Conversational Pointing Gestures for Virtual Reality Interaction: Implications from an Empirical Study

Thies Pfeiffer* Marc E. Latoschik† Ipke Wachsmuth‡

AI Group, Faculty of Technology
Bielefeld University, Germany

Abstract

Interaction in conversational interfaces strongly relies on the system's capability to interpret the user's references to objects via deictic expressions. Deictic gestures, especially pointing gestures, provide a powerful way of referring to objects and places, e.g., when communicating with an Embodied Conversational Agent in a Virtual Reality Environment. We highlight results drawn from a study on pointing and draw conclusions for the implementation of pointing-based conversational interactions in partly immersive Virtual Reality.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction Styles; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques

1 Introduction

Deictic expressions (such as "put *that there*") are fundamental in human communication to refer to entities in the environment. In situated contexts, deictic expressions often comprise pointing gestures directed at regions or objects. One of the primary applications of Virtual Reality (VR) is the manipulation of visually perceivable objects. Therefore the system's capability to select relevant objects is crucial. VR research has thus focused on developing metaphors optimizing the tradeoff between a swift and precise selection of objects. For an overview see [1]. However, these approaches are mainly targeted at direct manipulation tasks.

When the interaction with the system is mediated, e.g., by an Embodied Conversational Agent (ECA), the primary focus lies on a smooth understanding of natural communication. In such systems, users communicate their goals to the ECA, which represents the system. And they will inevitably use deictic expressions and pointing gestures. Thus the system needs to infer the semantic/pragmatic extension of a detected pointing gesture, i.e., the demonstrated objects or regions. We report results from a study on pointing at objects conducted in collaboration with linguists [3]. Although the study investigates pointing behavior in a real world context, it provides insights for improvements of conversational VR interfaces.

2 Background

There is excellent work on object selection in VEs for direct manipulation, which can be roughly summarized as following either ray casting, occlusion, or arm extension approaches (e.g. [7]), providing rich insight into the way users deal with the technical interfaces. However, targeting at human-like conversational interfaces, genuineness is preferred over performance: We aim at models for interpreting pointing that apply to human-human as well as human-machine communication, both in real and virtual reality.

*e-mail: thies.pfeiffer@uni-bielefeld.de
†e-mail: marcl@techfak.uni-bielefeld.de
‡e-mail: ipke@techfak.uni-bielefeld.de

IEEE Virtual Reality 2008
8-12 March, Reno, Nevada, USA
978-1-4244-1971-5/08/$25.00 ©2008 IEEE

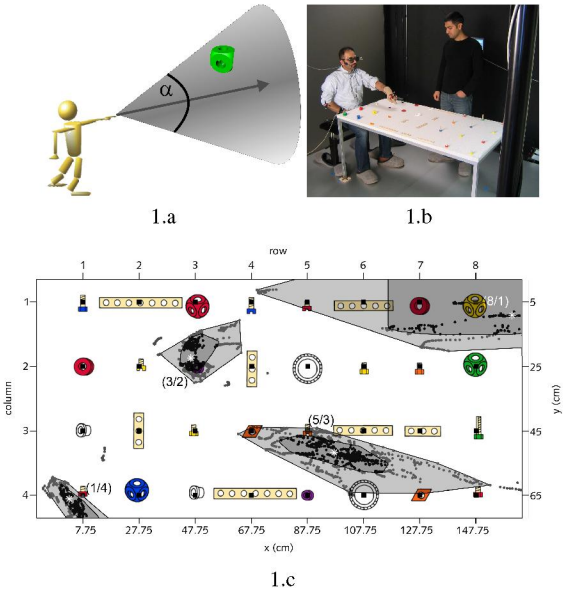

Figure 1: To refine cone-based models for the extension of pointing gestures (1.a), a study has been conducted where participants engaged in an identification game (1.b) over a set of objects arranged on a table. For selected demonstrations from the recorded data, (1.c) shows bagplots of the intersections of the pointing rays with the table for all participants (pointing from left to right).

Basic shapes have successfully been used to model the extension of a modality (i.e. the region of influence), e.g., the *SenseShapes* [5] used in a multimodal interaction architecture [2] resembling the VIENA system [6]. We reviewed and mined data from a study on pointing to extract parameterizations for such shapes. The work we are presenting builds upon our own work on deixis over years (VIENA [6], SGIM [4]). Deixis has also been one of the central topics within the interdisciplinary Collaborative Research Center 360, "Situated Artificial Communicators", running at Bielefeld from 1993 to 2005, in the context of which the study presented in the next section was conducted [3].

3 Method

We conducted a study with 22 pairs of unacquainted participants playing an object identification game (Figure 1.b): one pointing, the other identifying the objects pointed to (for details see [3]). The objects in this game were arranged on a table as shown in Figure 1.c. The participants were recorded using video (2 perspectives) and motion capturing (nine camera optical tracking system from *Advanced Real-time Tracking GmbH*). The games have been annotated and altogether 704 direct pointing gestures, restricted to one per game, have been identified.

Table 1: **Optimal apex angles and performances** per row r for the Gaze-Finger Pointing and Index-Finger Pointing, one for hits and one for successes (see text). The lower part of the table shows angles with the best overall performance for the proximal (p), distal (d), and for both (b) areas. The best performances are highlighted for each category and row.

r	IFP				GFP			
	hit		success		hit		success	
	α	perf.	α	perf.	α	perf.	α	perf.
1	84	**70.27**	120	**98.65**	86	68.92	143	**98.65**
2	80	61.84	109	**100**	68	**75**	124	**100**
3	71	71.43	99	**94.81**	69	**81.82**	94	93.51
4	60	53.95	109	**98.68**	38	**65.79**	89	93.42
5	36	43.84	72	**97.26**	24	**57.53**	75	94.52
6	24	31.15	44	**91.8**	25	**42.62**	50	90.16
7	14	**23.26**	38	**86.05**	17	**23.26**	41	67.44
8	10	7.14	31	52.38	10	**14.29**	26	**69.05**
p	79	56.11	120	**98.02**	69	**67.66**	143	96.37
d	35	27.68	72	**92.66**	23	**40.11**	75	86.44
b	71	38.54	120	**96.04**	61	**48.12**	143	92.71

4 Results

In simulations we tested cone-based extension models with varied parameters against the recorded data. For the orientation of the cone, we differentiated between *Index-Finger-Pointing* (IFP), were the cone is projected into the direction of the index finger, and *Gaze-Finger-Pointing* (GFP), were the direction is determined by projecting a ray from an imaginary cyclop's eye located between the eyes of the user aiming over the tip of the index finger.

For the interpretation it is also relevant whether the pointing cone is sufficient to single out the target object, i.e., only the target object lies within the cone, which we call a direct *hit*, or if additional heuristics are needed. We restricted us to simple heuristics weighting the angular distance between the objects and the axis of the cone. If the target object stands out based on the heuristics, we call it a *success*.

We tested these parameterized models with varying apex angles α (Figure 1.a). The results are shown in Table 1, condensed to an entry per row (Figure 1.c) and per region (proximal 1-4, distal 5-7, all 1-7), clustered by being within/without grasping reach. The last row was excluded from the regions, as the participants showed a very specific border-of-the-domain behavior, which we will have to investigate further.

5 Conclusion and Future Work

From these results we draw conclusions, highlighted in bold. Comparing successes with hits, our simulations confirm that in our setting even **a simple heuristics performs better than the one-or-nothing pointing cone**. Consequently, pointing gestures have to be interpreted pragmatically and are not available for an earlier semantic analysis. This has to be considered in the linguistic theory underlying any multimodal interface.

Over the full region, IFP performs as well as GFP. The latter being less precise than IFP (successes), which might be due to an amplification of jitter, as two tracked modalities contribute to GFP. When accuracy is needed (hits), GFP performs better than IFP. Thus the **effort for including gaze into the extension model has to be considered carefully**. At least in the setting used in our study, with widely spaced objects (20 cm), it can be ignored when going for high overall success.

The results (Table 1) and our observations of the video recordings suggest that it is useful to **distinguish between proximal and distal pointing** (Figure 2) using differently parameterized shapes. This fits nicely into the dichotomy common in many languages

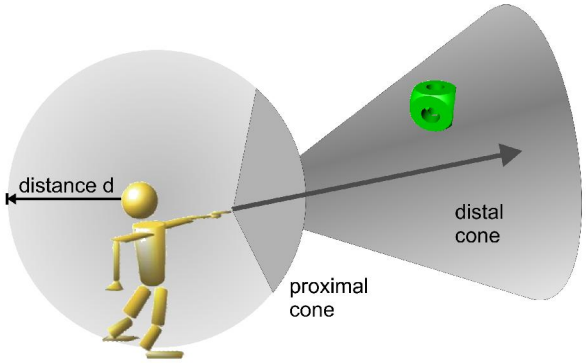

Figure 2: Distinguishing proximal and distal pointing improves performance when using appropriate cones for the interpretation of pointing gestures. The range of the proximal cone is determined by the radius d of the proximal personal area, i.e., the grasping space.

with deictic expressions (*here* vs. *there*). The distinctive feature is whether the user is part of the area (proximal) or not (distal).

In the presented study, two humans communicated over a set of real objects, tracked by VR interaction technology. We are currently mining the data and prepare a more thorough report. A study to follow will replicate the setting with users partly immersed, communicating with an ECA about virtual objects. The extension models derived from the real world setting will then be evaluated in the virtual setting and vice versa, aiming at a generalized model for both.

Acknowledgements

This work has been partly funded by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center 360, "Situated Artificial Communicators" and the EC in the project PASION, FP6 - IST program - reference number 27654.

References

[1] D. A. Bowman, E. Kruijff, J. Joseph J. LaViola, and I. Poupyrev. *3D User Interfaces – Theory and Practice*. Addison-Wesley, 2005.

[2] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 12–19. ACM Press, 2003.

[3] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and M. Staudacher. Measuring and Reconstructing Pointing in Visual Contexts. In D. Schlangen and R. Fernández, editors, *Proceedings of the BRANDIAL 2006 - The 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 82–89, Potsdam, 2006. Universitätsverlag Potsdam.

[4] M. E. Latoschik. A Gesture Processing Framework for Multimodal Interaction in Virtual Reality. In *Proceedings of the 1st International Conference on Computer Graphics, Virtual Reality and Visualisation in Africa, AFRIGRAPH 2001*, pages 95–100. ACM SIGGRAPH, 2001.

[5] A. Olwal, H. Benko, and S. Feiner. SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System. In *Proceedings of The Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*, pages 300–301, Tokyo, Japan, October 7–10 2003.

[6] I. Wachsmuth, B. Lenzmann, T. Jörding, B. Jung, M. Latoschik, and M. Fröhlich. A Virtual Interface Agent and its Agency. *Proceedings of the First International Conference on Autonomous Agents*, pages 516–517, 1997.

[7] C. A. Wingrave, D. A. Bowman, and N. Ramakrishnan. Towards Preferences in Virtual Environment Interfaces. In *EGVE '02: Proceedings of the Workshop on Virtual Environments 2002*, pages 63–72, Aire-la-Ville, Switzerland, Switzerland, 2002. Eurographics Association.