A Cognitive Perspective on Gestures, Manipulations, and Space in Future Multi-Device Interaction

Conference Paper · April 2014

CITATIONS
0

READS
180

1 author:

Hans-Christian Jetter
University College London
71 PUBLICATIONS **428** CITATIONS

SEE PROFILE

A Cognitive Perspective on Gestures, Manipulations, and Space in Future Multi-Device Interaction

Hans-Christian Jetter
Intel ICRI Cities, University College London
Gower Street, London, WC1E 6BT, UK
h.jetter@ucl.ac.uk

ABSTRACT
In this position paper, I introduce my view of gestures, manipulations, and spatial cognition and argue why they will play a key role in future multi-device interaction. I conclude that gestural input will greatly improve how we interact with future interactive systems, provided that we fully acknowledge the benefits of manipulations vs. gestures, do not force users to interact in artificial gestural sign languages, and design for users' spatial abilities.

Author Keywords
gestures; manipulations; gesture sets; space; spatial memory; multi-device; cross-device; ad hoc.

ACM Classification Keywords
H.5.2. User Interfaces: Input devices and strategies.

INTRODUCTION
During the last decade, the rapid advances in sensor and display technology, CPUs, GPUs, and wireless networks have enabled a new generation of novel computing devices with a great variety of form factors and interaction styles, e.g., smart phones, smart TVs, mobile tablets, mobile projectors, tabletop computers, large multi-touch and pen-enabled whiteboards and tables, wearable computing such as smart watches or augmented reality glasses. Through these new devices, we have advanced from the *"personal computer era"* with *"one computer per user"* to the "*mobility era"* with *"several computers per user"* and we expect to advance to the "*ubiquity era"* with *"thousands of computers per user"* in 2020 and beyond [8].

In this coming era, the core challenge of HCI will be to understand, design, and implement user interfaces for a "natural" and efficient interaction with not only a single screen, device, or gadget. Instead we will have to focus on a seamless cross-device interaction with always changing *ad hoc* communities of several or many co-located devices. Current HCI research already reflects the critical role that a seamless interaction with multiple displays and devices will play, e.g., the use of second or third screens while viewing TV [2], cross-device interactions and collaboration using multiple co-located tablets or phones [9, 21, 22, 24, 27] and how to track their spatial configuration [10, 20, 23, 24], *proxemic interactions* based on spatial relations such as distance, orientation, and movement of devices [7], cross-device interaction in multi-tablet environments for active reading [1], or also my own work on collaborative sensemaking in multi-device environments with large screens and tabletops [12, 18]. The future will bring us an even greater wealth of different interactive devices that we will use in concert in countless different contexts within our workplaces, homes, and public spaces of our future cities.

Our challenge will be to design a seamless and efficient gestural interaction with these always changing *ad hoc* communities of interactive devices while providing the necessary flexibility, scalability, and robustness for unanticipated uses [15, 16]. I believe the key to addressing this challenge is a better understanding of and a cognitive perspective on the roles of gestures, manipulations, and the physical space in which they are performed. Therefore, in the following, I introduce my view of how these different concepts are related and I illustrate how it affects our current thinking about gestures, manipulations, and space.

THE STARTING POINT: GESTURES
There are many reasons for using gestures as user input in HCI. The most obvious ones are those based on context-specific or domain-specific requirements: For example, for a screen showing medical images in a sterile operating theatre or for a non-touch public display, gestural UIs with waving or pointing are a natural choice, because touch input has to be avoided or is impossible. Another example is the games domain for which Microsoft's Kinect demonstrates how gestural and full-body interaction of multiple co-located players can entirely change the players' experience and introduce a novel source of fun and motivation into gaming. However, another sort of argument for gestures is far less convincing and, in my opinion, does not hold after closer scrutiny: Technologists frequently claim that gestural input makes computing more "natural" by enabling

communication with a computer the same way we also communicate with one another. I call this the *"gestures make computers more human"* hypothesis.

The flaw in this hypothesis is that the goal of letting a computer interpret natural human gesturing touches on unsolved grand challenges of computer science such as passing the Turing test or achieving strong AI. It is even more difficult than making a computer *understand* natural language. Developmental psychologist Michael Tomasello whose research is in the areas of cognitive development and comparative psychology between human and non-human primates argues that, compared with conventional human languages, gestures are very weak communicative devices, as they carry much less information "in" the communicative signal itself [32]. He illustrates this with the example of pointing at a bicycle in front of a library: When a person draws the attention of another person to the bike this way, this can mean anything ranging from "this is a nice bike" or "the library is still open" to "your ex-boyfriend is here", depending on the persons' context and shared experience. This resonates with linguistic anthropologist Charles Goodwin, for whom pointing is a *"situated interactive activity"* that is *"constituted as a meaningful act through mutual contextualization"* [6]. Consequentially for making a computer understand natural human gestures, we would have to make it understand the users' context and share experiences with the user. This is, at best, a very distant vision.

GESTURE SETS

A simple solution to this problem is to define gesture sets that constrain which gestures users can use and a system must recognize. Gesture sets assign a meaning (or function) to each gesture regardless of the users and their context. By this, gestures become unambiguous but are also reduced to mere symbols within a context-free sign language that is far less expressive than natural human gesturing and whose gestures have to be learned first. Thus, with regard to the *"making computers more human"* hypothesis, even the best gesture sets are only as close to natural human gesturing as chatbots are to understanding the meaning of natural language input, passing the Turing test, or strong AI.

We can, however, accept gesture sets as something not necessarily "natural" but as an artificial language that, if properly designed, most users will be able to adopt. Applications such as Matulic & Norrie's impressive tabletop system for document editing with pen and touch gestures demonstrates the great potential of application-specific gesture sets [25]. In the best case, the interaction with such gesture sets achieves a yet unequalled feeling of flow, control, and directness. In a mediocre case, users only save a few seconds (assuming that a gesture is faster than selecting a menu item or recalling and entering keyboard shortcuts), users are not interrupted (assuming that there is no need to switch between mouse/touchpad/keyboard anymore), and UI designers can save screen estate by leaving out menus or other administrative controls. However, in the worst case, discovering functions becomes guesswork and learning and remembering gestures turns out to be as difficult as using the keyboard shortcuts or command line languages, in particular if they are not used frequently. Often it is simply not possible to design a clear, unmistakable, and easy-to-remember mapping between commands and gestures for all application scenarios. This also shows in the relatively small agreement rates when multi-touch and/or pen gestures for tabletops or multi-display environments are elicited from users [26, 27, 30, 34]. In my eyes, it is therefore highly unlikely that a single self-explanatory, easy-to-learn standard gesture set is possible at all, at least unless future UIs are redesigned to be mainly controlled by *manipulations* not gestures.

FROM GESTURES TO MANIPULATIONS

Manipulations are a class of gestural input that is apparently easier to learn and to agree on. For example, in [26], there was a *"a clear trend towards higher agreement scores on actions that could be performed through direct manipulation and lower agreement scores on actions that were symbolic in nature"*. This resonates with George & Blake [5] who differentiate between two classes of gestural input: *gestures* and *manipulations*. For them, gestures are *"symbolic interactions"* for *"discrete, indirect, intelligent interaction"* while manipulations are *"literal interactions"* for *"continuous, direct, environmental interaction"*. Examples for *gestures* are symbolic stroke gestures such as "✓", "✗" for accepting or rejecting an item, or Apple's three-finger-tap to do a lookup on the word under your cursor. *Manipulations*, however, are continuous actions in space, e.g., dragging or flicking of an object across the screen, pinch-to-zoom, or two-finger-rotate. In [13], I also argue for this dichotomy of gestures vs. manipulations and that using too many gestures could result in a pseudo-natural UI which is close to a command line interface. My argument was that successful gestural input (e.g. the very popular pinch-to-zoom) is actually mimicking how physical objects could be manipulated in the real-world and that this can be explained with Hutchins et al.'s classic cognitive account of *direct manipulation* with two major metaphors for the nature of human-computer interaction: the *conversation metaphor* vs. the *model-world metaphor* [11].

The *conversation metaphor* implies that the user interface of a computer system is a language medium and that interacting with a computer means that users can converse with the system and tell it what to do, ideally in a natural way. However, in many cases, a conversation about what should happen is inefficient compared to direct action or direct manipulation to make it happen. Just imagine how cumbersome it would be to drive a car from the backseat by having to tell the driver about every necessary action. This becomes even more cumbersome, if we first have to learn the vocabulary and language of the driver. The opposed *model-world metaphor* for the nature human-computer

interaction is not based on the idea of a conversation [11]: *"In a system built on the model-world metaphor, the interface is itself a world where the user can act, and which changes state in response to user actions. The world of interest is explicitly represented and there is no intermediary between user and world. Appropriate use of the model-world metaphor can create the sensation in the user of acting upon the objects of the task domain themselves".* Using this metaphor, Hutchins et al. explain how *direct manipulation* user interfaces (e.g. the GUI) replaced the command line by making better use of the users' cognitive resources and their perceptual, spatial, and motor skills and relying on physical action with immediate visual feedback instead of conversation. *Direct manipulation* reduces the cognitive distance (the *gulfs of execution* and *evaluation*) between the forms of user input and system output. Users get the feeling of directness from the commitment of fewer cognitive resources.

SPACES OF BLENDED INTERACTION

It is important to notice that *direct manipulation* and the *model-world metaphor* are by no means restricted to mouse-operated GUIs with real-world metaphors like the "desktop". Already in the 1980s, *direct manipulation* was realized with different input devices such as pens or game paddles and in many "non-real-world" application or game UIs [31]. In the coming era of ubiquity, the principle of *direct manipulation* will extend beyond the boundaries of a few desktop or mobile devices into our entire environment that increasingly will be augmented with touch, gesture, and motion detection, tangible user interface elements, flexible displays, ubiquitous projectors, and many other sorts of interactive or "smart" objects connected by the so-called "Internet of Things". The world itself becomes one large model-world user interface and it will enable gestures and manipulations not only across devices but also across the physical and digital realms, e.g., picking up the content of a physical note by touching it with a finger and then pasting its content on a tablet or smart phone with another touch.

In [17], I provide a novel and more accurate description of the nature of human-computer interaction in such spaces called *Blended Interaction* that is based on recent findings from embodied cognition and cognitive linguistics. It uses Fauconnier and Turner's conceptual blends and conceptual integration [4] to explain how users always rely on familiar and real-world concepts whenever they learn to use new digital technologies. Designers should consider using and blending the vast amount of concepts that we as humans share due to the similarities of our bodies, our early upbringing, and our sensorimotor experiences of the world before resorting to elaborate conscious analogies such as the desktop metaphor. Similar to Dourish's embodied interaction [3], *Blended Interaction* draws on the way the everyday world works or, perhaps more accurately, the ways we experience the everyday world, instead of drawing on seemingly familiar artifacts.

SPACE AND ACTION

A good starting point for *Blended Interaction* is our shared experience and awareness of space and our skills to act and navigate in it. However, according to embodied cognition, spatial cognition happens not only in our head but is inseparable from our actions, gestures, manipulations, or movements in space.

For example, Wesp et al. [33] found that, unlike the commonly held belief that the sole role of gestures is to communicate meaning, gestures also serve a cognitive function and help speakers to maintain spatial images in short-term memory. This raises the question if HCI can use gestural input to help users better memorize locations or states of objects. We therefore conducted a study of gestural vs. mouse input for panning a map-like UI. We found that the accuracy of memorized object locations in the map was significantly better when the map was panned with touch instead of mouse [14]. We assume that the proprioceptive feedback during touch navigation with a 1:1 control-display ratio supported the encoding of locations in spatial memory. This effect cannot be observed without a 1:1 ratio, e.g., when using a mouse or when users use zooming in addition to panning. In a further study, we also investigated body movements for peephole map navigation [29]. We found that users are able to physically navigate a large map (292x82cm) with just a small tablet-sized peephole (23.5x13.2cm) with almost the same efficiency as when seeing the entire map. We attribute this to the proprioceptive feedback of peephole navigation and the absolute mapping between physical and map space. However, we found no significant difference in spatial memory performance [28]. This demonstrates both the potential but also the difficulty of understanding the interplay between physical action, space, and cognition.

CONCLUSION

In [19], David Kirsh describes the great potential of understanding the complex interplay between cognition, space, and physical action for interaction design. I fully agree with his vision of *"cognitively informed designers"*. I believe that gestural input will greatly improve how we interact with future interactive systems, provided that a new generation of *"cognitively informed designers"* fully acknowledges the benefits of manipulations vs. gestures, does not force users to interact with systems using artificial sign languages, and designs for users' spatial abilities.

REFERENCES

1. Chen, N., Guimbretiere, F. and Sellen, A. Designing a multi-slate reading environment to support active reading activities. *ToCHI 19*, 3 (2012), 1-35.
2. Courtois, C., Schuurman, D. and Marez, L. D. Triple screen viewing practices: diversification or compartmentalization? *In Proc. EuroITV*, ACM (2011).
3. Dourish, P. *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press, 2004.

4. Fauconnier, G. and Turner, M. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2002.
5. George, R. and Blake, J. Object, Containes, Gestures, and Manipulations: Universal Foundational Metaphors of Natural User Interfaces. *In Natural User Interfaces (a CHI 2010 Workshop)*, Atlanta, USA, 2010.
6. Goodwin, C. Pointing as Situated Practice. *In Kita, S. (ed.) Pointing: Where Language, Culture, and Cognition Meet*, Lawrence Erlbaum (2003), 217-243.
7. Greenberg, S., Marquardt, N. and Ballendat, T., Diaz-Marino, R., Wang, M. Proxemic interactions: the new ubicomp? *interactions 18*, 1 (2011), 42-50.
8. Harper, R., Rodden, T., Rogers, Y. and Sellen, A. *Being human: Human-computer interaction in the year 2020*. Microsoft Research Ltd, Cambridge, England, 2008.
9. Hinckley, K. Synchronous gestures for multiple persons and computers. In Proc. UIST '03. ACM (2003).
10. Huang, D.-Y., Lin, C.-P., et al., MagMobile: enhancing social interactions with rapid view-stitching games of mobile devices. *Proc MUM '12*, ACM (2012).
11. Hutchins, E. L., Hollan, J. D. and Norman, D. A. Direct manipulation interfaces. *Human–Computer Interaction 1*, (1985), 311-338.
12. Jetter, H.-C. Design and Implementation of Post-WIMP Interactive Spaces with the ZOIL Paradigm. PhD Thesis, University of Konstanz, 2013.
13. Jetter, H.-C., Gerken, J. and Reiterer, H. Natural User Interfaces: Why We Need Better Model-Worlds, Not Better Gestures. *In Natural User Interfaces (a CHI 2010 Workshop)*, Atlanta, USA, 2010.
14. Jetter, H.-C., Leifert, S., Gerken, J., Schubert, S. and Reiterer, H. Does (Multi-)Touch Aid Users' Spatial Memory and Navigation in 'Panning' and in 'Zooming & Panning' UIs? *In Proc. AVI '12*. ACM (2012), 83-90.
15. Jetter, H.-C., and Rädle, R. Visual and Functional Adaptation in Ad-hoc Communities of Devices. *In Visual Adaptation of Interfaces (Workshop ITS '13)*, St. Andrews, Scotland, 2013.
16. Jetter, H.-C. and Reiterer, H. Self-Organizing User Interfaces: Envisioning the Future of Ubicomp UIs. *In Blended Interaction (a CHI 2013 Workshop,)* Paris, France, 2013).
17. Jetter, H.-C., Reiterer, H. and Geyer, F. Blended Interaction: understanding natural human–computer interaction in post-WIMP interactive spaces. *Pers. and Ubiq. Comp.*, DOI 10.1007/s00779-013-0725-42013).
18. Jetter, H.-C., Zöllner, M., Gerken, J. and Reiterer, H. Design and Implementation of Post-WIMP Distributed User Interfaces with ZOIL. *International Journal of Human-Computer Interaction, 28*, 11 (2012), 737-747.
19. Kirsh, D. Embodied cognition and the magical future of interaction design. *ToCHI, 20*, 1 (2013), 1-30.
20. Li, M. and Kobbelt, L. Dynamic tiling display: building an interactive display surface using multiple mobile devices. *In Proc MUM '12*. ACM (2012).
21. Lucero, A., Holopainen, J. and Jokela, T. Pass-them-around: collaborative use of mobile phones for photo sharing. *In Proc CHI '11*, ACM (2011), 1787-1796.
22. Lucero, A., Jones, M., Jokela, T. and Robinson, S. Mobile collocated interactions: taking an offline break together. *interactions, 20*, 2 (2013), 26-32.
23. Marquardt, N., Diaz-Marino, R., Boring, S. and Greenberg, S. The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. *In Proc UIST '11*. ACM (2011), 315-326.
24. Marquardt, N., Hinckley, K. and Greenberg, S. Cross-device interaction via micro-mobility and f-formations. *In Proc UIST '12*. ACM (2012), 13-22.
25. Matulic, F. and Norrie, M. C. Pen and touch gestural environment for document editing on interactive tabletops. *In Proc ITS '13*. ACM (2013), 41-50.
26. Mauney, D., Howarth, J., Wirtanen, A. and Capra, M. Cultural similarities and differences in user-defined gestures for touchscreen user interfaces. *In CHI EA '10*. ACM (2010), 4015-4020.
27. Ohta, T. and Tanaka, J. Pinch: An Interface That Relates Applications on Multiple Touch-Screen by 'Pinching' Gesture. *In Proc ACE '12*. Springer (2012), 320-335.
28. Rädle, R., Jetter, H.-C., Butscher, S. and Reiterer, H. The Effect of Egocentric Body Movements on Users' Navigation Performance and Spatial Memory in ZUIs. *In Proc ITS '13*. ACM (2013), 23-32.
29. Rädle, R., Jetter, H.-C., Müller, J. and Reiterer, H. Bigger is not always better: Display Size, Performance, and Task Load during Peephole Map Navigation. *to appear in Proc CHI '14*. ACM (2014).
30. Seyed, T., Burns, C., Sousa, M. C., Maurer, F. and Tang, A. Eliciting usable gestures for multi-display environments. In Proc ITS '12. ACM (2012), 41-50.
31. Shneiderman, B. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology, 1* (1982), 237-256.
32. Tomasello, M. *Origins of Human Communication.* MIT Press, Cambridge, MA, USA, 2008.
33. Wesp, R., Hesse, J., Keutmann, D. and Wheaton, K. Gestures Maintain Spatial Imagery. The *American Journal of Psychology, 114*, 4 (2001), 591-600.
34. Wobbrock, J. O., Morris, M. R. and Wilson, A. D. User-defined gestures for surface computing. *In Proc. CHI' 09*. ACM (2009), 1083-1092.