systems the keyboard, mouse, remote-control unit, and touch display are still the main ways of interaction. But in spatial AR these input devices are no longer suitable. The idea of spatial AR is that the users are moving around a room while interacting with various screens and projections. Hence, fixed controls like a keyboard and mouse would not be suitable as they would limit the users' freedom of mobility.

Neither would the use of remote controls be a good choice. On the one hand the advantage of detaching the user from the technology would get lost and on the other hand the need for frequent gaze shifts between the remote control and the screen can be really annoying and time consuming. Direct touch is also not applicable as the user often stands at some distance from the screen. On top of that there are environments where touch can not be used even if the screen is within reach because they have to remain sterile (e.g. in surgery or laboratory settings). Thus we think that touchless interaction is the most promising way of interaction in spatial AR and we will therefore focus on these input devices here.

The second important aspect is that the user has to interact with screens and projections at various distances. As mentioned above, spatial AR implies that the user is not sitting in front of a display but is moving around. From neuropsychology it is known that the brain constructs multiple spatial representations in order to interact with the world around us and that different spaces around a person are characterized by different functions, e.g. manipulation of objects, visual search, spatial orientation, etc. (see chapter 3). The question is if and how various distances influence the type of interaction in spatial AR.

2 Touchless Interaction

We first offer an overview of the input modalities that are mainly used for touchless interaction – gaze, speech and gesture. Their advantages and disadvantages are outlined from a user-centred perspective. Additionally, the question of unimodal vs multimodal use is discussed.

2.1 Gaze

Gaze is the only reliable predictor of the area of visual attention. During human-human communication it is used by the listener to identify about what (e.g. which person or which object) a person speaks [6], [7]. In addition, gaze is sometimes also directly used to point at objects, people, etc that are around [7]. Since now the research focus of gaze as input modality is the interaction with close displays. Apart from research it is mainly used for people with motor disabilities.

Amongst the advantages are that selection of objects via gaze on a screen seems to be faster than hand-eye coordination (e.g. using a mouse) [8] and that users do not have to learn a difficult gaze control compared to hand-eye coordinated devices [6]. This is due to the fact that users have to look at the target anyway before they act mechanically [9], [10]. Accordingly, Ware and Mikaelian [9] could show that gaze interaction is faster than regular input devices. However, if using gaze as a mode of input it should be considered that the eye is primarily a sensory organ for perception and should not be overloaded with other tasks [11], [12].

In most cases, an object is selected by gaze direction and activated by dwell time, blink or an additional keystroke. The dwell-time method uses the fixation on an object to activate or to select it. The critical point for this common method is the setting for the duration of the fixation. If the dwell-time is too short, the so called “Midas touch” problem occurs [10], i.e. every glance unintentionally activates a command. The user can not look anywhere without selecting or issuing a command. If the dwell-time is too long, the time advantages from fast eye movement disappear [12]. And even with a longer dwell-time, the problem of activating an object accidentally still remains as the user always looks somewhere.

2.2 Speech

To communicate via voice is natural for humans and compared to gaze or gesture it is the modality which is used mostly for direct communication. In human-machine interaction, a main advantage is that speech can be used where the hands are occupied and eyes are busy. Additionally, speech is not only independent from static input devices like keyboard and mouse, but also works during motion and at a distance, which gives the user more freedom and offers more possibilities to solve several tasks at once.

In a lot of empirical studies speech input is compared to standard manual input devices like keyboard and mouse. Benefits are a faster task completion, particularly for tasks that are command intensive [13], and lower error rates than by typing. Martin [14] gives an overview of various studies and also points out that using speech

increases user productivity by providing an additional response channel. Users spend less time looking at keyboards when speech is available which also cuts performance time.

Draper et al. [15] tested these advantages with pilots performing a continuous flight/navigation control task while completing various data entry tasks. Speech input was significantly better than manual input in terms of task completion time, task accuracy, flight/navigation measures, and pilot ratings. Also the data entry time was reduced considerably with speech input. In addition, users prefer to speak rather than type if natural language interaction is chosen as a modality of human-computer interaction [16].

However, there are also disadvantages for using speech as input device. There are many situations where users do not want to be overheard by others - whether for privacy reasons or to avoid disturbances, e.g. in an office setting. Depending on the interface, another disadvantage is that voice commands often have to be learned and remembered, which places additional cognitive demands on the user. Voice commands are not always short and intuitive. In some applications the commands are quite long, detailed and circuitous, e.g. if spatial specifications are required [17]. Additionally speech is not always under full conscious control. During the production of speech, lexical disfluencies (e.g. content self corrections, false starts, verbatim repetitions, spoken pauses) may appear [18] which a user can not completely avoid.

Current research on speech as input modality focuses more on integrating speech in a multimodal setting than using speech on its own (e.g. [19], [20]).

2.3 Gesture

Gestures are an important part of the human-human communication. They are used in everyday life and are self explanatory. Following Fikkert [21] this leads to a joyful and natural use of gestures in human-machine interaction. However, in human-human communication, they normally highlight what is being said and are seldom used on their own. Furthermore, the only gesture which is natural is pointing [7]. All others are learned through cultural influences, and during communication are rarely used on a conscious level.

Hence, for utilizing gestures in human-machine interaction the problem arises that they have to be especially designed and are not necessarily intuitive. Nielsen et al. [22] could show that different user groups produce different gestures for the same task. On top of that the participants used two to three different gestures for the same instruction, although their choice was at a conscious level. If there is not even a real consistency in one person, it seems to be impossible to find gestures that are really intuitive. Thus, gestures in human-machine interaction always have to be learned and remembered by the user and are therefore cognitively demanding. Jetter, Gerken and Reiterer [23] even argue that “symbolic gestures are indirect and resemble learning an artificial sign language to converse with a system about triggering actions”. Therefore, the research focus should not be on symbolic gestures but rather on manipulations as these are more natural.

Apart from that, it should also be taken into account that gesticulating may be physically stressing and tiring if used frequently or for a long time. Users prefer gestures which require a minimum of effort and rate fast and easy switching between resting and interacting as very important [21].

2.4 Multimodality

Instead of using gaze, speech and gesture as unimodal input devices – with their advantages and disadvantages – they can also be combined and used to provide a multimodal input. In human-human communication, speech and gesture are normally used multimodaly for complementary information. Speech typically specifies subject, verb, and object information, while manual gestures convey location, number, manner of movement and related information [24]. As described earlier, gaze is also sometimes used to point at objects or people.

In human-machine interaction, users have also a strong preference for multimodal rather than unimodal interaction across a wide range of various application domains [25]. Here again modalities are used to complement one another and not redundantly. Bolt [26] already used speech and gesture complementary in his “Put-That-There” interface. The user selected an object or location on the interface by pointing at it and gave a spoken command (e.g. "Create a blue square there.").

However, the preference is very task dependent. Oviatt, DeAngeli and Kuhn [17] conducted a task analysis of speech and pen input for a dynamic map system. They found that the use of multimodal commands depend on the type of action which has to be performed. Users were most likely to use speech and pen in spatial location commands (adding, moving, modifying and calculating distances). A moderate use of multimodal interaction occurred during selection commands (querying, deleting, labeling and zooming), while during general action commands (e.g. scrolling, printing and locating) a marginal level of multimodality was found (i.e. less than 1% of the time).

Tan et al. [27] tested gaze and speech as a data entry method and showed that users prefer the multimodal use of gaze and speech compared to other methods like handwriting, mouse and keyboard, or speech only. The multimodal input was neither the fastest nor the most accurate, but obtained the highest hedonic quality and naturalness according to the users. Besides, Oviatt [28] reports that errors - in multimodal interaction - lead to less subjective frustration than errors in unimodal interfaces.

Comparing speech-only input with speech-gesture input while performing map-based tasks, the multimodal input results in more efficiency of speech (fewer spoken words), fewer task performance errors, fewer disfluencies (e.g. content self corrections, false starts, spoken pauses), and faster task performance [29]. Also Oviatt et al. [17] showed that speech commands get shorter and more explicit if used to complement another modality.

Thus there are a lot of advantages adding complementary multimodality to an interface. It is quite important to consider which modalities to combine and for which subtask each modality should be used in combination with other ones. For instance, studies show performance benefits when users indicate spatial objects and locations (e.g. points, paths, areas, groupings and containment) through gestures instead of speech [28], [29], [30]. On the other hand, speech is more useful than gesturing for specifying abstract actions [20].

Summarizing, we have shown that gaze, speech and gesture have different advantages and disadvantages for use as input modalities. Their benefits and weaknesses should be taken into consideration when designing new interaction

techniques for spatial AR environments. Thus using only one modality does not seem to be appropriate. Furthermore, as the interaction takes place in 3D space most of the tasks will have a spatial component. Therefore we think that a more multimodal approach is the most promising for interaction in spatial AR.

3 Interaction in 3D Space

As already mentioned, an important challenge for interaction in spatial AR is that the user is no longer sitting in one place but is moving around in a 3D space, interacting with various screens and projections at different distances. Thus, spatial aspects of perception and interaction should always be considered in spatial AR environments.

Neuropsychological studies have discovered that the brain builds multiple spatial representations in order to interact with the world around us [31] and that particular representations are created to guide certain actions like eye movements, head movements or arm movements [32]. According to Previc [33] "our behavioral interactions in each region of the 3-D space are associated with a characteristic 3-D locus of operation, a select set of sensory-perceptual and motor operations, and a predominant coordinate system". In his work on 3D spatial interaction he integrated neuropsychological findings into a model of four major behavioral realms: (1) peripersonal, (2) focal extrapersonal, (3) action extrapersonal, (4) ambient extrapersonal (see Fig. 1).

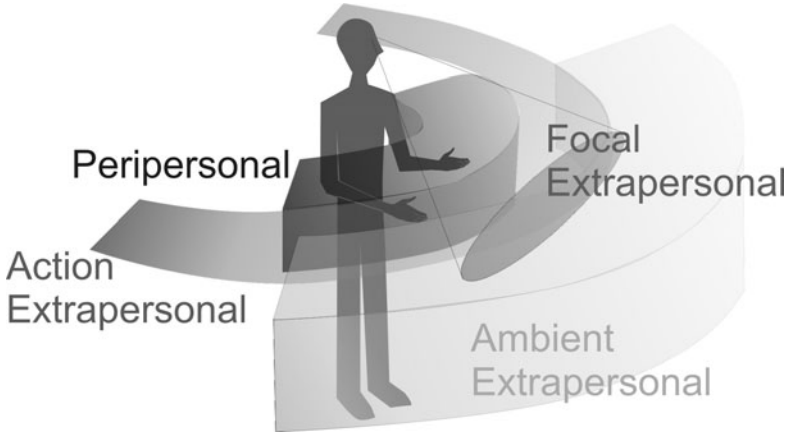

Fig. 1. A theoretical model of 3D spatial interaction, adapted from Previc [33]

The peripersonal realm is within hand reach, the three extrapersonal realms are out of reach. Therefore the realms also depend on an individual's arm-length [34]. The peripersonal behavioral system is characterized by grasping, reaching and manipulating objects, while the extrapersonal behavioral systems are characterized by the visual search for and recognition of objects (focal extrapersonal), navigation and target orientation (action extrapersonal), and spatial orientation, postural control and locomotion (ambient extrapersonal).

For interaction at varying distances, the differentiation between the peripersonal and the extrapersonal space is most important. As the peripersonal realm is in reach, it is highly optimized for direct interaction because this is the only space in which we are naturally able to manipulate objects. Other authors refer to the manipulative realm [35] or grasping space [36]. It can also be extended by the use of instruments [36], [37].

Objects beyond this space can normally not be reached or manipulated without moving to them. Naturally the main functions of the extrapersonal realms are more perceptive and visual and are not constructed for direct interaction. But what happens if we have to select or manipulate an object on a distant screen, i.e. in extrapersonal space? Or if we have to interact with objects on a screen that is in the peripersonal space but which we are not able to touch? Is there a difference for interaction in different realms? Or is the only difference the ability to touch? These questions should be considered by developing new interaction concepts for spatial AR environments.

4 Conclusion

In many application domains such as industry, research, and even home environments, spatial AR systems are increasing. Interaction with these systems is one of the most basic but also critical components. Unfortunately, input devices designed for traditional computers are not always suitable for new technological approaches – particularly not for interaction in spatial AR.

Although a lot of research has been carried out on interaction techniques we still do not know how to interact with these environments in a suitable way because the research is often more technologically driven than user-centred. As a first step towards a more user-centred design, this paper pointed out and discussed certain aspects that are special for interaction in spatial AR. The following aspects should be considered for input devices:

First, as the users are moving around in a room while interacting with different screens and projections at various distances the most promising way of interaction will be touchless interaction.

Second, as the user is interacting in 3D space most or all of the tasks will have a spatial component. Thus the design of input devices should focus more on multimodal interaction as users prefer multimodality for spatial tasks.

Third, as the brain builds multiple spatial representations in order to interact with the world around us, and as different spaces around a person are characterized by different functions, it should be taken into account that different modes of interaction might be more appropriate for various distances.

Many challenges remain for the interaction in spatial AR environments and there is an urgent need for new interaction concepts and systematic research from a user-centred perspective.

Acknowledgments. We want to thank the Deutsche Forschungsgemeinschaft, the Chair of Human-Machine Systems Berlin, and the research training group prometei for their support.

References

1. Azuma, R.T.: A survey of augmented reality. Presence-Teleoperators and Virtual Environments 6, 355–385 (1997)
2. Olwal, A., Gustafsson, J., Lindfors, C.: Spatial augmented reality on industrial CNC-machines. In: SPIE Electronic Imaging, vol. 6804 (2008)
3. Bimber, O., Raskar, R.: Spatial Augmented Reality: Merging Real and Virtual Worlds. A.K. Peters, Wellesley, Massachusetts (2005)
4. Zhou, F., Duh, H.B.-L., Billinghurst, M.: Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 193–202. IEEE Press, Washington (2008)
5. Streitz, N., Kameas, A., Mavrommati, I.: The Disappearing Computer. Springer, Heidelberg (2007)
6. Kaur, C.S., Tremaine, M., Huang, M., Wilder, N., Gacovski, J., Flippo, Z., Mantravadi, F.: Where is it? Event synchronization in gaze-speech input systems. In: ICMI, pp. 151–158. ACM, Vancouver (2003)
7. Poggi, I.: Mind, hands, face and body. In: A goal and Belief View of Multimodal Communication. Weidler, Berlin (2007)
8. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: SIGCHI Conference on Human Factors in Computing Systems CHI 2000, pp. 281–288. ACM Press, New York (2000)
9. Ware, C., Mikaelian, H.H.: An evaluation of an eye tracker as a device for computer input. ACM SIGCHI Bulletin 17, 183–188 (1987)
10. Jacob, J.K.: The Use of Eye Movements in Interaction Techniques: What You Look At is What You Get. ACM Transactions on Information Systems 9, 152–169 (1991)
11. Zhai, S., Morimoto, C., Ihde, S.: Manual and gaze input cascaded (MAGIC) pointing. In: SIGCHI Conference on Human Factors in Computing Systems CHI 1999, pp. 246–253. ACM Press, New York (1999)
12. Drewes, H., Hußmann, H., Schmidt, A.: Blickgesten als Fernbedienung. In: Mensch & Computer, pp. 79–88. Oldenburg-Verlag, Weimar (2007)
13. Karl, L., Pettey, M., Shneiderman, B.: Speech-Activated versus Mouse-Activated Commands for Word Processing Applications: An Empirical Evaluation. International Journal of Man-Machine Studies 39, 667–687 (1993)
14. Martin, G.L.: The utility of speech input in user-computer interfaces. International Journal of Man-Machine Studies 30, 355–375 (1989)
15. Draper, M., Calhoun, G., Ruff, H., Williamson, D., Barry, T.: Manual versus speech input for unmanned aerial vehicle control station operations. In: Human Factors & Ergonomics Society's 47th Annual Meeting, pp. 109–113 (2003)
16. Cohen, P.R., Oviatt, S.L.: The role of voice input for human-machine communication. National Academy of Sciences of the United States of America 92, 9921–9927 (1995)
17. Oviatt, S., DeAngeli, A., Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction. In: SIGCHI Conference on Human Factors in Computing Systems CHI 1997, pp. 415–422. ACM Press, New York (1997)
18. Oviatt, S.: Multimodal interfaces for dynamic interactive maps. In: SIGCHI Conference on Human Factors in Computing Systems Common Ground - CHI 1996, pp. 95–102. ACM Press, New York (1996)
19. Kammerer, Y., Scheiter, K., Beinhauer, W.: Looking my way through the menu: the impact of menu design and multimodal input on gaze-based menu selection. In: Symposium on Eye Tracking Research & Applications, pp. 213–220 (2008)

20. Tse, E., Shen, C., Greenberg, S., Forlines, C.: Enabling interaction with single user applications through speech and gestures on a multi-user tabletop. In: Working Conference on Advanced Visual Interfaces - AVI 2006, pp. 336–343 (2006)
21. Fikkert, F.W.: Gesture Interaction at a Distance. University of Twente, Centre for Telematics and Information Technology (2010)
22. Nielsen, M., Störring, M., Moeslund, T.B., Granum, E.: A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 409–420. Springer, Heidelberg (2004)
23. Jetter, H.-C., Gerken, J., Reiterer, H.: Natural User Interfaces: Why We Need Better Model-Worlds, Not Better Gestures. In: CHI 2010 Workshop - Natural User Interfaces: The Prospect and Challenge of Touch and Gestural Computing, pp. 1–4 (2010)
24. McNeill, D.: Hand and Mind. In: What Gestures Reveal about Thought, The University of Chicago Press, Chicago (1992)
25. Oviatt, S.: Multimodal Interfaces. In: Jacko, J.A., Sears, A. (eds.) A Handbook of Human-Computer Interaction, pp. 414–428. Lawrence Erlbaum Associates, New Jersey (2002)
26. Bolt, R.A.: 'Put-that-there': Voice and gesture at the graphics interface. In: 7th Annual Conference on Computer Graphics and Interactive Techniques, pp. 262–270. ACM Press, New York (1980)
27. Tan, D.S., Gergle, D., Scupelli, P., Pausch, R.: With similar visual angles, larger displays improve spatial performance. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 217–224. ACM Press, New York (2003)
28. Oviatt, S.: Ten myths of multimodal interaction. Communications of the ACM 42, 74–81 (1999)
29. Oviatt, S.: Multimodal interactive maps: Designing for human performance. Human–Computer Interaction 12, 93–129 (1997)
30. Cohen, P.R., Johnston, M., Mcgee, D., Oviatt, S., Box, P.O.: QuickSet: Multimodal Interaction for Distributed Applications. In: 5th ACM International Conference on Multimedia, pp. 31–40. ACM Press, New York (1997)
31. Colby, C.L.: Action-oriented spatial reference frames in cortex. Neuron 20, 15–24 (1998)
32. Colby, C.L., Duhamel, J.R.: Spatial representations for action in parietal cortex. Cognitive Brain Research 5, 105–115 (1996)
33. Previc, F.H.: The Neuropsychology of 3-D Space. Psychological Bulletin 124, 123–164 (1998)
34. Longo, S.F., Lourenco, M.R.: Space perception and body morphology: extent of near space scales with arm length. Experimental Brain Research 177, 285–290 (2007)
35. May, M.: Raum. In: Funke, P.A., Frensch, J. (eds.) Handbuch der Allgemeinen Psychologie – Kognition, pp. 66–74. Hogrefe, Göttingen (2006)
36. Grüsser, O.-J.: Multimodal structure of extrapersonal space. In: Hein, A., Jeannerod, M. (eds.) Spatially Oriented Behavior, pp. 327–352. Springer, New York (1983)
37. Holmes, N.P., Spence, C.: The body schema and the multisensory representation(s) of peripersonal space. Cognitive Processing 5(2), 94–105 (2004)