DISTANT POINTING IN DESKTOP COLLABORATIVE VIRTUAL ENVIRONMENTS

A Thesis Submitted to the College of
Graduate Studies and Research
In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
In the Department of Computer Science
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

By

NELSON WONG

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science

University of Saskatchewan

Saskatoon, Saskatchewan S7N 5C9

ABSTRACT

Deictic pointing—pointing at things during conversations—is natural and ubiquitous in human communication. Deictic pointing is important in the real world; it is also important in collaborative virtual environments (CVEs) because CVEs are 3D virtual environments that resemble the real world. CVEs connect people from different locations, allowing them to communicate and collaborate remotely. However, the interaction and communication capabilities of CVEs are not as good as those in the real world. In CVEs, people interact with each other using avatars (the visual representations of users). One problem of avatars is that they are not expressive enough when compare to what we can do in the real world. In particular, deictic pointing has many limitations and is not well supported.

This dissertation focuses on improving the expressiveness of distant pointing—where referents are out of reach—in desktop CVEs. This is done by developing a framework that guides the design and development of pointing techniques; by identifying important aspects of distant pointing through observation of how people point at distant referents in the real world; by designing, implementing, and evaluating distant-pointing techniques; and by providing a set of guidelines for the design of distant pointing in desktop CVEs.

The evaluations of distant-pointing techniques examine whether pointing without extra visual effects (natural pointing) has sufficient accuracy; whether people can control free arm movement (free pointing) along with other avatar actions; and whether free and natural pointing are useful and valuable in desktop CVEs.

Overall, this research provides better support for deictic pointing in CVEs by improving the expressiveness of distant pointing. With better pointing support, gestural communication can be more effective and can ultimately enhance the primary function of CVEs—supporting distributed collaboration.

PUBLICATIONS FROM THIS DISSERTATION

Wong, N., & Gutwin, C. (in submission). Support for Deictic Pointing in CVEs: Still Fragmented After All These Years? Submitted to *European Conference on Computer-Supported Cooperative Work (ECSCW 2013)*. (Chapter 7)

Wong, N., & Gutwin, C. (2012). Controlling an Avatar's Pointing Gestures in Desktop Collaborative Virtual Environments. In *Proceedings of the International Conference on Supporting Group Work, GROUP 2012* (pp. 21–30). (Chapter 6)

Wong, N., & Gutwin, C. (2010). Where Are You Pointing? The Accuracy of Deictic Pointing in CVEs. In *Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI 2010* (pp. 1029–1038). (Chapters 4, 5)

ACKNOWLEDGEMENTS

I would like to express my greatest gratitude to everyone who has helped and supported me throughout my research.

To Carl Gutwin: thank you, thank you, thank you. You are an awesome supervisor. Thank you for guiding me through my long PhD journey with your wisdom, intelligence, and patience. I will never forget your five questions of research, how you helped me by brilliantly turning answers into questions, and the night you patiently taught me how to use |STAT with step-by-step instructions via a phone call. This dissertation would not have been possible without your mentorship.

To Regan Mandryk, Kevin Stanley, Julita Vassileva, Veronika Makarova, and Gerald Penn, my committee members: thank you for your interest and encouragement. I could not ask for a better committee.

To Sheelagh Carpendale: thank you for giving me the courage to pursue my PhD degree. Without you saying “you can do it”, I would not have even applied for a PhD. Also, thank you for helping me open the door to research.

To Judy Gartaganis and Paul Pospisil: thank you for being great teachers. You taught me the basics of computer science and helped me build a strong foundation to be a computer scientist.

To all my colleagues and friends of the Interaction Lab: thank you for your support and companionship. In particular, thanks to Brian de Alwis, Mike Lippold, and Stephen Damm for helping me solve programming problems; David Flatla for sharing an office and countless research ideas with me; André Doucette, Scott Bateman, Adrian Reetz, Kathrin Gerling, and Steve Sutcliffe for reading and commenting on my drafts, and listening and critiquing my practice talks; and Max Birk and Gregor McEwan for being my coffee buddies.

To Tony Lai, Johnny Saw, Thomas Tang, and Ben Tou: thank you for keeping me company whenever I came back to Calgary and for sharing the same passion with me for all these years. You guys rock.

To Foon Lu: thank you for helping me settle in Saskatoon and being a wonderful friend.

To Yue “Ha-Ha” Gao: thank you for sharing my laughter and sorrow with me, supporting me when I needed it the most, and being there for me. I will always remember how you took care of me after I injured my hand.

To Dai-Ma, Dai-Ba, and Ga-Jeh: thank you for your love and support for all these years.

To Mom, Dad, and Emily: thank you for everything. Mom and Dad, thank you for giving me the freedom to pursue my dream. Emily, thank you for your smiles and being a supportive sister.

To mom and dad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Deixis is a reference to a thing that is relevant to the context of an utterance. *Deictic pointing* is a pointing gesture that provides such a reference and is usually used with words like "this" or "that": for example, "look at that tree <points>." It is ubiquitous, natural, and simple in face-to-face communication. Although pointing has some cultural dependencies, many people from many different backgrounds naturally use deictic pointing in daily activities, such as showing people directions and indicating where things are.

Deictic pointing is useful and important. It enhances our communication by providing a non-verbal channel that simplifies and clarifies verbal descriptions. For example, "let's go over to the blue coffee shop at the corner across the street" can be simplified to "let's go over there <point>." In some situations, only using speech to communicate can be confusing, e.g., in Figure 1.1A, Alex is facing Bob and says "look at the monitor behind you." It is unclear to which monitor Alex is referring. The uncertainty exists even if Alex specifies the location: "look at the monitor behind you. The right one." It can mean the one on Alex's right, but it can also mean the one on Bob's right. Deictic pointing can clarify the confusion and simplify the utterance: "look at that monitor <point>." (Figure 1.1B)

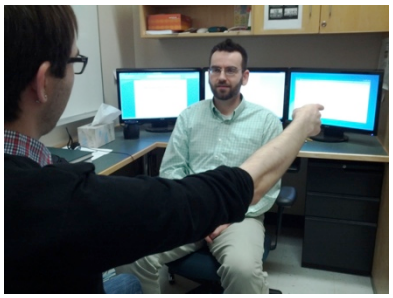
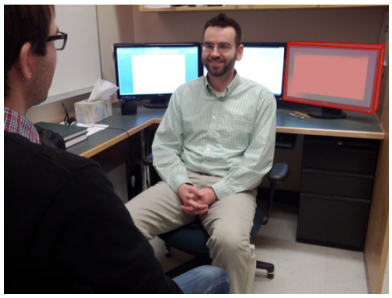
A  B  C 

Figure 1.1: A) Alex mentions a monitor behind Bob; B) Alex points at a monitor behind Bob; C) a monitor behind Bob is highlighted, but he cannot see it.

Deictic pointing is important in the real world; it is also important in collaborative virtual environments (CVEs) because CVEs are three-dimensional virtual worlds that resemble the real

world. Some of them have a higher degree of resemblance to the real world, e.g., Second Life (a social community); others have a lower degree, e.g., World of Warcraft (a fantasy-world game). CVEs have been moving from research labs into everyday settings and are becoming increasingly common. A main reason why CVEs have become popular is their ability to connect people from different locations, allowing them to communicate and collaborate remotely. While CVEs resemble the real world, however, the interaction and communication capabilities of CVEs are not nearly equal to those of the real world.

In current CVEs, people mainly rely on text channels and voice chat to communicate. People should also be able to use pointing gestures through their avatars (a human-like representation of themselves in the CVE) to communicate and interact with each other as naturally and freely as they can in the real world, but deictic pointing in CVEs has many limitations and is not well supported.

Deictic pointing in CVEs has limited expressiveness, is difficult to control, and often depends heavily on visual aids, such as object highlighting and 'laser beams,' which are not complete solutions to the problem of referencing in CVEs. For instance, pointing with 'laser beams' can cause confusion when multiple beams appear on the screen. Also, object highlighting would not help indicate an object if the addressee cannot see it. For example, if the monitor example above were situated in a CVE, Bob would not know which monitor Alex is referring to with only object highlighting because it is behind Bob (Figure 1.1C). The current pointing mechanisms in CVEs are not sufficient to support the rich information that is available with pointing gestures.

In order to keep this research effort manageable, I focus on distant pointing in desktop CVEs. *Distant pointing* is a type of deictic pointing where the target is out of reach. A *desktop CVE* is set up in a desktop environment. Figure 1.2 shows a typical desktop CVE that uses a desktop monitor as the display, and a keyboard and a mouse as input devices.

Figure 1.2: A desktop CVE.

1.1 Problem Statement

The problem addressed by this research is that *pointing in CVEs is limited in comparison with pointing in the real world*. The limitations are three-fold. First, the generation of pointing gestures in CVEs is discrete while real-world pointing gestures are created with continuous control. The gradual creation of pointing gestures provides important information for referential conduct. In CVEs, however, pointing is often immediate with discrete movement (e.g., with a "/point" command), resulting in missing crucial information, such as the speed of the gestures that may imply the importance of the conduct (e.g., a fast pointing gesture implies urgency and requires immediate attention). Second, pointing in CVEs is more difficult to generate than that in the real world. Raising an arm and index finger to generate a pointing gesture is generally easy to do in real life; however, doing so in CVEs with an avatar is much more complex. For example, controlling pointing in CVEs requires manipulating at least two extra degrees of freedom with the arms and hands, which are already busy with motion and direction controls. Third, pointing in CVEs is harder to observe than in the real world. This is primarily due to the difference in field of view width. The field of view in CVEs, especially desktop CVEs, is much smaller than that in the real world, making it more difficult to see pointing gestures generated by collaborators. Compared to observing a real person pointing, factors such as low avatar resolution and small screen size also reduce the visibility of pointing gestures.

Based on the research problem, the statement of thesis is that *the limitations of pointing in CVEs can be alleviated by three pointing methods: free, natural, and augmented*. *Free pointing* is pointing that is independent from other avatar actions such as moving or turning, *natural pointing* is pointing that does not have any visual effect other than the movement of the gesture,

and *augmented pointing* is pointing that has additional visual effects (see Chapter 3 for more details on free, natural, and augmented pointing).

1.2 Motivation

As more people use CVEs for work (e.g., Second Life), socialization (e.g., PlayStation Home), and entertainment (e.g., World of Warcraft), support for deictic pointing in CVEs becomes more important. When there is no pointing support, people need to rely mainly on voice and text for referencing. Even when avatars are able to point, problems still arise if the support is poor. For example, avatars cannot point at spaces between objects, pointing animations are fixed and predefined without variations, subtle pointing gestures cannot be generated, gestures cannot be paused once the command is executed, pointing gestures are difficult to synchronize with speech, and pointing commands are hard to remember. Given the ubiquity of referential communication, these limitations mean that much effort would be wasted and errors would be introduced during communication in CVEs. Therefore, it is important to provide proper and sufficient support for deictic pointing in CVEs.

1.3 Solution

My solution to the research problem is to design techniques and provide design guidelines for improving the expressiveness of pointing gestures in CVEs. I do that with free, natural, and augmented pointing (see Chapter 3 for more details). This solution has the following steps:

1. ***Develop a framework of distant pointing***. The framework identifies important stages and characteristics of distant pointing (from analysis of previous work).
2. ***Identify important aspects, such as accuracy requirements, of distant pointing***. I observe and analyse how people point at distant objects in the real world.
3. ***Develop pointing techniques***. Based on the insights gained from the observational study, I design and develop new pointing techniques for CVEs.
4. ***Develop a CVE that supports free and natural pointing, and other pointing techniques***. This CVE is the basic software for the rest of my studies (with slight modifications for each).

5. ***Examine if natural pointing has sufficient accuracy to be used in desktop CVEs***. The pointing technique needs to have adequate accuracy in CVEs; otherwise, it would not be useful. I compare pointing accuracy in the real world and CVE.
6. ***Provide ways for people to control free pointing along with other avatar actions***. I compare different input devices and configurations to find out the best input for controlling free pointing together with other avatar actions such as moving, turning, and looking.
7. ***Determine whether people can control free pointing together with other avatar actions***. The pointing method would not be useful if it cannot be controlled along with actions like moving and turning. I conduct a study to determine if people can control free pointing simultaneously with other avatar actions.
8. ***Verify the usefulness of free and natural pointing in desktop CVEs***. I observe how people use free, natural, and augmented pointing in a CVE with realistic collaborative tasks to compare the techniques, and to determine if free and natural pointing are useful even when augmented-pointing techniques are available.
9. ***Provide design guidelines for distant pointing in CVEs***. Based on the studies and the lessons learned, I develop a set of design guidelines to help CVE designers provide distant pointing support.

1.4 Contributions

The main contribution of this dissertation is *the design and evaluation of a set of techniques for improving the expressiveness of pointing gestures in CVEs which together demonstrate that pointing-based referential communication can successfully be integrated into CVEs*. My research also has the following minor contributions:

- *Provide a framework of distant pointing in desktop CVEs.*
- *Identify important aspects of distant pointing from observational work.* These important aspects include pointing accuracy requirements, types of pointing gestures, communication richness, attentional focus, and observer's location.

- *Show that natural pointing is accurate enough to be used in desktop CVEs.* In particular, people can interpret others' pointing gestures in a CVE almost as well as they can in the real world.
- *Determine the best way to control free pointing among five sets of commonly-available input devices.* The mouse and Wii controls are consistently better than the trackball, gamepad, and joystick.
- *Verify that free pointing can be controlled together with other actions of an avatar.*
- *Show that free and natural pointing are useful for distributed collaboration.* They are useful even when augmented pointing techniques (e.g., a laser beam) are available and are particularly important at early stages of pointing.
- *Provide a set of design guidelines for developing distant pointing in CVEs.*

1.5 Overview of the Dissertation

Chapter 2 provides background on three main areas of this research—deictic pointing, CVEs, and deictic references in CVEs. Chapter 3 details the framework of distant pointing. Chapters 4, 5, 6, and 7 provide details of the steps to the solution. Chapter 4 describes an observational study about how people point at distant targets in the real world in order to identify important aspects of distant pointing. Chapter 5 focuses on pointing accuracy. It describes a study comparing distant pointing accuracy in the real world and in a CVE, and shows that distant pointing is accurate enough to be used in CVEs. Chapter 6 is about controlling distant pointing. It details a study that compares different pointing control techniques and tests how well distant pointing can be used together with other avatar actions such as moving, turning, and looking. Chapter 7 verifies the usefulness of distant pointing for distributed collaboration in CVEs. It describes an observational study that shows how distant pointing is used to perform collaborative tasks in a CVE. Chapter 8 lists the design guidelines for distant pointing in CVEs, and discusses the solution to the research problem, the importance of free, natural, and augmented pointing techniques, and the limitations and generalizability of the research. Chapter 9 concludes the dissertation by summarizing the contributions and providing future directions for the research.

CHAPTER 2
BACKGROUND

This chapter provides fundamental understanding of three main areas that help improve the support of pointing in CVEs: deictic pointing, CVEs, and deictic references in CVEs.

2.1 Deictic Pointing

Deixis is a term that comes from a Greek word meaning display, indicate, and reference. Lyons (1977) wrote:

> By deixis is meant the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatiotemporal context created and sustained by the act of utterance and the participation in it, typically, of a single speaker and at least one addressee. (Lyons, 1977, p. 637)

There are three common kinds of deixis: person, time, and place (Lyons, 1977). For all of them, the context of the utterance is important to identify the *referent*—the actual thing being referred to. For example, in the sentence "she is happy", the word "she" can refer to different people depending on who utters the sentence (person-deixis); in "today is sunny", "today" refers to the day on which the sentence is said (temporal deixis); and in "that is a nice car", "that" refers to the car at a particular location (spatial deixis).

Spatial deixis usually requires indices to help identify the referents. An *index* is one of Peirce's classes of signs (Buchler, 1955) that has a physical connection to the object of interest. Clark (1996) presented the concept of signals and methods of signaling. A *signal* is "the presentation of a sign by one person to mean something for another." (Clark, 1996, p. 160) The method of signaling for an index is *indicating*.

Clark (1996) identified many ways of indicating. For example, when someone rings a doorbell, the sound of the doorbell indicates that a person is outside the house. People also use voices in

conversations to indicate things, e.g., saying "I" indicates identity, "now" indicates time, and "here" indicates location. More obvious ways of indicating are by using body parts, such as using the body to occupy something (sitting in a chair and saying "I'm sitting here"), using the eyes to gaze at someone ("I need to talk to you <look> and you <look>"), and using the head to nod at somewhere ("let's go over there <nod at the direction>"). Among all indicating methods, using a finger to point at something—*deictic pointing*—is the most common (e.g., "that is my book <point at the book>").

2.1.1 Pointing and Gestures

Pointing is a type of hand gesture (Krauss, Chen, & Gottesman, 2000; McNeill, 1992). Hand gestures can be ordered based on linguistic properties, and this ordering is called Kendon's continuum (1988): gesticulation, language-like gestures, pantomimes, emblems, and sign languages.

From gesticulation to sign languages, the obligatory presence of speech decreases while the presence of language properties increases (Kendon, 1988; McNeill, 1992). *Gesticulation* is idiosyncratic spontaneous hand and arm movements during speech. An example is pushing the palm forward while saying "he pushes the door open". *Language-like gestures* are similar to gesticulation, but grammatically integrated in the utterance. For example, "the car fell off the bridge and <gesture to show overturning>". The gesture here is replacing part of the speech. *Pantomimes* are gestures without speech. Multiple successive pantomimes can produce complex and sequential demonstration. *Emblems* are well-formed gestures that need to be performed in some specific ways, such as the OK sign that needs to be made by connecting the thumb and index finger while extending all other fingers. A *sign language* is a set of gestures that has a full linguistic system.

In addition to Kendon's continuum (1988), gestures can also be categorized using McNeill's classification (1992). The classification has five types of gestures: iconics, metaphorics, beats, deictics, and cohesives.

Iconic gestures depict the appearances of objects or actions of events, and have a close formal relationship to the semantic content of speech, e.g., moving both hands up and down while

saying “she climbed up that ladder”. *Metaphoric* gestures are pictorial gestures that show abstract ideas, e.g., when describing different game genres, a gesturer raises the hands to create sections of area as if each genre belongs to a section. *Beats* are hand movements that move along with the rhythm of the speech and have only two movement phases, e.g. in and out, up and down. They are used to emphasize the significance of words or phrases. *Deictics* are pointing gestures for indicating objects or events either concrete or abstract, e.g., saying “look at the tree” while pointing at the tree. *Cohesive* gestures are used to link together temporarily separated parts of a discourse that are within the same theme; they can be iconics, metaphorics, beats, or deictics.

While deictic pointing can be performed by “the furthest point of the body part that has been extended outward into the environment” (Kendon & Versante, 2003, p. 112) and can be done using foot, head, lip, etc. (Sherzer, 1973), my research focuses on deictic pointing as a type of hand gestures.

2.1.2 Phases of Pointing

When we generate a pointing gesture, our arms and hands go through a sequence of actions. Kendon (1980) classified the sequence in terms of gesticular unit, gesticular phrase, and gesticular phase.

A *gesticular unit* (G-Unit) starts when a limb moves from a *rest position* (e.g., one’s lap or the arm rest of a chair) and ends when the limb moves back to another rest position. Each G-Unit contains one or more *gesticular phrases* (G-Phrases), and each G-Phrase is composed by different phases: preparation, stroke, and recovery. A *preparation* is the movement of a limb from a rest position to the beginning of a stroke. A *stroke* is an accented movement with a distinct peak of effort in the sense of dance movements (Dell, 1977). A *recovery* describes the limb moving back to a rest position or becomes ready for another stroke. A stroke is an obligatory phase in a G-Phrase, whereas preparation and recovery are optional.

McNeill (1992) and Kita (1990) point out that there are holding phases in a G-Phrase: pre-stroke hold and post-stroke hold. A pre-stroke hold is a period where the gesture waits for the speech to occur, and a post-stroke hold is to extend the period of a stroke.

In 1998, based on Kendon's phases of gestures, Kita et al. (1998) proposed a more complete classification of movements. He pointed out that the stroke phase can be omitted in a G-Phrase. To capture this concept, he introduced the idea of an *expressive phase* that contains either a stroke or a stroke-less hold (also called independent hold). Kita also broke the preparation phase down into three different phases. First, the *liberating movement* phase is for freeing the hands from some constrained location, e.g., from an interlocking finger position. Second, the *location preparation* phase moves the arm to the starting position of an expressive phase. Third, the *hand internal preparation* phase is for shaping and orienting the hand for an expressive phase. The two preparation phases can occur at the same time. Figure 2.1 shows the flow of different movement phases.

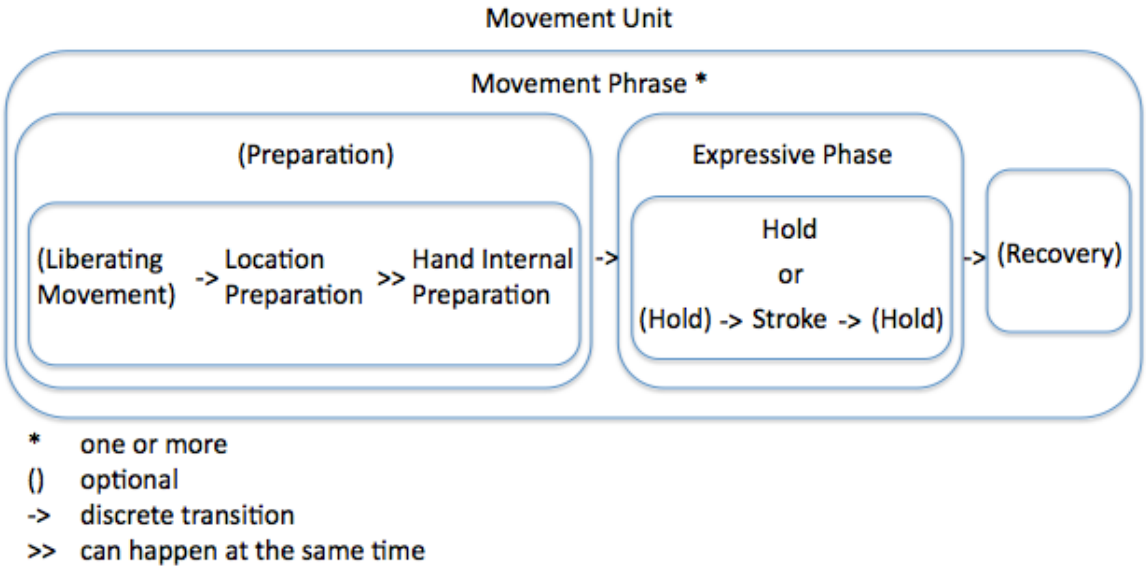


Figure 2.1: Movement phases (modified based on (Kita et al., 1998, pp. 26–27)).

2.1.3 Factors Affecting Pointing

The success of pointing gestures depends on more than how the gestures are generated. There are other relating interaction factors—such as speech, gaze, and orientation—that affect how well people can understand the pointing gestures (Goodwin, 1981; Heath, 1986; Hindmarsh & Heath, 2000). How people communicate is based on the situation at hand (Clark, 2003; Goodwin & Goodwin, 1996; Goodwin, 2000; Heath & Luff, 1991; Heath, 1986; Hindmarsh & Heath, 2000). Researchers found that speech, gaze, orientation, and pointing gestures are mutually related in

many communication settings, for example, a playground (Goodwin, 2000), a medical consultation room (Heath, 1986), the line control rooms of the London Underground (Heath & Luff, 1991), the restoration control office of British Telecom (Hindmarsh & Heath, 2000), an airport (Goodwin & Goodwin, 1996), and a university (Kita, 2003).

When observing children playing hopscotch, Goodwin (2000) found that human interaction using speech and gestures was tightly related to the environment where the interaction took place. For example, when the children were discussing an illegal move in a game, the interactions (including their gaze direction, body orientation, gestures, and speech) between them only made sense when they were standing inside the hopscotch grid. Similar observations can also be found in other reports, e.g., (Goodwin & Goodwin, 1996; Heath & Luff, 1991; Heath, 1986; Hindmarsh & Heath, 2000).

Heath (1986) observed how patients and doctors interacted during medical consultations. He found that when a patient looked at the doctor, the patient enabled a "display of recipiency" which invited the doctor to start a conversation. This highlights Goodwin's findings in the observation of human discourse. Goodwin (1981) found that gaze between speakers and hearers is the fundamental element that initiates conversations. Whether and when a speaker obtains a hearer's gaze can change the dynamic and even content of their conversation. For example, the speaker may wait until receiving the hearer's gaze to begin speaking, and the speaker may pause or restart a sentence to request the hearer's gaze.

While gaze often initiates conversations, mutual orientation is also an important element supporting human interactions. Goodwin (1981) found that once the speaker and hearer obtain mutual orientation, they can withdraw their gaze without much impact to the conversation, by maintaining their awareness through body motion (e.g., nodding the head) and vocal responses (e.g., "mmhm"). The importance of mutual orientation can also be seen in other research projects. For example, Heath and Luff (1991) found that the workers in the line control room of the London Underground relied on mutual orientation to accomplish their tasks. When having mutual orientation, workers can maintain awareness of each other's actions; thus, they can obtain enough information to work on their individual tasks without interrupting each other. Furthermore, in the restoration control office of British Telecom, Hindmarsh and Heath (2000)

found not only that having mutual orientation is crucial in communication, but also that being able to refer to objects by pointing is also important. The workers can correctly identify objects in complex collaborative settings because they can point at the object while maintaining mutual orientation.

In studies done by Kita (2003) outside a university library in Tokyo, he underscored the importance of the interplay between orientation, gaze, speech, and pointing gesture. He asked people outside the university library to describe directions to places that are out-of-view (occluded or too far to see) and to point at visible targets without speech. He found that all those components of communication are interrelated and crucial in communication.

2.1.4 Pointing in Video-Mediated Environments

While deictic pointing is commonly used in face-to-face activities, it can also be used in distributed settings, such as CVEs and video-mediated environments. In video-mediated environments, videos of hand gestures (Kirk, Rodden, & Fraser, 2007; Tang & Minneman, 1991b; Tang, Neustaedter, & Greenberg, 2006) and full-scale people (Heath, Luff, Kuzuoka, Yamazaki, & Oyama, 2001; Luff, Yamashita, Kuzuoka, & Heath, 2011; Tang & Minneman, 1991a) can be projected into collaborators' workstations to enhance distributed collaboration.

Tang and Minneman (1991b) developed a shared drawing tool, called VideoDraw, that allows two people to collaborate remotely. VideoDraw uses a video camera to capture hand gestures and drawings, and shows the video on the partner's workstation. Being able to see each other's gestures (e.g., point at certain parts of the drawings) while working together remotely provides a new sense of co-presence between collaborators.

Tang and Minneman (1991a) later developed another shared drawing tool called VideoWhiteboard. The main difference between the two tools is that instead of showing hand gestures in full-colour video, VideoWhiteboard shows the shadows of collaborators' upper bodies. Video cameras are placed at the back of the projection screens to capture users' shadows. Tang and Minneman found that pointing gestures that require precise locations can be difficult to perceive when users stand far away from the screen because the shadows are blurry. However,

users are willing to exaggerate gestures and stay close to the screen to compensate for the difficulty, showing the importance of pointing gestures.

Kirt et al. (2007) observed collaborators (one worker and one helper) performing a Lego block assembly task using a system similar to VideoDraw. They found that pointing gestures help establish mutual references and reduce overlapped speech, thus smoothing turn-taking and improving task performance.

Studies using other video-mediated systems also show the importance and frequent use of pointing gestures. For example, Tang et al. (2006) used VideoArms, images of local collaborators' arms that are redrawn at a remote location, to share hand gestures. They found that gestures are often used as a substitute for speech. Luff et al. (2011) observed how users worked on collaborative tasks in a system, called t-Room. A t-Room has multiple displays, video cameras, and tables. It allows full-scale videos of remote collaborators to be shown on local displays. While pointing gestures were used throughout the study, imperfect camera locations and viewer perspectives hindered the effectiveness of pointing. Heath et al. (2001) observed how two distributed collaborators arrange furniture. In the study, a collaborator was physically in a living room with a robot that was controlled by a remote partner. The remote partner could see what happened in the room through video cameras on the robot. Many pointing gestures were used during the study. However, it was difficult for the collaborators to fully understand each other and to know where referents were because they did not know whether their own conduct could be seen by their partner. Heath et al. (2001) also found that pointing often became the centre of discussion instead of a resource for collaboration. These studies show that while deictic pointing is useful and important, the setting and environment where pointing is used are crucial to its effectiveness.

2.1.5 Pointing in Different Cultures

People generate and use pointing gestures differently due to different cultural influences. In North America, pointing at objects with index finger extended is common, and is only considered impolite when pointing at people. In some Asian countries, such as Japan, Malaysia, and Indonesia, index-finger pointing in general is considered rude (Etiquette in Asia, 2012). Pointing gestures also have some specific rules in some cultures. For example, pointing with the

left hand is a taboo for Ghanaians (Kita & Essegbey, 2001), and pointing gestures with different hand shapes have different meanings for Arrernte people (Wilkins, 2003). In this section, I show how different cultures influence pointing gestures.

Much research has shown that cultures have influences on pointing gestures even at early stages of the development of human communication. Salomo and Liszkowski (2012) observed daily activities of 48 infants (from 8 to 15 months) and their interlocutors across Yucatec-Mayans (Mexico), Dutch (Netherlands), and Shanghai-Chinese (China) cultures. The observation was focused on index-finger pointing. The authors found that Chinese infants used the most number of gestures, followed by Dutch infants, and then by Mayan infants. Index-finger pointing was used more frequently than other gestures. Also, the number of infants pointing with their index finger was significantly different across the three groups (Chinese was the most and Mayans was the least). The results suggested that infants' prelinguistic pointing behavior is influenced by the amount of social-interactional experience (Chinese infants had the most interactions with their caregivers while Mayan infants had the least). Zlatev and Andrén (2009) observed how Swedish and Thai children use pointing gestures between 18 and 27 months of age. Index-finger pointing occurred more frequently with Swedish children than with Thai. The findings suggested that a norm to avoid index-finger pointing, especially directed at people, in Thailand has influences on how children point. Pettenati et al. (2012) observed 22 Italian and 22 Japanese toddlers (from 25 to 37 months) performed picture-naming tasks. They found that some differences of how gestures were used with speech between the two groups could be related to cultural differences. For example, Italian children used more gestures without speech because Italian adults often use gestures as emblems, and Japanese children created gestures that were closer to actions shown in pictures because learning by observation is more common in Japanese culture.

Some cultural norms and traditions influence how pointing gestures are generated. For example, index-finger pointing is a taboo in American Indian culture; thus pointing with the lip is a common way for referential communication (Kirch, 1979; Labarre, 1947). Also, pointing and gesturing with the left hand is considered provocative and disrespectful in Ghana (Kita & Essegbey, 2001). Kita and Essegbey observed how Ghanaians give route directions. They found that when Ghanaians point to the left, they often cross the right hand in front of the face or around the neck to avoid using the left hand. For another example, using different hand shapes

for pointing can have different specific meanings for Arrernte people (Wilkins, 2003). Pointing with a horned sign (index and little finger are extended, and middle and ring finger are contracted) is for showing the destination of motion (e.g., "go over there <point with a horned sign>"). Pointing with one finger extended is to indicate a single object. When referring to multiple objects or a region, Arrernte people would point with all fingers extended and spread out. When showing path segments and turns, one would use a flat hand with all fingers held together.

2.2 Collaborative Virtual Environments

Collaborative virtual environments (CVEs) can be seen as a combination of the work from the fields of Computer Supported Cooperative Work (CSCW) and Virtual Worlds (Benford, Greenhalgh, Rodden, & Pycock, 2001). Virtual worlds (or virtual environments) are computer generated three-dimensional environments that resemble the real world. Ellis (1994) defined virtual environments as "interactive, virtual image displays enhanced by special processing and by nonvisual display modalities, such as auditory and haptic, to convince users that they are immersed in a synthetic space." (Ellis, 1994, p. 17)

CSCW is a field that involves many different disciplines, such as Human-Computer Interaction, social sciences, and networking (Bannon & Schmidt, 1989; Grudin, 1994). While CSCW has a multidisciplinary nature, basically it can be described as "a field which covers anything to do with computer support for activities in which more than one person is involved." (Bannon & Schmidt, 1989, p. 359)

In traditional two-dimensional systems that support CSCW, users are normally represented by telepointers (Dyck, Gutwin, Subramanian, & Fedak, 2004; Greenberg, Gutwin, & Roseman, 1996), which are replicated pointers that track the locations of other users' mouse cursors. In three-dimensional virtual environments, users can be represented by more complicated embodiments called avatars (Benford, Bowers, Fahlén, Greenhalgh, & Snowdon, 1995; Fraser, Hindmarsh, Benford, & Heath, 2004; Salem & Earle, 2000; Wadley & Ducheneaut, 2009). Avatars are three-dimensional representations of the users and are commonly shown as human-like shapes. Users can use their avatars to interact with others and objects in the virtual world.

When adding CSCW properties to virtual environments, the environments become CVEs. CVEs are computer generated three-dimensional worlds that resemble the real world, and allow people to interact with one another and objects in the environment via their avatars.

2.2.1 Types of Collaborative Virtual Environments

There are two main types of CVEs—immersive and non-immersive. Biocca and Delaney (1995) defined *immersive* as:

> the degree to which a virtual environment submerges the perceptual systems of the user in computer-generated stimuli. The more the system capivates the senses and blocks out stimuli from the physical world, the more the system is considered immersive. (Biocca & Delaney, 1995, p. 57)

In general, immersive CVEs give users a feeling of *being in* a virtual environment; whereas, non-immersive CVEs support the feeling of *looking at* the environment (Kjeldskov, 2001; Shneiderman, 1998). The sense of *being in* and *looking at* is largely dependent on the devices being used (Bowman, Datey, Ryu, Farooq, & Vasnaik, 2002; Draper, Kaber, & Usher, 1998; Otto, Roberts, & Wolff, 2005, 2006; Pausch, Proffitt, & Williams, 1997; Slater & Steed, 2000). An immersive CVE usually uses motion-tracking devices along with a head-mounted display (HMD) or a spatially immersive display (SID), and a non-immersive CVE primarily uses a desktop display with a mouse and a keyboard.

A fully immersive head-mounted CVE requires the user to wear a HMD. A typical HMD is a helmet-like device with a display in front of each eye (e.g., Figure 2.2). Some HMDs use the same image for both eyes while others use different images to produce a stereoscopic effect. HMDs are also equipped with a tracker to detect head movements. By tracking head location and orientation, the scene shown on the displays can be changed accordingly to provide a correct view for the user.

Figure 2.2: A headed-mounted display.

A spatially immersed CVE uses a SID that surrounds the user. A typical SID has four to six sides arranged like a cube (Figure 2.3 shows a four-sided SID). Images are projected from outside the cube. The user needs to wear a head tracker so that the virtual environment can be changed corresponding to the user's point of view, and normally wears stereoscopic glasses for 3D effects. The CAVE (Cruz-Neira, Sandin, & DeFanti, 1993) developed at the Electronic Visualization Laboratory at the University of Illinois is a commonly used SID.

Figure 2.3: A four-sided spatially immersive display.

A desktop CVE only uses a desktop monitor. It can be used in regular home and office settings, and is the most common kind of CVE (see Figure 1.2).

2.2.2 Avatars

In CVEs, avatars are three-dimensional human-like representation of users (Benford et al., 1995; Meadows, 2007). Users interact with each other and elements in the virtual world via avatars. They are primarily used to reflect users' actions, and to mediate social interactions (Bowers, Pycock, & O'Brien, 1996; Hindmarsh, Fraser, Heath, & Benford, 2001; Hindmarsh, Fraser, Heath, Benford, & Greenhalgh, 1998; Meadows, 2007; Moore, Ducheneaut, & Nickell, 2007). For example, Bowers et al. (1996) observed how people interact during virtual meetings and pointed out that the orientation of an avatar helps identify who will talk next in a conversation, and that avatars' expressiveness has a significant role in turn-taking in conversations. Moore et al. (2007) observed how people play video games and found that the limited control mechanisms and expressiveness of avatars cause problems in social interactions, e.g., difficult in understanding intentions of others, creating actions with precise timing, and coordinating activities. To improve social interactions in CVEs, it is important that avatars have high expressiveness and appropriate methods to control their actions (Bowers et al., 1996; Moore, Ducheneaut, et al., 2007; Salem & Earle, 2000).

2.2.2.1 Expressiveness

To make an avatar more expressive, one can increase its flexibility by having more movable components and more variety in its appearance. In this section, I outline research about appearances, facial expressions, gaze directions, postures, and gestures of avatars.

Appearance. Avatars are not only used to represent users' actions, but also their identity (Turkay & Adinolf, 2010; Turkle, 1995). Identity, personality, and other personal attributes, such as fashion and body shape, can be expressed through the appearance of an avatar (Bessière, Seay, & Kiesler, 2007; Ducheneaut, Wen, Yee, & Wadley, 2009; Pace, 2008; Salem & Earle, 2000; Turkay & Adinolf, 2010; Turkle, 1995). Many commercial CVEs allow users to customize avatar appearance. Users can choose different races, genders, body shapes, outfits, hairstyles, and many other properties for their avatars. Figure 2.4 shows the avatar customization screen in City of Heroes/Villains.

Figure 2.4: Avatar customization in City of Heroes/Villains.

Turkay and Adinolf (2010) conducted a survey with players of World of Warcraft (WoW) and City of Heroes/Villains. They found that avatar appearance is an enjoyable feature, is one of the favourites among other customizations (e.g., sound, graphics, and user interface), and can affect players' interest in the games. Another survey was conducted by Bessière et al. (2007) with players of WoW and found that they tend to create avatars that look like an idealized version of themselves, and this phenomenon is often seen among players with high levels of depression and low self-esteem. The finding was further confirmed by Ducheneaut et al. (2009) who performed a study investigating appearance customization in WoW, Second Life (SL), and MapleStory. In addition, they found that avatars reflect the players' physical properties and limitations, and that hairstyle and colour were consistently considered to be the most important features in avatar appearance.

Facial Expressions. People often convey their emotion through facial expressions (Ekman, 1992). They are particularly useful when collaborators are in close proximity in CVEs because the details can be clearly seen (Salem & Earle, 2000). Users are able to identify, feel, and understand the mental states of other users in a CVE from observing the facial expressions of avatars (Fabri, Moore, & Hobbs, 2004; Fabri & Moore, 2005). Salem and Earle (2000) developed a system to synchronize text chat and expressions by storing some predefined expressions, such as blinking and smiling, and map them to strings of text. This method has been widely used in modern CVEs, e.g., WoW and SL. However, it raises problems such as users are required to memorize certain commands, only a limited set of expressions can be produced, and the duration of the expressions is difficult to control (Moore, Ducheneaut, et al., 2007).

Gaze Directions. In face-to-face interaction, gaze direction has useful functions such as providing feedback and regulating turn-taking (Kendon, 1967). Garau et al. (2003; 2001) compared random gaze directions and gaze directions that correspond to turn-taking in conversations. They found that showing meaningful gaze directions for avatars helped improve quality of communications. This finding was further supported by Steptoe et al. (2009) with a study of gaze influence on performance in object arrangement tasks. Furthermore, an avatar's gaze can be used as deictic reference (Duchowski et al., 2004). Duchowski et al. (2004) developed a system allowing users to use their real-world head or gaze direction to control where their avatars look (indicated by a red dot in the CVE). They found that using a red dot is an effective way to show gaze direction, and that using eye-control is more effective than head-control in manipulating an avatar's gaze. Using a red dot is not the only way to represent gaze. In other CVEs, gaze can be shown using view frustums with wire-frames (Fraser, Benford, Hindmarsh, & Heath, 1999) and semi-transparent colour (Fraser et al., 2004).

Postures. People often use postures in face-to-face communication (Heath, 1986), e.g., to indicate availability to engage in a conversation. In CVEs, the posture of an avatar can be used to indicate the state of the user (Moore, Ducheneaut, et al., 2007; Moore, Gathman, Ducheneaut, & Nickell, 2007; Salem & Earle, 2000). Moore et al. (2007) pointed out that when an avatar appears to be like a 'lifeless zombie' (i.e., standing still for an extended period of time with no other activities like text chat or voice chat), collaborators often assumed that the user of the avatar is away from the keyboard or busy at other things like navigating through menus. Moore et al. (2007) also suggested that creating meaningful postures to represent user actions that are hidden from other users can improve coordination between collaborators. For example, an avatar holding up a map can represent the user looking at a map of the CVE, and tilting the head and closing eyes can represent the user being away from the keyboard.

Gestures. One of the most important non-verbal communication channels in CVEs is gestures (Hindmarsh, Fraser, Heath, Benford, & Greenhalgh, 2000; Kirk, Crabtree, & Rodden, 2005; Kirk et al., 2007; Moore, Ducheneaut, et al., 2007; Salem & Earle, 2000). However, these gestures are mostly predefined with limited sets of variations (Salem & Earle, 2000) and are difficult to use effectively for communication (Moore, Ducheneaut, et al., 2007). While expressive gestures are often suggested by researchers (Bowers et al., 1996; Moore, Ducheneaut, et al., 2007; Salem &

Earle, 2000), using basic gestures to indicate objects is challenging (Fraser et al., 1999; Fraser & Benford, 2002; Hindmarsh et al., 2001, 2000). In order to make gestures easier to see and interpret, Fraser and colleagues (Fraser et al., 1999; Fraser & Benford, 2002) exaggerated the gestures by extending the length of the avatar's arm, and Hindmarsh et al. (2000) and Linebarger et al. (2003) used a thin line to connect the arm and the object of interest.

2.2.2.2 Inputs and Controls

There are many ways to control avatars. Some use specialized equipment such as body sensors (Peinado et al., 2009), eye trackers (Duchowski et al., 2004), and Omni-directional locomotion systems (Bouguila, Ishii, & Sato, 2002). Others involve more commonly-available devices, e.g., keyboards and mice (Mackinlay, Card, & Robertson, 1990) and gamepads (Templeman, Sibert, Page, & Denbrook, 2007).

Body sensors and trackers. Some researchers use body motions and gestures to control avatars by attaching sensors and trackers to the users. For example, Lee et al. (1998) used a hand-gesture recognition system along with tracking body orientation to control movements such as sitting, walking, and jumping; Peinado et al. (2009) put optical markers on users so that they can use their whole body to control an avatar.

Eye trackers. These are primarily used to control avatars' gaze directions. For example, Steptoe et al. (2008, 2009) developed a system called EyeCVE allowing users to control avatars' gaze with their own gaze. Duchowski et al. (2004) used an eye tracker for object selection in a CVE.

Omni-directional locomotion system (ODLS). An ODLS allows users to control avatar's movement by walking and turn in any direction in-place. Using ODLS can prevent users from walking out of range and from turning out of view. These systems can be built by using a turntable (Bouguila et al., 2002), slidable footwear (Iwata & Fujii, 1996), two overlapped perpendicular treadmills (Darken, Cockayne, & Carmein, 1997), or a giant sphere surrounding the user (Virtusphere, n.d.).

Keyboard and mouse. In the early 90's, Mackinlay et al. (1990) used a mouse and on-screen icons to control an avatar's body and gaze locations. Later, Salem and Earle (2000) used predefined text strings from a keyboard to control the facial expressions and gestures of avatars.

Nowadays, many modern PC-based CVEs use keyboard and mouse combination as default input devices.

Gamepad. Console-based CVEs (e.g., PlayStation Home) primarily use gamepads for controlling avatars. Some researchers explore customized gamepad settings to improve performance. Templeman et al. (2007) altered the conventional control mapping of a gamepad to make it more suitable for tactical movement (moving and aiming at the same time), by separating the control of avatar movement and viewing direction.

2.2.3 Awareness

Awareness is "an understanding of the activities of others, which provides a context for your own activity" (Dourish & Bellotti, 1992, p. 107). It is important in CVEs because awareness information helps coordinate group activities and is crucial to the success of collaboration (Benford, Bowers, Fahlen, & Greenhalgh, 1994; Benford & Fahlen, 1993; Dourish & Bellotti, 1992; Endsley, 1995; Fraser et al., 1999; Gutwin & Greenberg, 2002; Hindmarsh et al., 1998). In this section, I describe the concept of situation awareness, the framework of workspace awareness, and the model of focus and nimbus.

2.2.3.1 Situation Awareness

Endsley defined *situation awareness* as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, p. 792). Her definition has three levels (Endsley, 1995):

Level 1: Perception of the Elements in the Environment. An actor needs to gather relevant information about what is happening in the environment. This involves the kinds of elements in the surroundings, as well as their locations, attributes, and status. For example, a driver needs to know where other vehicles are, how fast they are moving, whether there are traffic signals, and the states of the signals.

Level 2: Comprehension of the Current Situation. An actor needs to synthesize the elements from level 1 and understand their importance based on the current situation. For example, a

driver notices that three vehicles in close proximity change lanes one after another and understands that the two vehicles at the back may be following the one in the front.

Level 3: Projection of Future Status. An actor needs to predict how the elements will change in the near future based on the knowledge of the elements (from level 1) and comprehension of the situation (from level 2). For example, when a driver knows that someone is drunk driving, they can avoid a possible accident by staying away from the drunk driver.

2.2.3.2 Workspace Awareness

Gutwin and Greenberg provided a framework of *workspace awareness*, which they defined as "the up-to-the-moment understanding of another person's interaction with a shared workspace" (Gutwin & Greenberg, 2002, p. 412). Workspace awareness can be seen "as a specialization of situation awareness, one that is tied to the specific setting of the shared workspace" (Gutwin & Greenberg, 2002, p. 417). A *shared workspace* is a bounded space that allows users to share and manipulate artifacts within a real-time distributed groupware system.

The framework was oriented towards small groups (three to five people) of users working together in a medium-sized shared workspace, and was built around the information of workspace awareness: the components of information involved, the mechanisms for gathering and maintaining the information, and the ways people use the information (Gutwin & Greenberg, 2002).

The elements of workspace awareness can answer the "who, what, where, when, and how" questions: e.g., who (Who is there? Who is doing that?), what (What is that? What are they doing?), where (Where are they? Where are they looking?), when (When did that happen?), and how (How did that happen?).

Gutwin and Greenberg (2002) state that one reason why workspace awareness is important is that it helps collaborators interpret visual signals such as deictic references. Knowing who is creating the reference, what is being referenced, and where the reference happens, are crucial to the understanding of deictic references. Although the framework of workspace awareness was developed mostly based on Gutwin and Greenberg's experience with 2D distributed groupware systems, it can be applied to other groupware such as CVEs.

2.2.3.3 Focus and Nimbus

Benford et al. (1994; 1993) developed a spatial model that manages awareness in CVEs. The key concepts of the model include aura, focus, nimbus, adapters, and boundaries. An *aura* bounds the presence of an object (Fahlén & Brown, 1992). If the auras of two objects intersect, they may interact with each other depending on the environment. *Focus* is the attention of observers, and *nimbus* is the projection of the information from the person being observed (Benford et al., 1994; Benford & Fahlen, 1993). The amount of awareness depends on whether and how focus and nimbus overlap. For instance, Ben is looking at the back of Alex. Alex's focus and Ben's nimbus do not overlap, so Alex has no awareness of Ben. However, Ben's focus intersects with Alex's nimbus; therefore, Ben has awareness of Alex (Figure 2.5). *Adapters* are objects that change one's focus and nimbus (Benford et al., 1994; Benford & Fahlen, 1993), e.g., a loudspeaker increases one's nimbus, while *boundaries* are dividers of areas that affect the properties of aura, focus, nimbus, and interactions between objects (Benford et al., 1994), e.g., walls and windows. MASSIVE-2 is a CVE system built based on these concepts (Benford, Greenhalgh, & Lloyd, 1997; Greenhalgh & Benford, 1995).

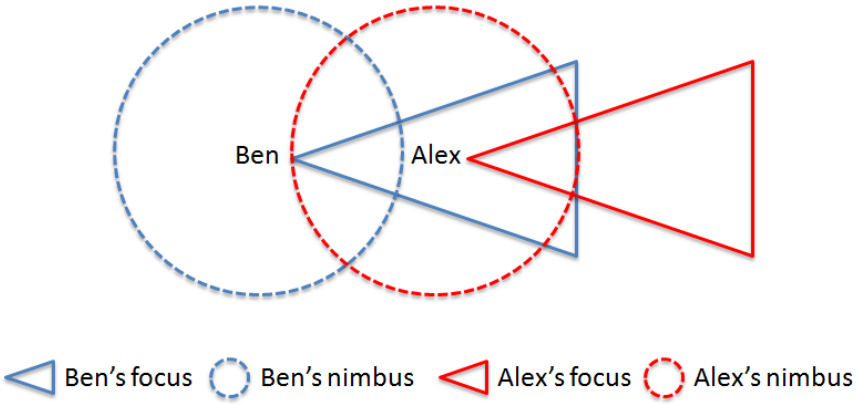


Figure 2.5: Focus and Nimbus.

2.2.4 Views

Establishing proper views in CVEs is critically important for the success of collaboration in CVEs. This primarily involves issues of field of view, perspective, and mutual orientation.

Fields of View. CVEs (especially desktop CVEs) have smaller fields of view (FoV) compared to the real world (Ellis, 1995). Narrow FoV limits one's range of focus, and so decreases one's

awareness of others. It also causes the problem of 'fragmentation'—the screen cannot display all the relevant things needed for communication (Hindmarsh et al., 1998). For example, an avatar and the objects related to the avatar's actions (such as a chair being pointed at) cannot be seen simultaneously. Users need to change their view multiple times to look for their collaborators and the objects of interest to make sense of conversations and activities, disrupting the flow of collaboration. Fragmented views also slow down collaboration because users often need to explicitly describe the action in more detail to compensate for fragmentation (Fraser et al., 1999; Hindmarsh et al., 1998, 2000).

Hindmarsh, Fraser, and colleagues (Fraser et al., 1999; Hindmarsh et al., 2000) attempted to solve the problem of fragmentation with peripheral lenses (Robertson, Czerwinski, & Van Dantzich, 1997). Peripheral lenses are small columns on the left and right edges of the screen. The lenses have high FoV showing the peripheral view, and so the screen can show a wider scene with the total FoV increased. While this technique helps show a larger scene and improves awareness, the images in the peripheral lenses are compressed (high FoV in a small area) causing much distortion. This makes it difficult to perceive other avatars' orientation and what other can really see (Fraser et al., 1999).

Perspectives. There are two commonly used perspectives in CVEs: first- and third-person view. In a first-person view, the camera of the scene is located at the eyes of the avatar (Figure 2.6A). In a third-person view, the camera is moved behind the avatar (Figure 2.6B). The main visual difference between the two views is that in a first-person view users are able to see the virtual world from the avatar's perspective but cannot see the avatar (they may be able to see very limited parts of the avatar such as hands in some CVEs); whereas, in a third-person view users can see the whole avatar and all of its actions (Rouse, 1999). Rouse (1999) discussed the two perspectives in games in different genres (e.g., shooting, adventure, and role playing). He mentioned that players with a first-person view can have a better sense of immersion and can associate more closely with the game characters, but players with a third-person view can have a much stronger sense of the characteristics of the game characters. He suggested that different perspectives are better in different situations depending on the goals of the games. His suggestions aligned with the results of Salamin et al.'s study that compared the two perspectives in virtual and augmented reality (Salamin, Thalmann, & Vexo, 2006). They found that different

views are better in different situations. In particular, a first-person view is good for moving objects and a third-person view is good for avatar movement. While the study was not done with a CVE, they argued that the results can be generalized to CVEs. Recently, Bateman et al. (2011) compared the two views in a driving simulation, but they did not find any significant differences of first- and third-person views in performance with driving tasks.

A  B 

Figure 2.6: Perspectives in Second Life: A) first-person view and B) third-person view.

Mutual Orientation. Establishing mutual orientation is important for communication in the real world. To have a successful conversation, people often need to see the person they are having a conversation with and the things they referring to (Goodwin, 2000; Heath & Hindmarsh, 2000; Hindmarsh & Heath, 2000). This also applies in CVEs. People arrange their avatars so that they can see each other's actions and the objects of interest (Fraser et al., 1999; Hindmarsh et al., 1998, 2000). This process is particularly difficult when the CVE has a narrow FoV, because it is unlikely that the object of interest and the avatars are in the same view at the onset of a conversation. Although it is not easy to establish mutual orientation in CVEs, Hindmarsh et al. (1998, 2000) found that people take time and effort to arrange their avatars to compensate for the narrow FoV and to maintain awareness of each other. To help establish mutual orientation, Fraser et al. used wire-frame models (Fraser et al., 1999) and semi-transparent colours (Fraser et al., 2004) to show the view frustums of avatars. However, users found it was confusing when the view frustums were used with peripheral lenses.

2.3 Deictic References in Collaborative Virtual Environments

Deictic referencing has an important role in collaboration in CVEs. It helps in grounding communicational conduct (Clark, Schreuder, & Buttrick, 1983; Heath et al., 2001; Hindmarsh et al., 1998; Luff et al., 2011; Moore, Ducheneaut, et al., 2007) and can simplify complex verbal descriptions (Bangerter, 2004; Fussell et al., 2004). Deictic referencing can be based on virtual pointers or avatar based. Virtual-pointer-based references can be controlled by users directly without an avatar. Avatar-based references, on the other hand, are based on various actions of an avatar. The user needs to control the avatar to create the references.

2.3.1 Virtual-Pointer-Based References

There are two main types of virtual pointers: virtual cursors and ray casting.

Virtual Cursors. A virtual cursor is a 3D cursor in a virtual environment that allows users to control its location in 3D (Hinckley, Pausch, Goble, & Kassell, 1994). Poupyrev et al. (1996) developed an interaction technique called Go-Go to control a hand-shaped virtual cursor. Zhai et al. (1994) applied semi-transparency to a rectangular virtual cursor called Silk Cursor. With transparency, users are able to know if an object is in front of, within, or behind the cursor, thus making object selection easier. However, it is difficult to select a single target from a group of dense objects. Vanacken et al. (2007) developed 3D Bubble Cursor that is based on the ideas of Silk Cursor (Zhai et al., 1994) and Bubble Cursor (Grossman & Balakrishnan, 2005). The Bubble Cursor is a 2D area cursor (a cursor that covers an area of a 2D surface) that can dynamically change its size to always select only the closest object. Vanacken et al. applied a 3D effect and transparency (from the Silk Cursor) on the Bubble Cursor to form the 3D Bubble Cursor allowing users to select a single target in a dense environment.

Ray Casting. A visual technique of projecting a ray from the user's hand or input devices is called ray casting. The most basic kind of ray casting allows users to select the first object it intersects (Liang & Green, 1994; Mine, 1995). This method is straightforward and easy to use, but is limited to only one selectable object at a time and cannot select occluded objects. To overcome these problems, many variations of ray casting were developed. Liang and Green (1994) changed the traditional ray to a cone-shaped selection volume, called Spotlight, with the

apex at the input device. With the increased volume, users can select multiple objects. Wyss et al. (2006) developed the iSith technique that uses two rays for object selection. The user controls the direction of one ray with each hand. An object can be selected by intersecting the rays at the object. This allows users to select objects that are located behind other objects. Olwal and Feiner (2003) introduced a flexible pointer that allows users to bend an arrow-shaped ray. Users can bend the arrow to refer to partially occluded objects at the back of the environment without intersecting with the occluding objects in the front. Vanacken et al. (2007) developed the Depth Ray by adding a depth marker to a basic ray. Users can control the location of the marker along the ray by moving the hand forwards or backwards. The object that is intersected with the ray and is the closest to the depth marker can be selected. When using the Depth Ray with transparency applied on the objects near the ray, users can select occluded objects.

2.3.2 Avatar-Based References

Avatar-based references are typically accomplished by avatar pointing gesture, gaze direction, and orientation.

Pointing Gestures. Although few CVEs allow avatars to point with their arms, two methods have been suggested to aid referential communication: object highlighting and connecting lines. Hindmarsh, Fraser, and colleagues suggested that visually highlighting objects in CVEs could be used to help collaborators identify objects of interest (Fraser et al., 2004; Hindmarsh et al., 1998, 2000). Object highlighting is implemented in Second Life (SL). In SL, users can control an avatar to point at individual objects by clicking on them. The outline of the object will be highlighted. While highlighting makes objects easier to identify, this method restricts what can be pointed at because some referents are not objects. Referents can be directions, spaces between objects, or general areas, which are not supported by this pointing method. In addition, the granularity of highlighting can lead to misunderstanding. For example, when a user wants to point at the door of a house, the house could be highlighted instead of the door if the door is not a predefined ‘highlightable’ object. On the other hand, if the door is ‘highlightable’ and the user wants to point at the house by clicking the door (which is part of the house), the door will be highlighted instead of the house.

Another suggestion by Hindmarsh et al. (1998, 2000) to help identify objects is to connect the avatar to the object by a line. This method is also used in SL in that a dotted line connects the avatar's hand and the thing being pointed at. However, the problems of limited 'pointable' objects and granularity still exist with this method. In addition, when many avatars perform this kind of pointing simultaneously, lines would overlap creating confusion. Hindmarsh, Fraser, and colleagues also built a variation of this technique by lengthening the arms so that they reach the objects (Fraser et al., 1999; Hindmarsh et al., 2000), but they found that this method slowed down the flow of collaborations and caused confusions between collaborators (Fraser et al., 1999).

Gaze Direction. Duchowski et al. (2004) explored how deictic referencing can be performed through gaze direction of an avatar in a CVE. A red dot represented the location of where the avatar was looking. They compared two ways to control gaze direction. The user could control the avatar's gaze by his/her head movement or gaze direction. They found that using gaze direction to control visual deictic reference in CVE caused less confusion than using head movement especially when the user's line of sight is different from the head direction.

Orientation. Avatar orientation can often help the process of grounding conversational conduct (Hindmarsh et al., 1998, 2000; Moore, Ducheneaut, et al., 2007), thus make understanding deictic references easier. In the real world, people's orientation is a strong indicator of what they can or cannot see. We often assume that people can see the objects that their body is facing. However, we cannot assume the same in CVEs. Moore et al. (2007) pointed out that some CVEs allow users to change their view without moving their avatar, and therefore, the user may not be looking at where the avatar is facing. For example, Figure 2.7A shows what the user of an avatar is actually seeing. However, this is not known by other people. They might think the user has the view as in Figure 2.7B, which is aligned with the avatar's orientation. This can cause confusion if someone is referencing an object in front of the avatar and assumes that the user of the avatar can see it.

A  B 

Figure 2.7: A) The view of the avatar user; B) the view that other people think the user sees.

CHAPTER 3
A FRAMEWORK OF DISTANT POINTING

To better understand distant pointing and to improve the expressiveness of pointing gestures in CVEs, I construct a conceptual framework that describes the stages and enactment of distant pointing, and provides a list of questions that help in designing distant-pointing techniques. I then discuss the problems of existing distant-pointing methods in CVEs and provide design principles derived from those problems.

3.1 What Is Distant Pointing?

Distant pointing is a type of deictic pointing where the referent is out of reach. For example, pointing at a restaurant across the street and saying "I had lunch at that restaurant <point>" is distant pointing. In distant pointing, the gesturer builds a connection between him/herself and a referent without touching it. Because the gesturers are not able to touch the referents, distant pointing is much different from non-distant pointing (deictic pointing where the referent is within reach).

One major difference is that gesturers in distant pointing may not be pointing directly at the objects that they want to reference. Because referents are out of reach and can be far away, the gesturers may point at objects that are near to the referent. Whether the gesturers are actually pointing at the referent can be ambiguous, thus affecting how people communicate. Although this problem can be alleviated by speech, distant pointing is not as accurate as non-distant pointing where the gesturers can point at the referent up close or even put their finger on the referent.

Achieving mutual understanding is therefore more difficult when communicating with distant pointing. The gesturer and observer need to be able to see each other, the pointing gesture, and the referent for effective communication. When the gesturer points at a referent from a distance, the observer has to switch focus from the pointing gesture to the referent (and sometimes back and forth between them). When a referent is nearby, the observer does not need to switch focus

in this way because the referent is right next to the gesturer. These characteristics of distant pointing may not seem to be problematic in the real world; however, they can hinder gestural communication in CVEs.

3.2 Stages of Distant Pointing

In order to use distant pointing effectively in communication, people need to pay attention to more than hand gestures—therefore, distant pointing consists of body movements other than just hand gestures. Distant pointing has four stages: orientation, preparation, production, and holding. These stages are the fundamental building blocks of expressive distant pointing in both the real world and CVEs.

Orientation. Establishing mutual orientation (Fraser et al., 1999; Hindmarsh et al., 1998, 2000) is critical to pointing especially distant pointing. When a gesturer wants to indicate a referent to an observer with distant pointing, the gesturer needs to see where the referent is in order to point at it. The observer must also see the referent as well as the pointing gesture to understand what the gesturer wants to show. So, the gesturer needs to orient him or herself so that the gesturer and observer can see both the pointing action and the referent (Figure 3.1).

Figure 3.1: Orientation.

Preparation. After achieving mutual orientation, the gesturer makes preparatory motions that often indicate to the observer that a pointing gesture is about to be made. This stage is similar to the preparation phase as described in Section 2.1.2. In addition to liberating movements (for freeing the hands from a constrained locations), location preparation (to move the arm to the

starting position), and hand internal preparation (for shaping and orienting the hand), the preparation stage described here includes other body movements. These movements can be moving the head, turning the torso, and changing gaze directions. All of these preparation actions not only set the stage for the pointing gesture, but also provide useful information to the observer letting the observer knows that a pointing gesture is about to be made (Figure 3.2).

Figure 3.2: Preparation.

Production. The gesturer is ready to produce the pointing gesture after preparatory actions are done. The production of a pointing gesture (the stroke phase as described in Section 2.1.2) is not immediate (see Figure 3.3). The gradual production of the action, together with information from previous stages and conversational content, allow the observer to predict the general direction of the gesture before it is completed (Moore, Ducheneaut, et al., 2007).

Figure 3.3: Production.

Holding. Once a pointing gesture is produced, the gesturer holds the gesture to make sure the observer has seen it (Figure 3.4). This holding stage (as described in Section 2.1.2) is the most

important stage because it provides the most information to the observer. In this stage, the pointing gesture creates and maintains a connection between the gesturer and the referent. The gesturer needs to hold the gesture until a mutual understanding of the referent has been achieved.

Figure 3.4: Holding.

Figure 3.1 to Figure 3.4 illustrate how the four stages of distant pointing are used in a conversation (Alex is asking Ben the direction to a classroom). First, they need to establish mutual orientation (Figure 3.1). They orient themselves to the general direction of the referent in such a way that they are also slightly facing each other. Doing so allows them to be able to aware of each other's actions and to make sure they mutually know what the referent is going to be. Ben then proceeds to the preparation stage (Figure 3.2). He takes his hands out of his pockets and gets ready to point. At this time, Alex knows that Ben is about to show him where the classroom is, thus shifting attention from the map he is holding to Ben. Once Ben knows that he has Alex's attention, he produces the pointing gesture (Figure 3.3). Alex then follows the gradual production of the gesture and turns his head toward the general direction of the classroom, although he does not yet know its exact location. Finally, Ben points at the classroom and holds the pointing gesture (Figure 3.4) making sure that Alex knows where it is by allowing him to check between the gesture and the referent.

3.3 Enactment of Distant Pointing

Enactment of pointing (i.e., how it occurs) can be characterized in terms of properties, such as speed, flexibility, movement, and visual effect:

- Speed—how fast a gesture is generated. Gestures may have different meanings when generated with different speed. For example, a quickly-generated gesture shows confidence while a slowly-generated gesture shows hesitation.
- Flexibility—how flexibly a gesture is generated. The ways the elbow, wrist, and fingers move when generating a pointing gesture can affect how the gesture is perceived. Also, generating a pointing gesture with the arm straight may have different meanings than with the arm bent.
- Movement—how freely the arm can move. Whether the arm can freely move to point at any directions greatly affects how people use pointing gestures. For example, if someone has an arm injury and can only point to the side, they need to turn and then point in order to point at something that is originally in front of them.
- Visual effect—how different pointing gestures appear visually. Visual effects, such as a laser dot (pointing with a laser pointer) and a light beam (pointing with a flashlight), can change how pointing gestures are generated and perceived. Without such visual aids, it would be difficult to precisely point at something at a distance.

Among these four properties, movement and visual effect are particularly important in the design of pointing techniques for CVEs. Arm movement, as seen in the previous section, is crucial to the production and holding stages, which provide most information about the location of referents. Visual effect is especially important in CVEs where many physical limitations of the real world do not exist. Many more visual effects, such as changing the appearance of the whole referent (object highlighting) and extending the length of a gesturer's arm (Fraser et al., 1999; Hindmarsh et al., 2000), can be used in CVEs but are not available in the real world. Therefore, I characterize the enactment of distant pointing in terms of movement and visual effect.

Movement describes how freely a person can point. When the pointing direction is independent from other actions of the gesturer (such as where the gesturer is facing or looking), this type of pointing is called *free pointing*. Free pointing is how people normally point in daily activities, and people can point anywhere they want, e.g., point at a referent while keeping eye contact with the interlocutor (Figure 3.5). The opposite of free pointing is called *restricted pointing*, which is commonly used in CVEs. Many pointing gestures in CVEs are command based and have a limited number of preset animations (Moore, Ducheneaut, et al., 2007) (Figure 3.6A). As a

result, most avatars can only point at the direction they are facing and where they are looking, i.e., the centre of their current view (Figure 3.6B). Restricted pointing is also seen in the real world although it occurs more rarely (e.g., a gesturer holding something or with arm injuries that restrict arm movements).

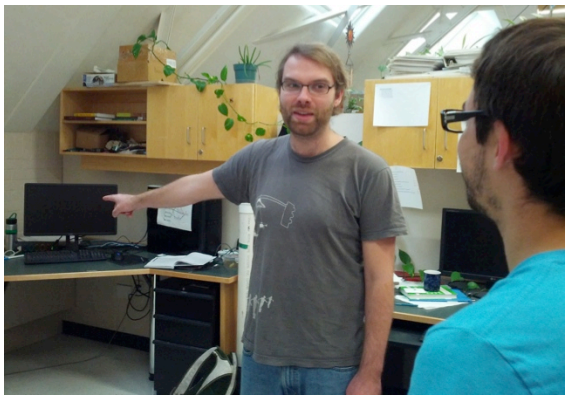


Figure 3.5: Free pointing: a gesturer is pointing at a referent while keeping eye contact with the interlocutor.

A  B 

Figure 3.6: Restricted pointing: A) the avatar can only point forward with fixed animations, and B) at the centre of the view.

Visual Effect describes how different pointing gestures appear visually. A pointing gesture that does not have any visual effect other than the movement of the gesture is called *natural pointing*. It is how people naturally point at things in the real world. Figure 3.7 shows examples of natural pointing in the real world and a CVE. Conversely, *augmented pointing* is pointing that has additional visual effects. For instance, when a gesturer in the real world uses a laser pointer to point at a referent (e.g., an object on a projector screen), a laser dot appears on the referent. As shown in Section 2.3.2, augmented pointing is used more often in CVEs with techniques such as object highlighting and connected lines between gesturers and referents (Figure 3.8).

Figure 3.7: Natural pointing: A) in the real world; B) in a CVE.

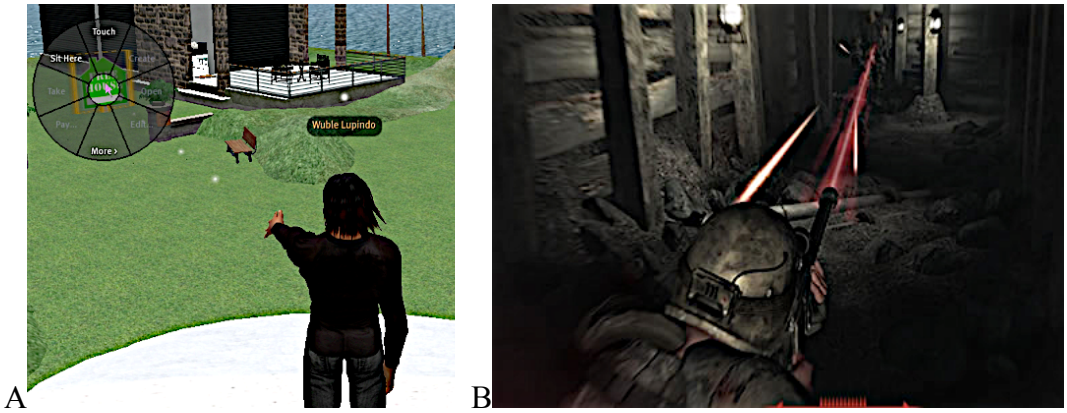


Figure 3.8: Augmented pointing in CVEs: A) object highlighting with a dotted line connecting the object and the avatar, and B) a laser gun.

3.4 Design Questions for Distant-Pointing Techniques

In this section I present four basic design questions (see Table 3.1) that need to be answered when designing and developing distant-pointing techniques. I then address these questions based on the stages and enactment of distant pointing.

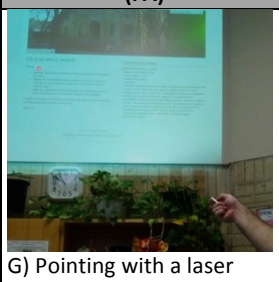
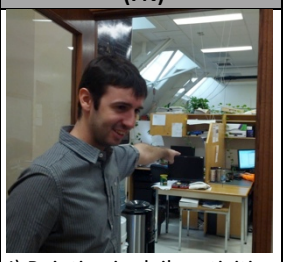
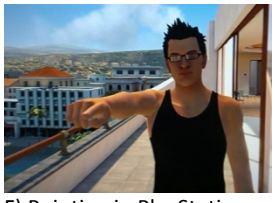
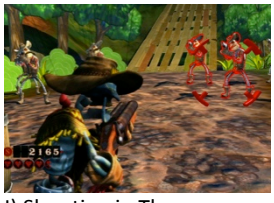
Table 3.1: Design questions for designing and developing distant-pointing techniques.

Design Questions for Distant-Pointing Techniques
1. *How accurate is the type of pointing?* 2. *How to control the type of pointing?* 3. *How visible is the type of pointing?* 4. *Can the stages (i.e., orientation, preparation, production, and holding) be shown by the type of pointing?*

3.4.1 Types of Distant Pointing

The design questions of Table 3.1 can be addressed by different types of pointing based on movement and visual effect. As described in Section 3.3, pointing gestures characterized by movement are either restricted or free, and pointing gestures characterized by visual effect are either augmented or natural. So, when combining movement and visual effect, there are four types of distant pointing: restricted and augmented (RA), restricted and natural (RN), free and augmented (FA), and free and natural (FN). Here, I provide some typical examples for each type of distant pointing in the real world and CVEs (see Table 3.2).

Table 3.2: Examples of distant pointing.

	Restricted and Augmented (RA)	Restricted and Natural (RN)	Free and Augmented (FA)	Free and Natural (FN)
Real World	 A) Pointing with a flashlight on a helmet	 D) Pointing while holding something	 G) Pointing with a laser pointer	 J) Pointing in daily activities
CVEs	 B) Shooting in Return to Castle Wolfenstein  C) Indicating point of interest in Portal 2	 E) Pointing in World of Warcraft  F) Pointing in PlayStation Home	 H) Highlighting an object in Second Life  I) Shooting in The Gunstringer	No current CVE support

Restricted-and-augmented (RA) pointing. In the real world, RA pointing is uncommon and is used only in some specific situations. For example, when a firefighter uses the flashlight mounted on the helmet to point (Table 3.2A), the pointing direction of the flashlight is bounded to where the head is facing (restricted) and flashlight forms a light beam towards the referent (augmented). In CVEs, for example, RA pointing occurs when firing a machine gun in Return to Castle Wolfenstein (Table 3.2B), a first-person shooter (FPS) game, and when using an indication mechanism in Portal 2 (Table 3.2C), a collaborative puzzle game. In both cases the avatar can only point at the centre of the screen, thus they are restricted. They are also augmented because firing creates tracers and bullet holes and the indication mechanism creates laser beams. Tracers, bullet holes, and laser beams are additional visual effects to arm movement.

Restricted-and-natural (RN) pointing. This kind of pointing occurs in the real world when arm movement is limited, e.g., when someone is holding something (Table 3.2D). In CVEs, command-based pointing with fixed animations, such as pointing in fantasy role-playing games (e.g., World of Warcraft in Table 3.2E) and social environments (e.g., PlayStation Home in Table 3.2F), is restricted and natural. The avatar can only point with some preset gestures and no extra visual effect was used.

Free-and-augmented (FA) pointing. A typical example of FA pointing in the real world is pointing at something on a projector screen with a laser pointer (Table 3.2G). There is no restriction on the arm movement (free) and the laser dot shows up on the screen (augmented). In CVEs, pointing with object highlighting, e.g., in Second Life (Table 3.2H), a social environment, and The Gunstringer (Table 3.2I), a 3D shooting game, changes the appearance of the referent but the hand-and-arm movements are not limited by other actions of the avatar, so object highlighting is free and augmented. Note that not everything can be highlighted in CVEs. Although object highlighting is considered a kind of FA pointing, it is limited by whether the objects are highlightable and where they are located.

Free-and-natural (FN) pointing. This kind of pointing has no restriction in where the avatar can point and does not use a visual effect. FN pointing is commonly used in real-world communication (Table 3.2J); however, it is not available in current CVEs.

3.4.2 Answers to Design Questions

Here I list the characteristics of the four types of distant pointing (RA, RN, FA, and FN; see Figure 3.9) in Table 3.3, and then answer the four design questions in Table 3.4.

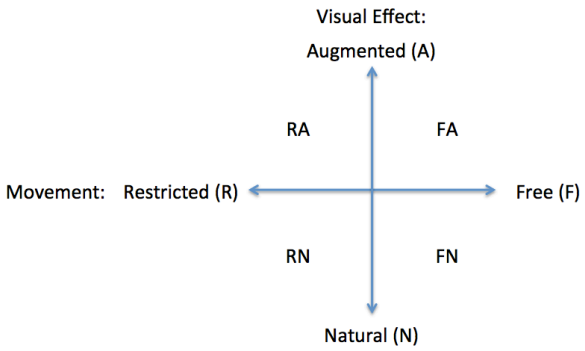


Figure 3.9: Distant pointing (RA, RN, FA, and FN).

Table 3.3: Characteristics of different types of distant pointing

Distant Pointing Characteristics
Restricted-and-augmented (RA) pointing • arm movement is unavailable or limited (no subtle control of movement and speed) • augmented techniques, e.g., object highlighting and a laser beam, can be used • accuracy is low because of restricted arm movement • accuracy can be improved by augmented techniques • visibility is depended on the augmented techniques used
Restricted-and-natural (RN) pointing • arm movement is unavailable or limited (no subtle control of movement and speed) • accuracy is low because arm movement is restricted and no extra visual cue is available • pointing and non-pointing states are difficult to distinguish (e.g., avatars with no explicit pointing gestures as in FPS games)
Free-and-augmented (FA) pointing • the arm can move freely • augmented techniques, e.g., object highlighting and a laser beam, can be used • accuracy is depended on the augmented techniques used (e.g., a laser beam is possibly more accurate than an elongated arm) • visibility is also depended on the augmented techniques used (e.g., object highlighting is possibly more visible than a laser dot)
Free-and-natural (FN) pointing • the arm can move freely • accuracy is low because no extra visual cue is available • visibility is also low for the same reason

Table 3.4: Answers to the design questions.

			Types of Distant Pointing			
			Restricted Augmented (RA)	**Restricted Natural (RN)**	**Free Augmented (FA)**	**Free Natural (FN)**
Examples of Type			- Pointing with a flashlight on a helmet - Shooting in FPS games - Indicating point of interest in first-person puzzle game	- Pointing while holding something - Pointing in World of Warcraft - Pointing in PlayStation Home	- Pointing with a laser pointer - Highlighting an object in Second Life - Shooting in The Gunstringer	- Pointing in daily activities
Design Questions	*1) How accurate is it?*		- Low (restricted pointing direction, but can be improved by using augmented techniques with higher accuracy)	- Low (restricted pointing direction)	- Vary (depending on augmented techniques)	- Low
	2) How to control it?		- No explicit control (fixed arm) - Text command - Key press (to activate augmented techniques)	- No explicit control (fixed arm) - Text command	- Move the arm with different input devices - Select the object of interest	- Move the arm with different input devices
	3) How visible is it?		- Vary (depending on augmented techniques)	- Low (also not obvious when there is no distinction between pointing and non-pointing)	- Vary (depending on augmented techniques)	- Low
	4) Can the stages be shown?	**Orientation**	Yes, but not suitable (may mislead observers that the referent is already being pointed at)	Yes	Yes, but not suitable (may mislead observers that the referent is already being pointed at)	Yes
		Preparation	No	No	Yes, but not suitable (the same reason as above)	Yes
		Production	Yes, if the arm is movable; no otherwise	Yes, if the arm is movable; no otherwise	Yes	Yes
		Holding	Yes	Yes	Yes	Yes

3.5 Problems of Existing Distant Pointing Methods in CVEs

The distant pointing methods mentioned above can be used to indicate referents in CVEs; however, these methods have several problems.

Pointing direction is tied to the centre of the screen (restricted pointing). While it is common to point where we are looking, the exact pointing direction may not necessarily be the centre of the screen. For example, imagine that Alex needs to point out five different referents to Ben. Alex adjusts his view so that the five referents are all on the screen with one in the bottom and two on each side (see Figure 3.10). He wants to point at each of them without losing sight of any of the referents. This cannot be done if the avatar can only point at the centre of the screen because Alex needs to change his view every time he points at a different object.

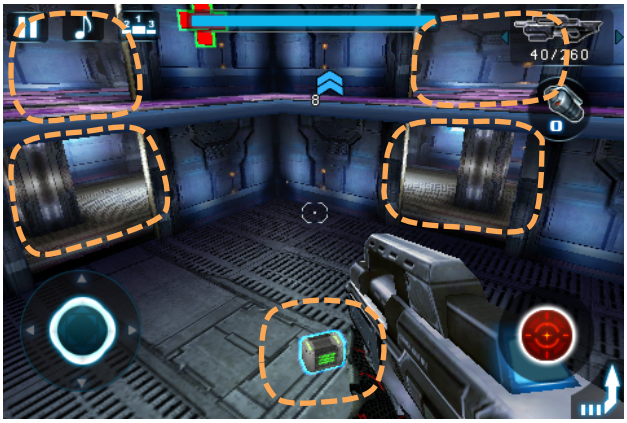


Figure 3.10: Five referents (highlighted with dotted lines) locate in different regions of the screen.

Users cannot control how pointing gestures are generated (restricted pointing). Restricted pointing is often generated with fixed animations and activated by command-based input. Once the command is executed, the pointing gesture cannot be paused or changed. Users have no control over how fast pointing gestures are generated. Therefore it is difficult for users to synchronize pointing gestures with speech, and difficult to generate subtle pointing gestures.

Users need to remember pointing commands (restricted and augmented pointing). Users need to use special commands, e.g., "/point" in World of Warcraft, or navigate through menus, e.g., selecting "Conversation" then "Point" in a pop-up menu in PlayStation Home (Figure 3.11), to activate pointing gestures and augmented-pointing techniques. These commands can be hard to remember especially when multiple pointing gestures and techniques are available.

Figure 3.11: Pointing in PlayStation Home: A) top-level menu, B) second-level menu, and C) pointing gesture.

Highlightable objects are predefined by CVE designers (augmented pointing). Users have no control of what can be highlighted. If the referent is an object that cannot be highlighted, the user needs to point at the closest object and rely on verbal description to clarify the referent (e.g., "the bus station is in front of the restaurant that I highlighted" instead of "the bus station is there <highlighting the bus station>"). Also, users have no control over the granularity of highlightable objects. Users may only want to refer to one part of the whole highlighted object or refer to the bigger object that contains the highlighted part. For example, in Figure 3.12, a single chair cannot be highlighted because all the chairs and the table are considered as one single highlightable unit in the CVE. They are highlighted together even when only one chair is selected. The users need to clarify the referent with speech. Another problem is that object highlighting cannot be used on empty space (e.g., areas between objects, or general directions).

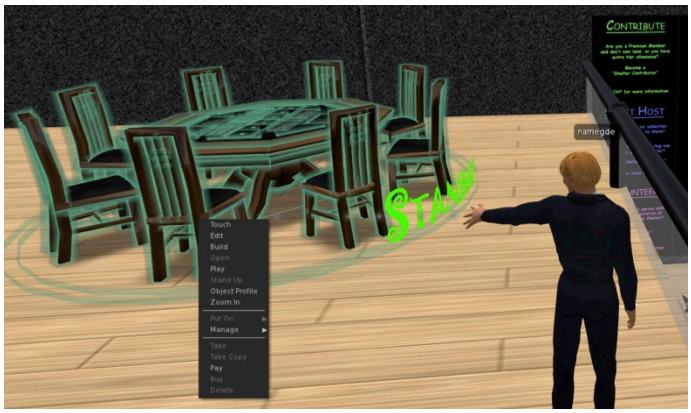


Figure 3.12: Users cannot control the granularity of object highlighting in Second Life.

3.6 Design Principles for Distant Pointing in CVEs

In this section, I discuss design principles for distant pointing in CVEs. These principles are derived from the problems of existing distant pointing mentioned above.

1. ***Pointing gestures should not be restricted by other avatar actions.*** People often point at things that are not at the centre of their focus. This happens more often during the holding stage of the pointing process. For instance, when the gesturer is holding the gesture for the observer to see, the gesturer looks at the observer to make sure he/she knows where the referent is (see the example discussed at the end of Section 3.2). Being able to separate pointing gestures with other avatar actions is important for establishing mutual understanding.

2. ***Users should be able to control the production stage of pointing gestures.*** There are different properties, such as speed and direction, in the production stage of pointing gestures. In order to successfully use pointing gestures for communication, the production of the gestures needs to be synchronized with other communicational conduct, e.g., speech and gaze direction (Goodwin, 1981; Heath, 1986; Hindmarsh & Heath, 2000). Without control over pointing speed and direction, timing and aiming gesture to match speech content becomes difficult. Being able to control these properties during the production of pointing is important to communication.

3. ***Pointing gestures should be easy to generate.*** Pointing gestures should be controlled by easy-to-use input devices and should have intuitive control schemes. Users should not need to memorize complex commands for generating pointing gestures.

4. ***Avatars should be able to point anywhere.*** Referents can be objects, areas, paths, directions, and empty space. Distant pointing should be able to point at all of these kinds of referents, not only to objects (as with object highlighting).

CHAPTER 4
OBSERVING DISTANT POINTING

To design and develop techniques for improving the expressiveness of gestural communication in CVEs, an understanding of the important characteristics of distant pointing is needed. However, there is little previous work available to provide this information. Therefore, I conducted an observational study to identify and explore different aspects of distant pointing. In the study, I observed the way that people point at distant referents in the real world, and the way that people interpret others' pointing gestures. In this chapter, I describe the study, explain five important aspects of distant pointing that the study identified, and discuss how the insights help designing and developing distant-pointing techniques in CVEs.

4.1 Setting

Distant pointing can be used to point at a wide range of referents. They include objects (e.g., a building and a road sign), areas (e.g., a parking lot, and an empty field), paths (e.g., a path from one building to another), and directions. Figure 4.1 shows examples of these referents. The referents can have varying visibility and distance from gesturers and observers. For example, a road sign can be close and fully visible (Figure 4.2), a car can be partially occluded (Figure 4.3), buildings can be distant but still visible (Figure 4.4), and a landmark can be too far away to see. This study explored how people generate and interpret pointing gestures for all of these referents.

4.2 Method

In this section, I describe the participants, experimental setup, procedure, and tasks.

4.2.1 Participants

I recruited four pairs of participants (7 male, 1 female); ages were between 21 and 40. There were two undergraduate students, five graduate students, and one faculty member. The

undergraduate students were from the Department of Chemistry and Kinesiology; the graduate students and the faculty were from the Department of Computer Science.

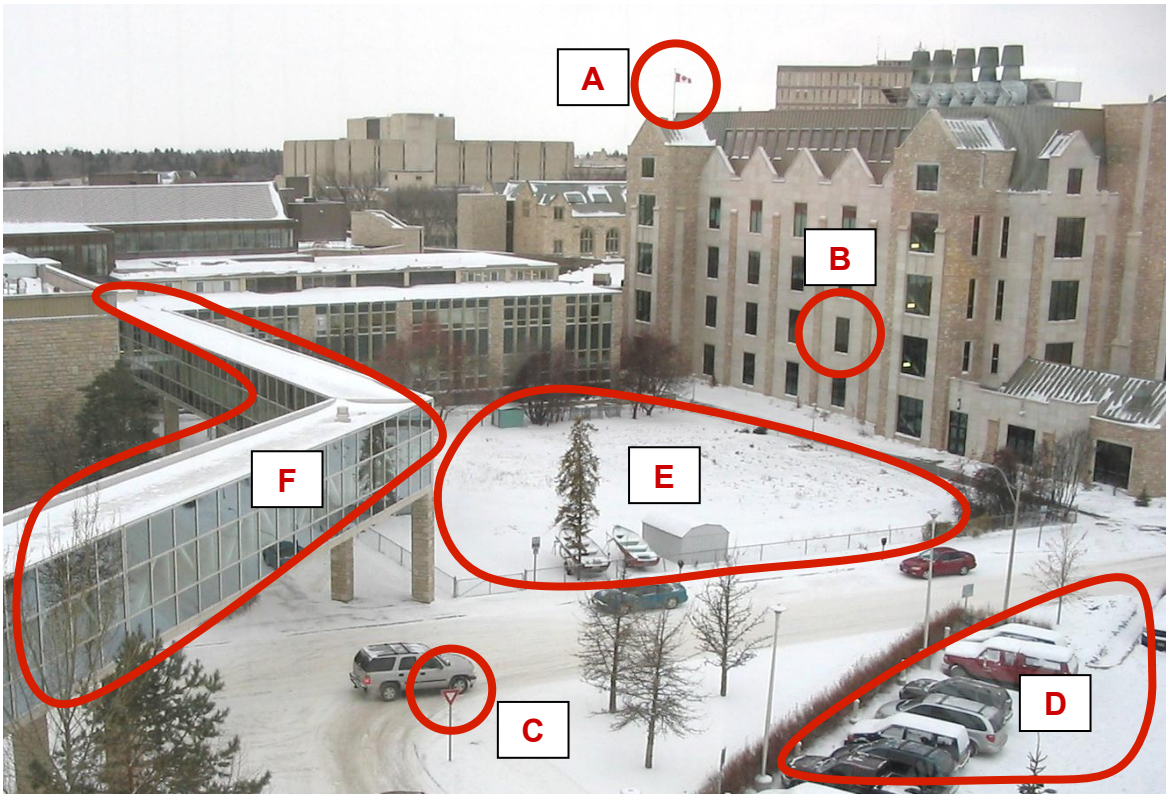


Figure 4.1: Object: A) a flag, B) a window, and C) a road sign; area: D) a parking lot, and E) an empty field; path: F) a skywalk connecting two buildings.

Figure 4.2: A road sign that is close and fully visible.

Figure 4.3: Cars that are partially occluded by trees.

Figure 4.4: Buildings that are far away but visible.

4.2.2 Experimental Setup

The study was carried out on the fifth floor of a building (Figure 4.5), with large windows overlooking a city. During the study, I observed how pointing gestures were generated and interpreted by the participants; I also video recorded the study for further analysis. I reviewed the videos and looked for instances that were important for referential communication.

Figure 4.5: A hallway where the study took place.

4.2.3 Procedure

For each session, the participants were informed with the purpose of the study, signed a consent form, and then did the tasks (described in the section below). At the end, the participants were debriefed and were given $10.00 as remuneration. The study took about one hour to complete.

4.2.4 Tasks

The study had three types of tasks: task 1 had different types of referents with different communication channels and visibility; task 2 involved users choosing their referents; and task 3 was a collaborative decision-making task. There were a total of 75 tasks and all of them involved distant pointing. Figure 4.1 through Figure 4.4 show some of the referents, and Figure 4.6 shows how a gesturer (who generates pointing gestures) and an observer (who interprets pointing gestures) were working on a task.

Figure 4.6: Participants were working on a task.

Task 1. In the first set of tasks, one participant was the gesturer and the other was the observer, and they interchanged roles halfway through the tasks. The gesturer was given photographs of different referents (e.g., the circled referents in Figure 4.1) that were outside the windows. I asked the gesturer to indicate the referents to the observer using three different communication channels (gestures only, gestures and written notes, or gestures and speech). The purpose of using these channels was to find out how people communicate when they were restricted by common CVE communication settings. *Gesture only* was used for simulating CVEs in which text and voice chat are unavailable; participants were not allowed to talk and pass notes to each other. *Gestures and written notes* was used to simulate text-chat-only CVEs; participants were not allowed to speak, but they could write notes on a notebook that was passed between them. *Gestures and speech* was used for simulating CVEs that support voice chat; participants were allowed to use gestures and talk to each other.

There were three kinds of referents: objects (e.g., a window, a flag, or a road sign), areas (e.g., a parking lot), and paths (e.g., a path between two buildings). The referents were either directly

visible, partially occluded, or completely out of view (e.g., a playground that was behind some buildings and places that were too far to see). Examples of these referents are shown in Figure 4.1 through Figure 4.4.

Task 2. The second set of tasks also involved a gesturer and an observer. The gesturer indicated ten different referents of their choice outside the building with only pointing gestures. The observer needed to figure out the referents one at a time and as quickly as possible. The observer continued stating to the gesturer what they thought the referent was until the gesturer confirmed that the answer was correct. The participants repeated the task with interchanged roles after the first ten referents were identified.

Task 3. For the third task type, I asked participants to collaboratively decide upon five different locations outside the building to hide imaginary objects. There were no restrictions on locations and communication methods. The participants were allowed to use gestures and talk to each other whenever needed to do so.

4.3 Observations

Participants used pointing gestures frequently throughout the study. In this section, I present five main findings relating to accuracy requirements, pointing gestures, communication richness, observer attention, and observer locations.

4.3.1 Accuracy Requirements

Participants' success in identifying the referent varied depending on the saliency and location of the referent. Observers were able to identify obvious referents quickly, but performed less well when referents were in a group, were hard to describe, or were partially occluded. When the referents were landmarks or obvious objects that were visible in the indicated direction, observers had no problem with identification. However, the difficulty of identifying referents dramatically increased when they were unobvious, such as pointing at a referent within a group of similar objects (e.g., any blue van in Figure 4.7 would be difficult to identify). In order to successfully identify referents, observers needed to understand the pointing gestures and then connect the gestures to the referents. Therefore, the varying difficulty of determining a referent

suggested that there are varying requirements for specificity in generating pointing gestures. I identified three canonical situations from the study that have different accuracy requirements.

Figure 4.7: A group of similar objects: a parking lot with many blue vans.

High accuracy requirement:

Pointing at an object in a group. The accuracy of pointing gestures appeared to be more important when an object is near or within a group of similar objects. Participants often had trouble identifying referents that were not obvious: for example, when someone pointed at a particular car in a full parking lot, it was difficult for observers to identify the referent. The problem arises both because of the density of the objects, and the difficulty of disambiguating the objects using speech. Observers would often have to guess at the referent within the cluster using a linear search. For example, when the referent was the car circled in Figure 4.8, one observer said "the black van in middle of the second row away from us", then "the black car next to it", and then "the next one". This linear-searching behaviour was seen most often when gesturers were not allowed to talk. When talking was allowed, gesturers would often give complex descriptions. For example, "it's a black car. Two cars to the right of the black van in middle of the second row away from us." These situations suggest that the more accurate a pointing gesture can be, the less verbal work will be required in these situations.

Figure 4.8: A parking lot full of cars.

Low accuracy requirements:

Pointing at distinct objects. Pointing accuracy was less critical when referents were distinct or easy to describe. Participants were able to easily identify referents such as a car that was the only vehicle in a parking lot; in these cases a general directional gesture and the phrase "the car" was usually enough for the observer to correctly identify the right referent. Even when speaking was not allowed, a general direction was sufficient if the referent was the only landmark in that direction (e.g., the flag at the top of the building in Figure 4.1). In these situations, pointing accuracy was not a major issue; participants needed only the general directions of the referents in order to successfully identify them through verbal cues.

Pointing at out-of-view referents. When referents were out of view, pointing accuracy also appeared to be less important. Participants tended to rely on the general directions for referents that they could not see from their view. For example, when gesturers wanted to indicate a parking lot that was behind other buildings, they might point at the direction of the parking lot and say, "somewhere over there." Also, when referring to something that was very far away and out of view, e.g., another city, accuracy was also less important. A general direction towards the far-away referent was appeared to be sufficient, but in these cases verbal communication was required.

4.3.2 Types of Pointing Gestures

Past research shows that people use a wide range of gestures in face-to-face settings (Bekker, Olson, & Olson, 1995; Tang, 1991). Similar behaviour was observed in this study. Participants used different types of pointing gestures for different referents. When the appearance of the referents can be easily depicted, participants might use one or both hands to illustrate the referents. For example, if the referent was the tip of a roof, the gesturer might generate a two-hand gesture forming a triangle, and extend the arms making the gesture in the direction of the roof (Figure 4.9). While these complex gestures were seen occasionally, the majority of the time when gestures were used, they were simple pointing gestures with a straight arm and an extended index finger. For example, when indicating a plainly-visible referent or general direction, the gesturer would raise the arm with an extended index finger as shown in Figure 4.6, and when showing an area, the gesturer would move the arm or wrist to create a circling gesture towards

the area. Sometimes different kinds of gestures were used for the same referent. For example, in Figure 4.10, the gesturer used a straight arm to point at a rooftop and the observer responded by using two hands to form a triangular shape towards it.

A  B 

Figure 4.9: A) a referent with a triangular shape; B) a gesture shaped like the referent.

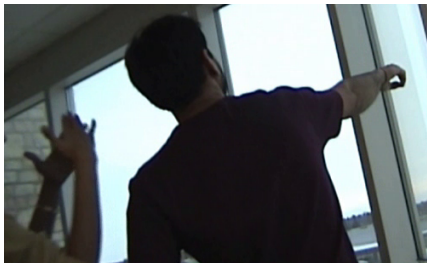


Figure 4.10: Participants used different gestures.

4.3.3 Communication Richness

There was a clear relationship between the richness of the communication channel and the type and complexity of pointing gestures. When communication channels were more restricted (written notes or no verbal communication), gestures were more detailed. For example, when only gestures were allowed, participants would form shapes with their hands more often. When communicating with gestures and written notes, participants needed to constantly switch between pointing and writing, which interrupted the flow of communication. When speech was allowed, gesturers used simpler gestures together with verbal descriptions and observers could identify referents much faster and easier.

4.3.4 Attentional Focus of Observers

Observers switched their attention depending on the stages and complexity of gestures and did not necessarily look directly at the gesturers or their pointing gestures. Before producing pointing gestures (i.e., during orientation and preparation stages), the participants would normally look at

each other, especially when speech was allowed. The attention of the observer was on the gesturer. When the gestures were produced, however, the observer would shift the focus to the referents and only saw the gestures in their peripheral view. Figure 4.11 shows an example of how the observer's attention was switched from the gesturer to the referent. In this example, the gesturer wanted to indicate a distant building. The observer was looking at the gesturer during orientation and preparation stages (Figure 4.11A and B), and switched the focus to the gesturer's arm when the pointing gesture was being produced (Figure 4.11C), and then changed the focus again to the building during the holding stage (Figure 4.11D). Although observers did not look at the gesture for most of the time, they were always able to determine the general direction of the pointing gesture. This happened more often when the pointing gestures were simple (raised arm and extended index finger). When the gestures were complex (such as forming the shape of the referent), observers appeared to pay much more attention to the gesture.

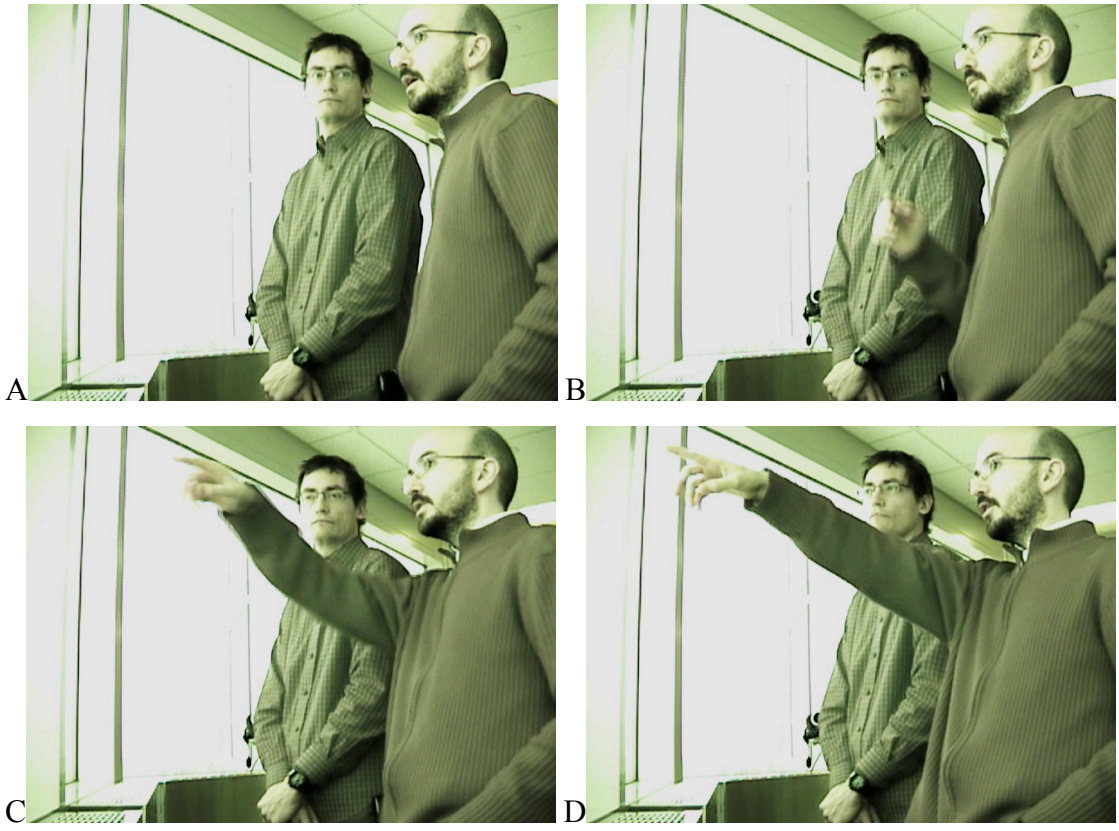


Figure 4.11: The observer changed his focus throughout the course of pointing: A) the observer was looking at the gesturer during the orientation stage; B) the observer was still looking at the gesturer when the gesturer was preparing to point; C) the observer switched his attention to the arm during the production of the gesture; D) the observer focused on the referent during the holding stage.

4.3.5 Locations of Observers

Observers changed their locations depending on how easy the referents could be identified. When referents were easily identified, observers would remain where they were standing. For example, in Figure 4.12, the observer would stay in position A, look at the gesturer, and then look at the referent (like the scenario described in the previous section). However, when observers had difficulty in finding the referents (e.g., cannot find the correct building after several attempts), the observers would move closer to the gesturers. There were several instances that observers even stood right behind the pointing arm to reduce parallax (Figure 4.12B and Figure 4.13).

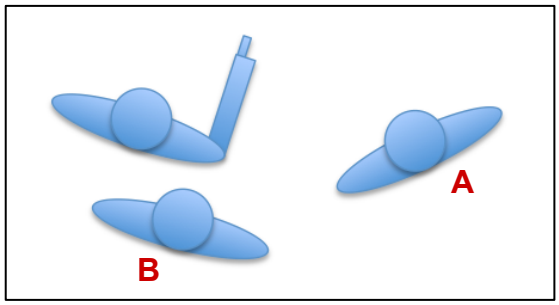


Figure 4.12: An observer stands at different locations: A) when referents can be easily identified; B) when referents are difficult to identify.

Figure 4.13: The observer stood behind the gesturer.

4.4 Discussion

In this observational study, I found that referents and communication richness have noticeable influences on pointing accuracy, type of gesture made, and observers' attention and position. In particular, I found five important aspects of distant pointing:

1. Pointing accuracy requirements differ depending on how obvious referents are;
2. Different gestures are used for different types of referents;
3. Communication richness affects how complex gestures are used;
4. Attention to gestures varies depending on the complexity of the gestures; and
5. Observers move to different locations based on how difficult it is to identify referents.

From these five points, I list lessons and implications that can be applied in CVEs:

1. ***Distant pointing has varying accuracy requirements.*** The accuracy required for different pointing gestures varied with the difficulty of the referential task. This means that designers can support different kinds of pointing with different mechanisms; for example, augmented pointing techniques (e.g., laser beams) may not be required for pointing at obvious referents or referents that cannot be seen.

   In Chapter 5, I will focus on pointing accuracy by investigating whether natural pointing (i.e., pointing without additional visual effects) is accurate enough to be used in desktop CVEs.

2. ***CVEs should support multiple types of pointing.*** People use different gestures during referential communication. For example, pointing with an extended arm is for indicating plainly-visible objects, circling gestures for general areas, and complex two-handed gestures for hard-to-describe objects. To allow this richness in CVEs, designers should provide much more expressivity than what is currently available.

   In the studies described in Chapters 5, 6, and 7, the avatar will be using free pointing (i.e., able to move the arm freely) for all referential activities. While the avatar cannot generate all the complex gestures (e.g., forming triangular shapes with two hands) as seen in this chapter, it can freely move the arm to point at objects, circle around areas, and draw along paths. The knowledge that will be gained from investigating free pointing can set the foundation for supporting more complex gestures in the future.

3. ***Speech is important for referential communication.*** The relationship between pointing complexity and communication richness suggests that in lower-richness CVEs (e.g., chat-based communication in environments like Second Life), the difficulty of constructing

referential statements puts the onus on pointing gestures to carry the reference. CVE designers should provide rich communication channels (e.g. enable voice chat) to make referential activities more effective.

In the study described in Chapter 7, I will use a CVE that allows collaborators to communicate with gestures and speech.

4. ***The importance of a wide field of view.*** Several situations in the study involved people focus on the referents instead of the pointing gestures, but clearly maintaining an awareness of the gesture in peripheral vision. CVEs with a wide field of view can allow users to maintain awareness of the gestures and focus on the referents at the same time.

   In Chapter 5, I will compare narrow and wide fields of view, and will investigate whether field of view affects pointing accuracy. In Chapter 7, I will explore two different ways to increase field of view by using three monitors and a third-person view.

5. ***Other avatar actions are also important.*** People need to move to different locations, change orientations, and face different directions during referential communication. While providing pointing support, CVE designers should keep other basic avatar actions available such as moving and turning.

   In Chapter 6, I will explore five different input mechanisms for controlling four avatar actions—pointing, moving, turning, and looking. In Chapter 7, I will conduct a study using avatars that are able perform all these actions.

CHAPTER 5
DETERMINING THE ACCURACY OF NATURAL POINTING IN CVES

Accuracy is a crucial element to the success of pointing. As discussed in the previous chapter, different referents have different pointing accuracy requirements in different situations. However, we do not know how accurately people can point and how accurately people can interpret pointing gestures in CVEs. Therefore, I conducted a study examining whether natural pointing (i.e., distant pointing that has no additional visual effects) is accurate enough to be used in desktop CVEs.

5.1 Setting

This study focused on the holding stage of distant pointing. During holding, the gesturer holds the gesture to ensure that the observer has seen it and has made a connection to the referent. This stage provides the most information to the observer when compared to other pointing stages (i.e., orientation, preparation, and production). Without the ability to convey direction in the holding stage, referential communication via pointing gestures cannot be successful. Therefore, I investigate how accurately people can point and interpret pointing gestures during the holding stage.

Natural pointing (pointing without additional visual aids) is the most basic way to point, and is the typical type of pointing people use in daily activities. Therefore, natural pointing was used in the study. If the study can show that natural pointing is accurate enough for referential communication in CVEs, it would suggest that other types of pointing that have added visual aids would also be accurate enough.

In order to determine whether natural pointing is accurate enough to be used in CVEs, I needed to compare pointing accuracy in the real world and in a CVE. If people can point in CVEs as

accurately as they can in the real world, pointing should have adequate accuracy in CVEs. Therefore, the study was held in two environments: the real world (RW) and a CVE.

In this chapter, I answer two main questions:

1. How accurately can people point at referents, in the RW and in a CVE?
2. How well can people determine the direction of a pointing gesture, in the RW and in a CVE?

In addition to the main questions, I also answer the following questions:

3. Does distance to the referent affect accuracy?
4. Does the observer's location affect interpretation?
5. Does field of view affect pointing?

5.2 Method

In this section, I describe the method of the study, including details of the participants, apparatus, experimental setup, conditions, procedure, and tasks.

5.2.1 Participants

Ten university students (6 male and 4 female) participated in the study. The mean age of the participants was 24. All participants were regular computer users, and four of them reported that they played 3D video games weekly.

5.2.2 Apparatus

The CVE used in the study was built using C# and XNA (see Figure 5.1). The CVE was set as a room with an avatar in it. Users were able to move the avatar's arm with a mouse. Moving the mouse up, down, left, and right pointed the arm in the corresponding directions. The system ran on a Windows XP PC with a Pentium 4 processor, and used a 22-inch LCD monitor with a 1680 x 1050-pixel resolution.

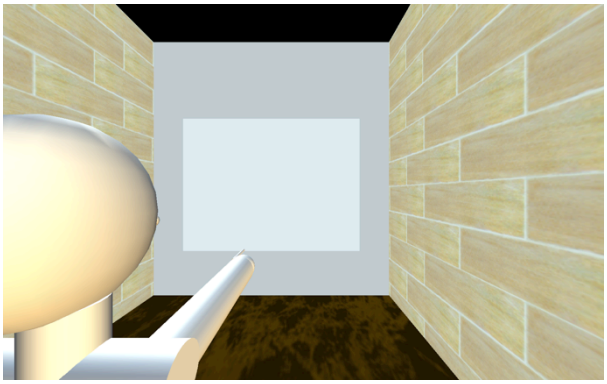


Figure 5.1: The CVE used in the study.

5.2.3 Experimental Setup

The study was conducted in two environments—RW and CVE. For the RW setting, the study was held in a 750cm x 400cm room. A projector with 1024 x 768 resolution was used to display referents on a 400cm-width wall. The projected area was 300cm x 225cm. The image was horizontally centred on the wall and 100cm above the ground (Figure 5.2). The locations where the participants and experimenter were standing are shown in Figure 5.3. For generation tasks, participants stood at location A (300cm from the wall) and B (600cm). For pointing-interpretation tasks, the experimenter (who had the role of a gesturer) stood at A and B, while participants stood at C, D, E, and F. Further explanation of the locations will be given in the following section.

Figure 5.2: The RW setup.

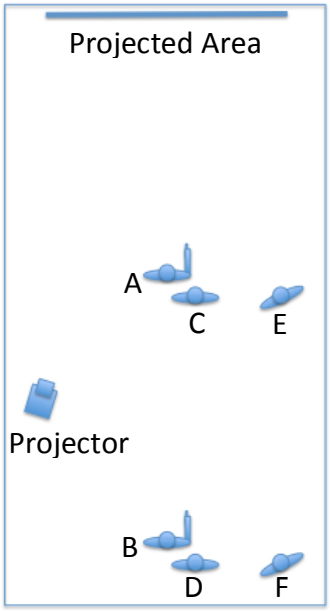


Figure 5.3: Top view of the experimental setup.

For the CVE setting, participants sat in a quiet room and did tasks in a CVE. The CVE replicated the real world setting: the virtual room was the same size as the real room, and participants placed their avatars in the same locations as in the real room. For gesture-generation tasks, participants used a mouse to control the avatar's arm movement. The avatar was at location A and B in Figure 5.3. For the pointing-interpretation task, the participant used the mouse to control the camera (the observer's view) at locations C, D, E, and F. The avatar in the role of the gesturer was located at A and B.

5.2.4 Conditions

The study had five factors: environments, task type, distance, observing location, and field of view (FoV); each factor had two levels.

Environment: *RW* and *CVE*. As explained earlier, I need to compare pointing accurate in the real world and in a CVE to determine whether natural pointing has adequate accuracy in CVEs.

Task type: *generation* and *interpretation* of pointing gestures. Pointing is a communicational act that involves someone generating a gesture and someone else interpreting the gesture. To examine the accuracy of natural pointing, I need to know how accurately people can generate and interpret pointing gestures.

Participants stood at location A and B in Figure 5.3 for *generation* task; at C, D, E, and F for *interpretation* task. Figure 5.4 shows the two task types in the RW and CVE settings.

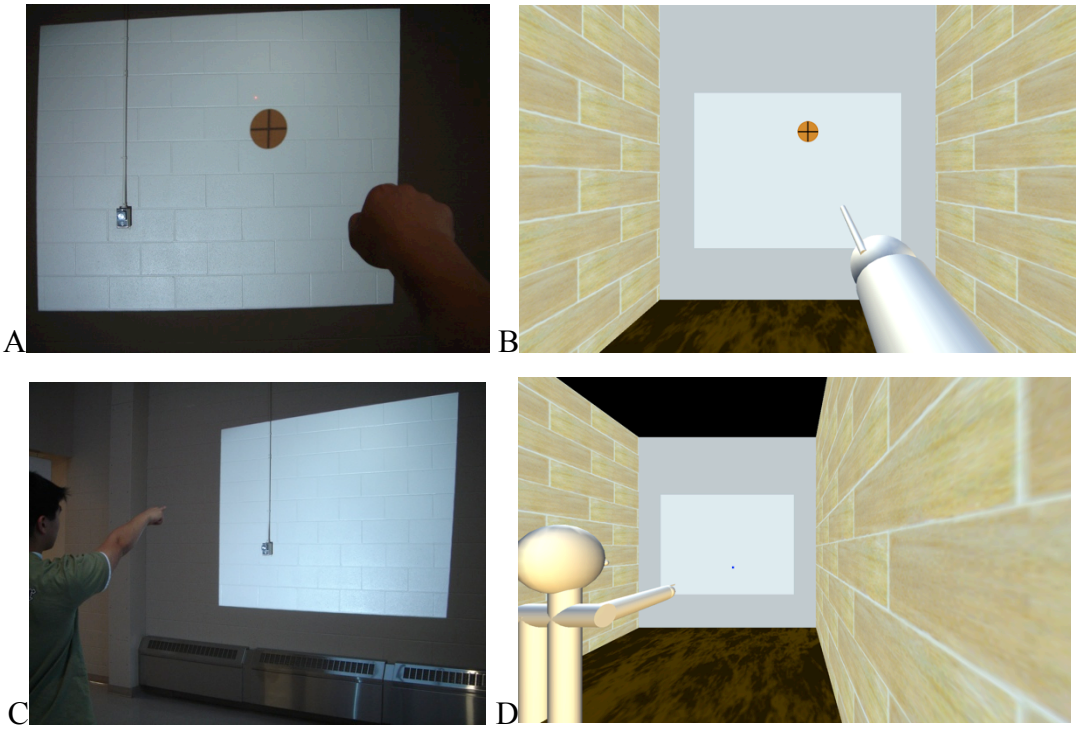


Figure 5.4: From participants' view: generation (A and B); interpretation (C and D).

Distance: *300cm (near)* and *600cm (far)* to the referents. How far referents are can affect how people communicate via pointing gestures (see Chapter 4). For example, far-away objects can be harder to point at than close ones. However, due to the difficulty in measuring accuracy for very far-away objects (e.g., the buildings in the cityscape in Chapter 4) and the limitation of the experimental room, I only compared pointing accuracy for referents that are 300cm and 600cm away.

Participants stood at location A, C, and E in Figure 5.3 for the *near* condition; at B, D, and F for *far* condition. Figure 5.5 shows the *near* and *far* conditions in the RW and CVE settings.

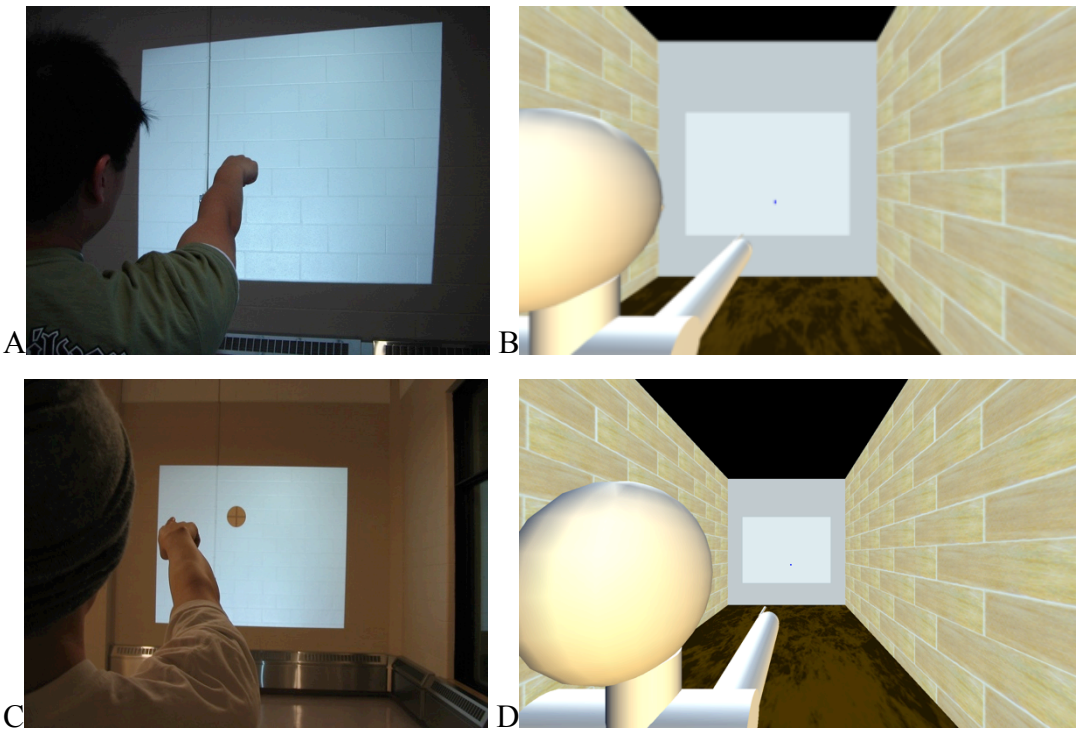


Figure 5.5: Distances to referents: near (A and B); far (C and D).

Observing location: *behind* and *beside* the gesturer. As described in Chapter 4, observers move to different location relative to the gesturers depending on the difficulty in identifying the referents. Standing behind (for hard-to-identify referents) and beside (for easy-to-identify referents) the gesturer are two typical observing locations. Note that this condition was only used for interpretation tasks because generation tasks did not have observers.

Participants stood at location C and D in Figure 5.3 for the *behind* condition; at E and F for the *beside* condition. Figure 5.6 shows the differences between *behind* and *beside* in the RW and CVE settings. The two locations were chosen to ensure that both the referents and the gesturer's arm will be able to be seen by the observers in a single view, so that the participants will not need to establish mutual orientation by themselves.

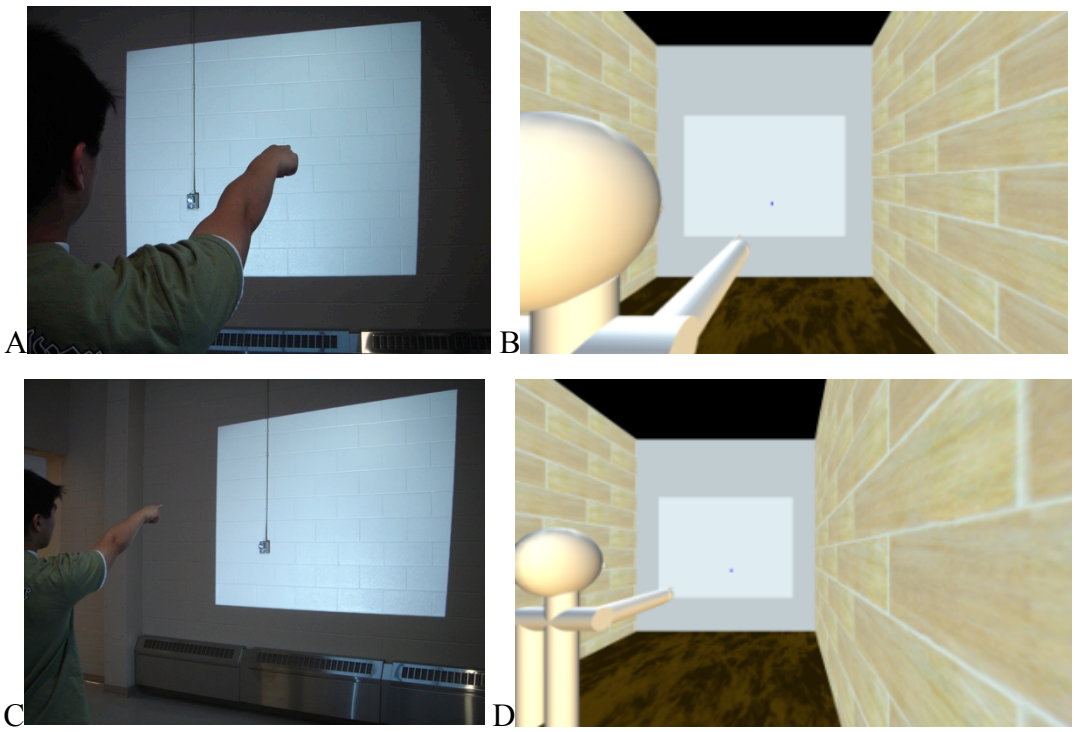


Figure 5.6: Observer views: behind (A and B); beside (C and D).

Field of view: *85° (small)* and *120° (large)*. As discussed in Chapter 4, it is important that CVEs have a wide field of view. Therefore, I compared two different fields of view to determine how field of view affect pointing accuracy. This factor was only used for the generation tasks in the CVE. Interpretation tasks in the CVE had an 85° field of view. Figure 5.7 shows the difference between small and large fields of view. Note that the monitor size did not change, so the widen field of view resulted in a compressed image.

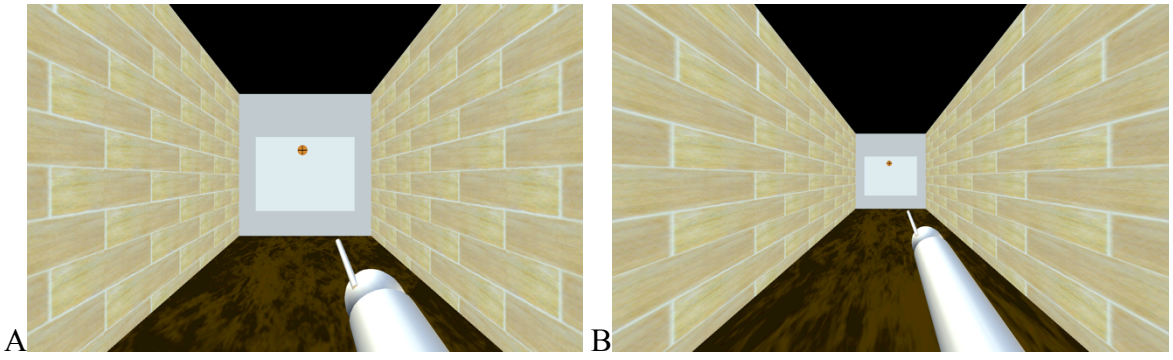


Figure 5.7: Different fields of view: A) small (85°); B) large (120°)

There were 15 trials per condition. The first five trials of each condition were marked as training. Table 5.1 summarizes the conditions used in the study. The study was a within-participants design and condition order was counterbalanced using a Latin square design.

Table 5.1: Number of trials in each experimental condition.

		CVE		RW	
		Near	Far	Near	Far
Gesture Generation	Small FoV	10	10	10	10
	Large FoV	10	10		
Gesture Interpretation	Behind	10	10	10	10
	Beside	10	10	10	10

5.2.5 Procedure

At the beginning of each session, the participant was informed of the purpose of the study, signed a consent form, and filled out a demographic survey. Then, the participant did the tasks and filled out a post-test questionnaire. Finally, the participant was debriefed and was given $10.00 as remuneration. The study took about one hour to complete.

5.2.6 Tasks

The study had two types of task: gesture *generation*, in which participants were asked to point as accurately as possible at a given referent; and gesture *interpretation*, in which they were asked to determine the direction of another person's pointing gesture.

Generation of pointing gestures. Participants were asked to point at the centre of referents that appeared on the wall in front of them. Referents appeared at random locations, one at a time. In the RW, participants pointed with a laser pointer; in the CVE, the participants controlled their avatar's arm with the mouse (Figure 5.8).

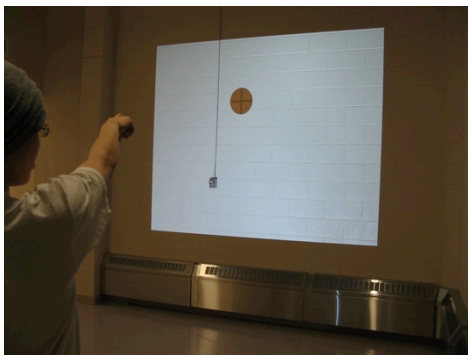
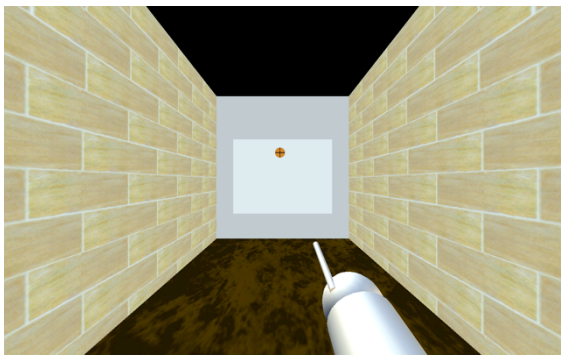
A  B 

Figure 5.8: Generation task: A) RW; B) CVE.

In the RW, participants were first given a laser pointer and asked to practice with it until they could point comfortably and consistently. Then they stood at the required position in the room, and pointed at referents with their arm straight. Participants were told not to turn the laser pointer on until they were confident that it was aiming at the referent. When the laser was switched on, the experimenter recorded the location of the laser dot.

In the CVE, a similar procedure was followed, except that participants used the mouse to control the avatar's arm direction, and clicked the mouse button to complete each trial. On each mouse click, a red dot appeared on the virtual wall to provide the same feedback about where the user had pointed as was given in the real world.

Interpretation of the direction of pointing gestures. For this task, participants were asked to observe a gesturer (the experimenter) who pointed at locations on the front wall, and then determined what location was being pointed at. Figure 5.9 shows the interpretation task in the RW and CVE settings.

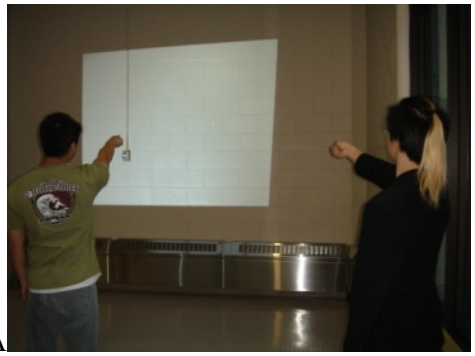
A  B 

Figure 5.9: Interpretation task: A) RW; B) CVE.

In the RW version of the task, the participant first turned away, and then a referent appeared on the wall. The gesturer produced and held a straight-arm pointing gesture towards the referent. When the pointing gesture was ready, the referent was hidden and the participant turned around. Then, the participant used a laser pointer (on at all times) to indicate where on the wall he/she thought the gesturer was pointing; this location was recorded by the experimenter.

The CVE version of the task was equivalent, but adapted to the desktop setting similar to the description of the generation task. The gesturer was a computer-controlled avatar that pointed at random invisible referents on the wall. The participant did not need to turn away from the gesturer because the referents were not shown on the screen. Instead of using a laser pointer, the participant used the arrow keys on a keyboard to control a small dot to indicate where the referents were.

5.3 Results

All referent locations and the locations where the participants pointed were recorded. Using these data and height of participant shoulders (measured and recorded before the tasks began), I calculated the angular error of each task (i.e., the difference in angle between imaginary lines drawn from the gesturer's shoulder to the referent, and from the shoulder to the participants' recorded location). Angular error is used as the measure of performance, rather than absolute error, because it is not affected by distance from the referents, and thus results can be comparable across different distances. Results are organized below based on the five research questions specified earlier.

5.3.1 How Accurately Can People Point at Referents, in the RW and in a CVE?

Using the data from the generation task, except the data with a large field of view for balancing the data points between environments, the mean angular error was 3.1°. Analysis of variance (ANOVA) showed a main effect of environment ($F_{1,9} = 31.56$, $p < 0.001$), with RW pointing (mean 1.8°) significantly more accurate than pointing in the CVE (mean 4.4°) (see Figure 5.10).

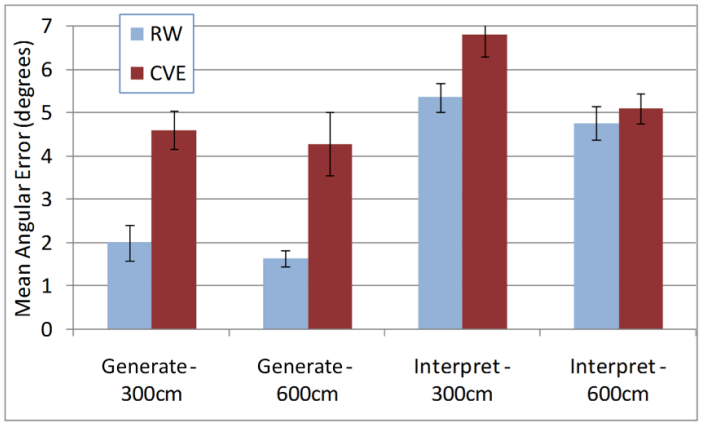


Figure 5.10: Mean error by environment, task, and distance (error bars represent standard error).

5.3.2 How Well Can People Determine the Direction of a Pointing Gesture, in the RW and in a CVE?

Using all conditions in the gesture-interpretation task, the mean angular error was 5.5°. ANOVA showed a significant main effect of environment ($F_{1,9} = 7.04$, $p < 0.05$), with errors in RW (mean 5.1°) less than in the CVE (5.9°).

There was also a significant interaction between environment and distance ($F_{1,9} = 7.38$, $p < 0.05$); as shown in Figure 5.10, the difference between the RW and CVE was much more pronounced at 300cm than at 600cm. There was no interaction between environment and observation location ($F_{1,9} = 3.04$, $p = 0.12$). In addition, generation of pointing gestures overall was more accurate than interpretation; ANOVA showed a main effect of task ($F_{1,9} = 169.47$, $p < 0.001$).

5.3.3 Does Distance to the Referent Affect Accuracy?

An ANOVA using data from both tasks (data of large field of view and standing beside conditions were excluded for balancing) showed a main effect of distance ($F_{1,9} = 12.31$, $p < 0.01$). However, the ordering of the two conditions was surprising: when standing 300cm from the referent, error was always more than when standing at 600cm (see Figure 5.10 and Figure 5.11). Although the difference was unexpected and small (less than 1°), the result was significant (see Section 5.4.2 for the explanation of the result). As reported above, there was also a significant interaction between distance and environment (with distance having more of an impact on interpreting pointing in the CVE).

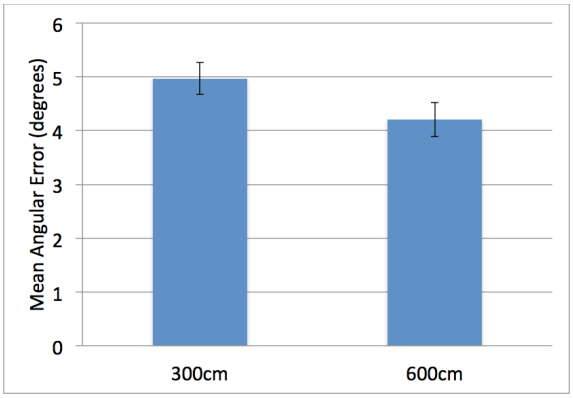


Figure 5.11: Mean error by distance (error bars represent standard error).

5.3.4 Does the Observer's Location Affect Interpretation?

An ANOVA showed a significant main effect of location on interpretation accuracy ($F_{1,9}$ = 14.32, $p < 0.01$). When observers stood behind the gesturer, error was less (mean 4.9°) than when standing beside (6.1°) (see Figure 5.12). No interaction was found between location and environment ($F_{1,9} = 3.04$, $p = 0.12$).

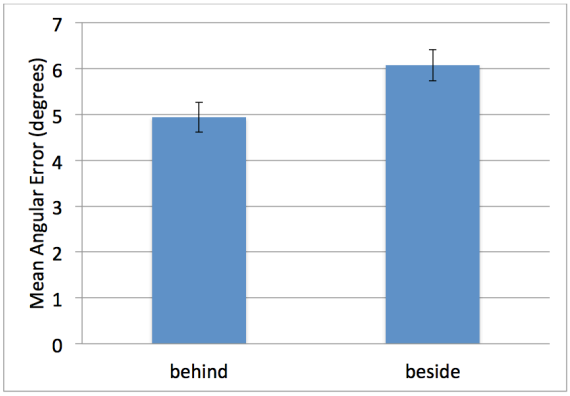


Figure 5.12: Mean error by observer's location (error bars represent standard error).

5.3.5 Does Field of View Affect Pointing?

ANOVA on the gesture-generation task did not show a significant main effect for field of view ($F_{1,9}$ = 1.53, p = 0.25), and no interaction between field of view and distance ($F_{1,9}$ = 1.84, p = 0.21). The mean error of the 85° view was 4.18°, and of the 120° view was 4.71° (see Figure 5.13).

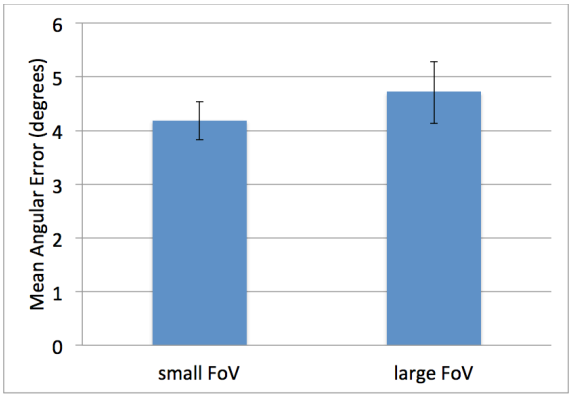


Figure 5.13: Mean error by field of view (error bars represent standard error).

5.3.6 Questionnaire Responses

All participants reported having more confidence in doing both tasks in RW as compared to the CVE; participants were also unanimous in stating that the tasks were more difficult in the CVE. Most participants (7 of 10) reported having more confidence when observing from behind the gesturer as opposed to beside.

5.4 Discussion

The main findings from the study organized by research questions are as follows:

1. *How accurately can people point at referents, in the RW and in a CVE?*
   Participants could generate pointing gestures more accurately in the real world than in the CVE.

2. *How well can people determine the direction of a pointing gesture, in the RW and in a CVE?*
   Participants could determine pointing directions more accurately in the real world than in the CVE. The difference between the environments for interpreting pointing direction was much smaller than expected—only 1.4° at 300cm, and only 0.33° at 600cm.

3. *Does distance to the referent affect accuracy?*
   Errors were larger (by approximately one degree) when people were nearer to the referent.

4. *Does the observer's location affect interpretation?*

   Observers were more accurate when interpreting a pointing gesture from behind than from beside (a difference of 1.13°).

5. *Does field of view affect pointing?*

   The different fields of view available in the CVE made little difference in generating pointing gestures.

5.4.1 Differences between RW and CVE

Although the differences between the real world and the CVE were significant (with people performing better in the real world), the most interesting and surprising feature of the study results is that the actual differences between the two environments are relatively small. To put the differences into real-world terms, Figure 5.14 and Figure 5.15 compare the error from the two environments, for the interpretation task. At 600cm from the referent, people would be able to differentiate objects that are 50cm apart in the real world, but in a CVE, referent objects would have to be 53.5cm apart. For example, people would not be able to differentiate the two crosses in Figure 5.15A in both the real world and CVEs. In Figure 5.15B, the crosses can be differentiated in the real world, but not in CVEs. In Figure 5.15C, the crosses can be differentiated in both environments.

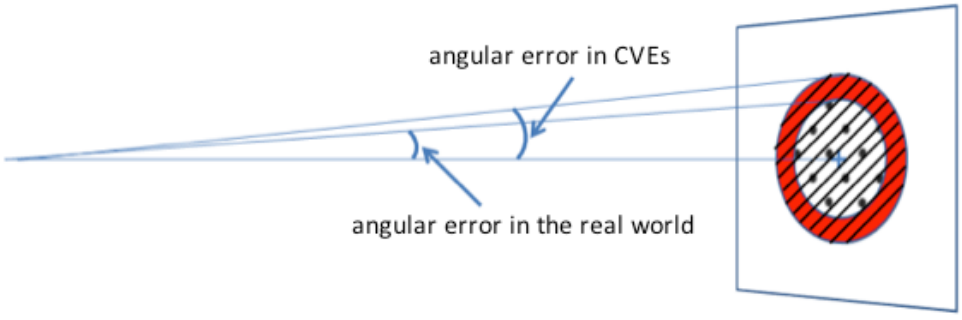


Figure 5.14: Comparison of error zones in RW and CVE (red area is the difference).

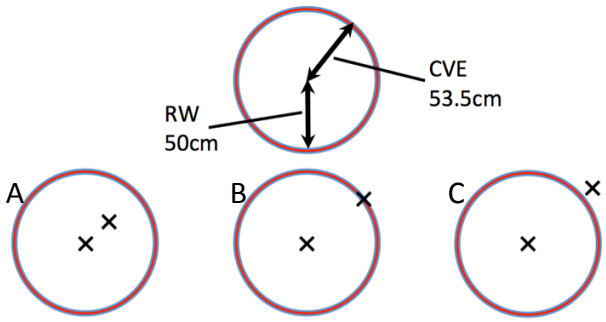


Figure 5.15: At 600 cm from the referents: A) the crosses cannot be differentiated in the real world and CVEs; B) the crosses can be differentiated in the real world, but not in CVEs; C) the crosses can be differentiated in both environments.

There are three possible reasons for the differences between the real world and CVEs:

Avatar control. People found it much easier to point in the real world than to control the avatar's arm in the CVE. One participant commented "[in the] real world, I just found I have more control over what I was doing and felt more confident in doing it." Although participants did not have any major difficulties with the input techniques, controlling the avatar in the CVE was definitely more difficult than moving one's own arm.

Size and resolution of the view. People receive much more information from their view of the real world than from a CVE displayed on a regular desktop monitor. In the real world, people have a view close to 180°, but had a much smaller view in our experimental CVE (a 22-inch screen). Furthermore, the visual resolutions of CVEs was dramatically less than the real world, and many details that could be seen in the real world may have been lost in the CVE.

Sense of distance and depth. Participants' comments indicate that the sense of depth in the CVE was not as strong as in the real world: for example, participants stated that "the real world was easier; it's easier to determine the distances," and "the distances are not as real compared to the real world… in the virtual world, I couldn't feel the distance difference." There were two potential differences in depth perception between the environments. First, whereas head movement (and the associated parallax that results from these minor view shifts) was a natural part of the real-world environment, the tasks in the CVE did not involve movement of the view. Second, some of the distance cues available in the real world (such as stereo vision) were not available in the CVE, reducing perception of depth.

5.4.2 Distance from Referents

It was surprising that people were more accurate when the referent was farther away. There are two possible explanations for this result. First, participants were asked to aim at the centre point of each referent, and distance does not affect the size of a point. Therefore, nearer referents do not have the advantage of appearing bigger. Second, parallax and the distance between the gesturer's eye and their shoulder may have an effect on accuracy. Near objects have larger parallax than faraway objects for both pointing and observing. The difference in angle between the eye-to-referent line and the shoulder-to-referent line becomes smaller as the gesturer moves farther from the referent.

5.4.3 Observer's Location

People were better at interpreting others' pointing gestures from behind rather than from beside. This is likely due to the fact that the view from behind the gesturer is more similar to the gesturer's view. The behind view was closer to the gesturer's shoulder, suggesting that the parallax issue described above may account for some of the difference between the two observer locations.

5.4.4 Field of View

The different fields of view available in the CVE made little difference in generating pointing gestures. It is possibly because all the necessary visual elements to carry out the pointing task (i.e., the gesturer's arm and the referent) were shown in both views at all times. This means that the view was not fragmented (Hindmarsh et al., 1998), and therefore the wider field of view made little difference. Field of view is clearly an important factor, however; if important elements are not in view, the pointing gesture cannot be either created or interpreted. Another possible reason is that both views were displayed on the same 22-inch monitor, which means that the larger field of view was compressed. Therefore, the benefit of having a larger field of view was not detected.

5.5 Lessons

There are three main lessons learned from the study.

1. ***Natural pointing in CVEs can be successful.*** The results show that people can interpret others' pointing gestures in a CVE almost as well as they can in the real world. Given that many types of pointing gesture do not require high accuracy, the results strongly suggest that naturalistic deictic reference without any additional visual effects can be used to a much greater degree than has been seen in current CVEs. In particular, both general directional pointing and more specific pointing where the referent is relatively easy to disambiguate through speech (as discussed in Chapter 4), should be possible in CVEs using natural pointing. However, these results are limited to situations where mutual orientation has already been established. I will address this limitation in Chapter 7. In the study of Chapter 7, participants will need to establish mutual orientation to do the tasks.

2. ***Pointing in CVEs is still less accurate than RW.*** Although people were more accurate overall for generation than for interpretation, generating gestures in the CVE was considerably less accurate than in the real world. There are several ways in which people could be assisted in pointing, e.g., put objects in the scene to help people estimate distance when generating a pointing gesture and use augmented-pointing techniques to provide visual aids. The studies in Chapter 6 and 7 will have more objects in the scene; the study in Chapter 7 will have a more realistic scene and the avatars will be able to use augmented pointing.

3. ***Compressed field of view does not aid accuracy.*** The process of mutual orientation, the ability to see both gesture and referent, peripheral awareness of gestures, and the visibility of pointing actions are all made difficult or impossible by restricted fields of view (Fraser et al., 1999). However, the results show that accuracy cannot be supported simply by compressing a larger view. Multi-monitor displays are a more likely solution to this problem, as they provide an uncompressed field of view increase. I will use multi-monitor displays in the study of Chapter 7.

CHAPTER 6
CONTROLLING AN AVATAR'S POINTING GESTURES

Pointing in common CVEs has limited expressiveness. To improve pointing expressiveness, I added free pointing—where the pointing direction is independent from other actions of the gesturer— to existing avatar actions (i.e., moving, turning, and looking). Users' hands, however, are already occupied for manipulating these basic actions. Also, existing input configurations for avatars may not work because of the extra pointing control. Therefore, I configured five commonly-available input devices so that they can each be used to control pointing, moving, turning, and looking. I then conducted a study to compare these configurations and answer the following two research questions:

1. Can people control free pointing while controlling other avatar actions?
2. What input device is the best for controlling free pointing along with other avatar actions?

6.1 Control of Deictic Pointing in Current CVEs

The expressiveness of deictic pointing in CVEs is limited compared to the real world. In this section, I provide examples of how pointing is performed in some common CVEs and describe the corresponding problems.

In World of Warcraft, pointing is performed by a text command (i.e., "/point"). The avatar can only point straight to the front, which is a preset and fixed animation. Once the command is executed, the user cannot change the pointing direction, pause the gesture, or control the speed. It is difficult for users to synchronize pointing gestures with speech. In addition, the user can click on an object and then execute the pointing command. The avatar will then generate the pointing gesture, and the game will register that the object is being pointed at. However, because the avatar can only point to the front, what is registered by the game may be different from what is visually shown in the virtual world. For example, in Figure 6.1A, the avatar on the right is visually pointing at an entrance. When other people see this pointing gesture, they would think

the referent is the entrance. However, this is different from what the system sees. The system registers that Orgrimmar Grunt (the avatar on the left) is the referent (see Figure 6.1B). This mismatch of information can cause confusions between players.

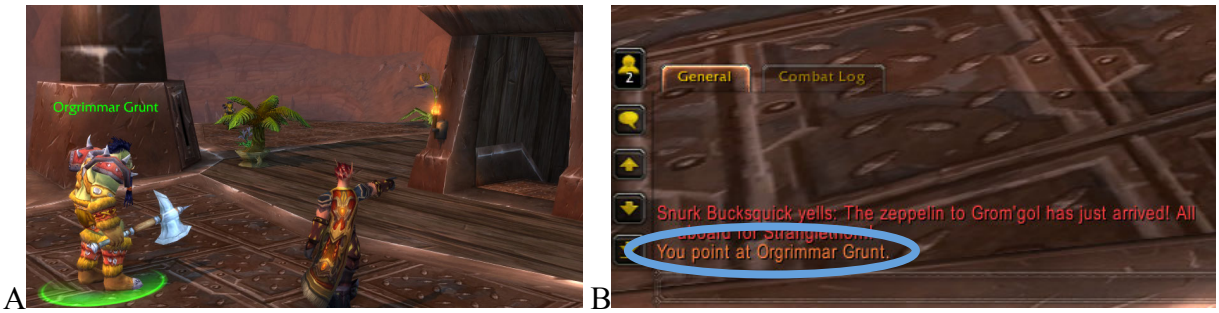


Figure 6.1: How a pointing gesture is seen in World of Warcraft by A) other players; and B) the game.

In Second Life, pointing gestures can be generated by right-clicking on objects. The objects will be highlighted and the avatar will point at the objects (Figure 6.2). However, highlightable objects are predefined by CVE designers. Users have no control of what can be highlighted. They may want to point at a single part of the whole highlighted object or point at the bigger object that contains the highlighted part. For example, in Figure 6.2, the user is not able to point at only one of the chairs because the system considers all the chairs and the table as one single highlightable unit. In addition to the granularity problem of highlightable objects, coupling pointing gestures to only objects is another problem. Users may want to point at empty space between objects, general directions, and paths between objects. However, these referents cannot be pointed at using this object-based pointing method.

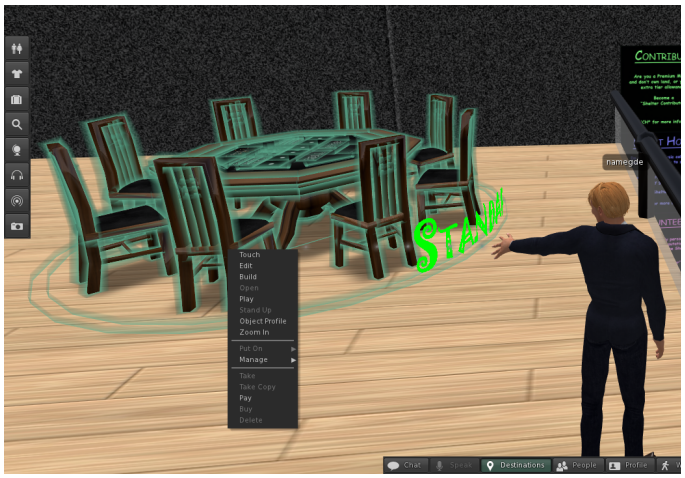


Figure 6.2: Pointing in Second Life.

In PlayStation Home, pointing is performed by navigating through menus (Figure 6.3). Users need to bring up the menu, choose "Conversation", and then select "Point". The avatar will then point to the front once the command is selected. One problem with this activation method is that it takes a long time to go through the menu, making pointing gestures difficult to synchronize with conversations. Another problem is that it is difficult to remember the command sequence. For example, in Figure 6.3, users need to select "Conversation" and then "Point". If users forget the correct sequence and think that "Point" command is under "Actions", they would select "Actions" instead of "Conversation" and unable to find the correct command.

Figure 6.3: Pointing by navigating through menus in PlayStation Home.

In many first-person shooter games, avatars can only point at the centre of the view and pointing is performed by controlling the avatar's view. Tying pointing direction to viewing direction is a problem because users may lose sight of important areas when pointing. For example, Bob needs to pay attention to the four exits on the sides and the box at the bottom (in Figure 6.4A), and also needs to tell a teammate to go to the top left exit. If Bob wants to point at the exit and say "go over there", he will need to change the view so that the top left exit is located at the centre of his screen (Figure 6.4B). He will then lose sight of other areas that he needs to pay attention to.

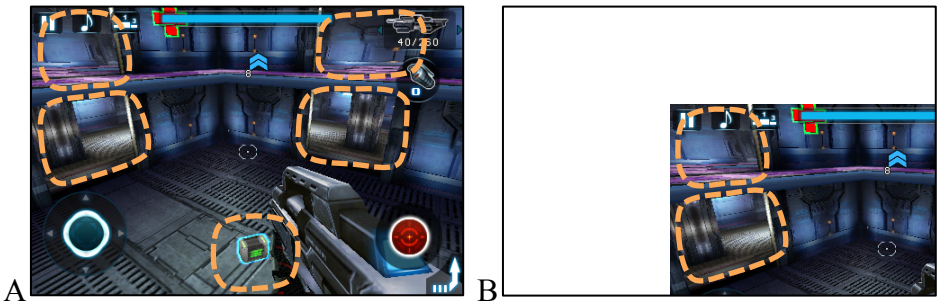


Figure 6.4: The screen of an FPS game: A) showing all important areas; B) showing only two areas.

6.2 Avatar Actions

To make gestural communication more effective in CVEs, it is important to solve the pointing problems mentioned above by improving pointing expressiveness. Therefore, I added a hand-and-arm based pointing control to basic avatar actions. Most avatars have three basic actions: moving, turning, and looking; and free pointing is added as the fourth action. Figure 6.5 shows the four actions. All actions can be controlled simultaneously, enabling the notion of parallel structure for object manipulation as discussed by Wang, MacKenzie, Summers, & Booth (1998). They suggested that "if the main goal of interface design is to achieve the 'naturalness' or realism such as virtual reality, remaining the natural structure of human object manipulation [that is being able to simultaneously control different avatar actions] will be particularly important." (Wang et al., 1998, p. 319)

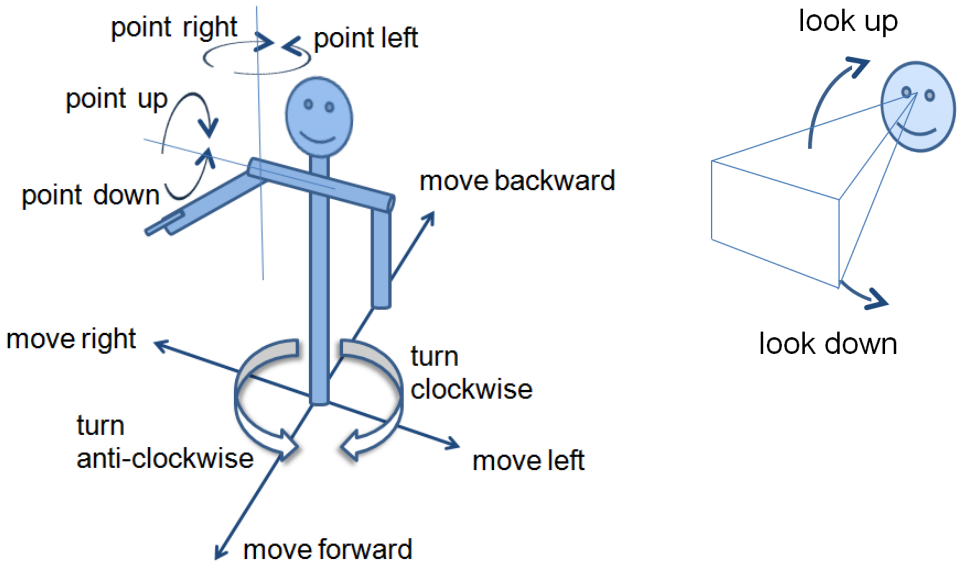


Figure 6.5: Four different actions of avatar: moving, turning, looking, and pointing.

Moving (body location). Moving an avatar involves translating its position on the surface of the CVE world (as would occur if the avatar walked sideways, forward, or backwards). Avatar movement also alters the avatar's view and pointing gesture, since the eyes and arm are moved along with the body. In many CVEs, avatar movement is accomplished using a four-directional keypad (e.g., the A-S-D-W keys on a keyboard).

Turning (body direction). Turning an avatar means rotating it around its vertical axis. Turning does not change the avatar's location, but does change the view and the direction of pointing. In addition, turning control is often used in concert with keypad-based movement in order to allow precise translation: that is, by turning as the avatar moves forward, better control over movement can be achieved. In many CVEs, turning is the only way to change the avatar's view left or right, as there is no separate control over horizontal view direction.

Looking (head direction). Looking involves changing an avatar's view, i.e., the direction of the view frustum originating from the avatar's eye position. Changing the avatar's view also changes what the user can see on the screen if a first-person view is used. Looking does not affect the avatar's location, rotation, and pointing direction. Many CVEs only provide dedicated control over up-down looking, with left-right looking tied to turning.

Pointing (arm direction). Pointing means extending the avatar's arm to indicate a particular direction relative to the avatar. I used a straight-arm gesture as the pointing action: the arm is held straight with a finger extended, and can be rotated horizontally and vertically at the avatar's shoulder. Free pointing does not change the avatar's location, rotation, or view. This type of free pointing is not supported in any current CVE.

6.3 Input Configurations

Input devices used in CVEs are not explicitly designed to control pointing together with other actions. To test if free pointing can be successfully controlled, I configured five widely-available input devices to allow control of all four types of avatar action (Figure 6.6 to Figure 6.10). These input devices are all commonly available; however, the input mappings are in some cases different from their conventional settings.

Each device combination has many possible mappings (e.g., the left thumbstick of gamepad can control looking while the right thumbstick can control pointing, or vice versa), and each mapping can have many different sensitivities. I performed a series of pilot tests to find out the best input mappings and sensitivities for all device combinations. Also, in all configurations, I combined looking and turning as described in the previous section. Because turning also changes the view, people can still look at all directions through the avatar's view.

Mouse and Keyboard. These devices are the most common configuration for controlling avatars on PCs. I retained the conventional mappings of the mouse and keyboard, but added a mode switch to control pointing: when the mouse button is pressed, the mouse controls pointing; when the button is up, the mouse controls rotation and view as normal. When pointing, forward and backward movement of the mouse moves the pointing arm up and down, and left and right mouse movement moves the arm left and right (see Figure 6.6). This configuration has the restriction that view control and pointing cannot be done at the same time.

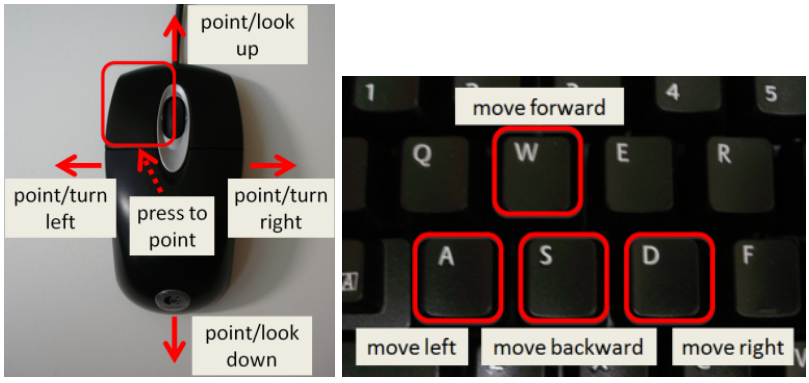


Figure 6.6: Mouse and keyboard.

Trackball, Mouse, and Keyboard. This combination is similar to the mouse-and-keyboard configuration, except that in place of a mode switch, an additional 2D input device (a trackball) was added to the top of the mouse in the normal mousewheel location (I used an Apple Mighty Mouse). Pointing can be controlled at any time by the trackball: rotating the ball forward, backward, left, or right changes the pointing direction up, down, left, and right (Figure 6.7). Other input mappings for the mouse and keyboard are as above. This configuration is similar to the mouse but allows simultaneous control of pointing.

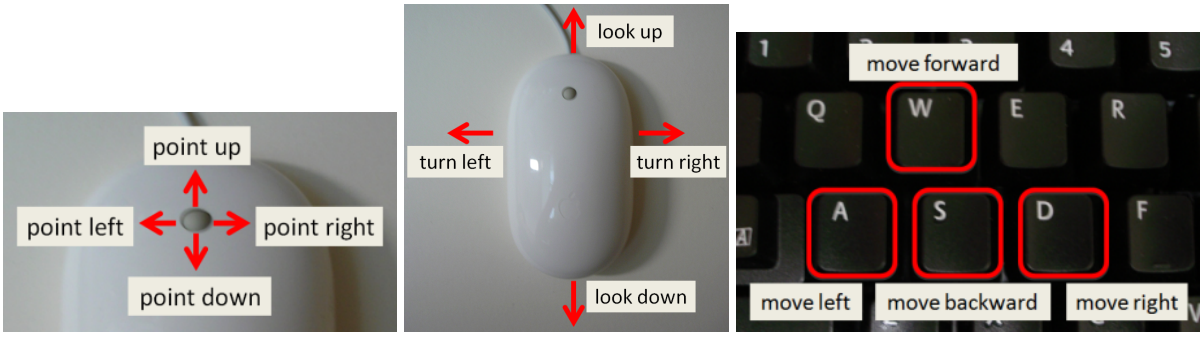


Figure 6.7: Trackball, mouse, and keyboard.

Gamepad. A gamepad is the primary input device for CVEs on game console systems. I used an Xbox controller with two thumbsticks and a directional control pad (d-pad). The right stick was used for pointing, the left stick for looking and turning, and the d-pad for moving (Figure 6.8). I used standard directional mappings of the thumbsticks and d-pad for pointing, looking, turning, and moving.

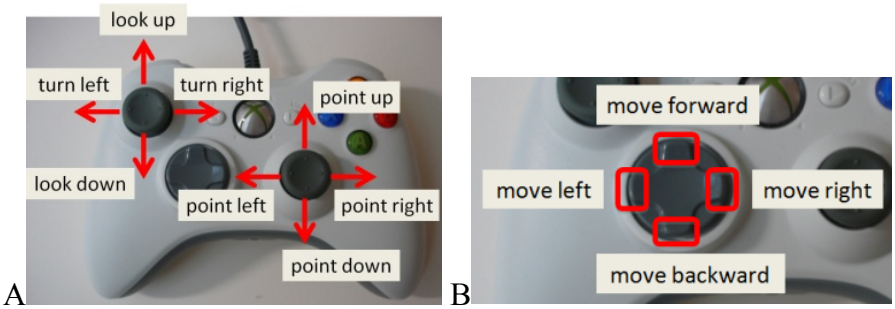


Figure 6.8: Gamepad: A) thumbsticks; B) d-pad.

Joystick. Many game joysticks allow control over multiple dimensions with a single device; I used a Microsoft SideWinder Precision 2 joystick to control all avatar actions (see Figure 6.9). The main stick (forward, back, left, right) is for pointing, and the 'hat' at the tip of the stick is for looking and turning. I used the four buttons on the base of the joystick (left of the main stick) to control moving.

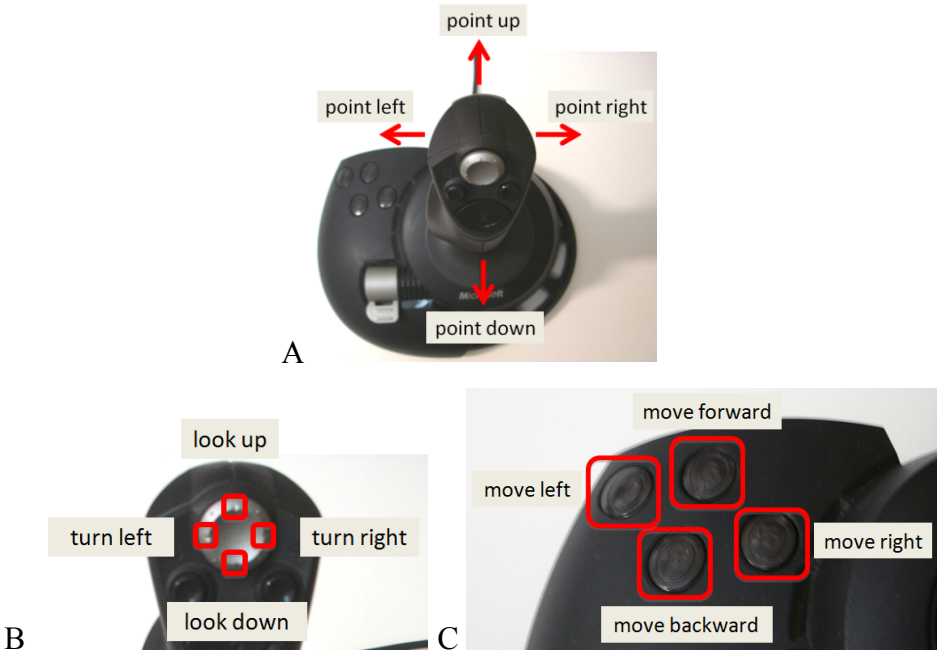


Figure 6.9: Joystick: A) main stick; B) hat; and C) buttons.

Wii Controls. The Nintendo Wiimote allows direct pointing at the screen, letting people point as they would in the real world. Pointing direction was controlled by the user's (real) arm: to point the avatar's arm in a certain direction, they moved the Wiimote in the corresponding direction. The thumbstick on the Nunchuk controller was used to control turning and looking (as with the thumbstick on the gamepad). The Wii Balance Board controlled the movement of the avatar. When seated, the user pressed on different parts of the board to move forward, back, left, and right (see Figure 6.10).

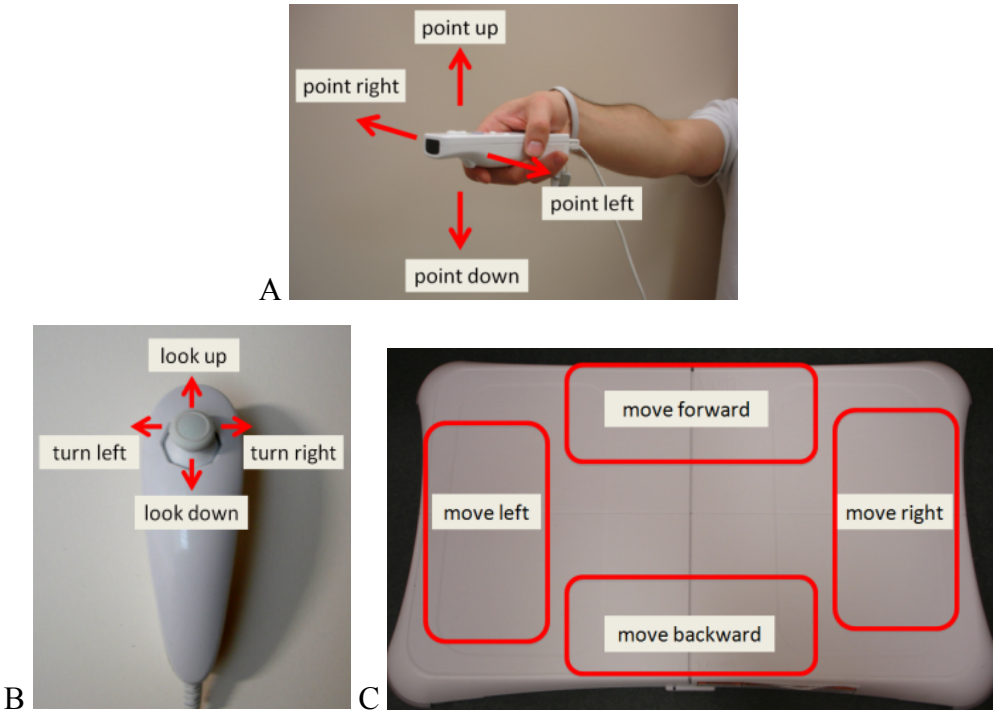


Figure 6.10: Wii controls: A) Wiimote, B) Nunchuk, C) Wii Balance Board.

6.4 Degrees of Freedom for Controls and Inputs

Moving, looking, and pointing involve continuous two-dimensional control, whereas turning involves continuous control over one dimension only. I used the same simplified controls as many game environments by reducing movement to simple four-direction control (forward, back, left, right), and by locking horizontal view direction to avatar rotation direction. Free pointing adds another two-dimensional space in which pointing direction must be specified: that is, vertical and horizontal rotation of the straight arm around the shoulder. The table below shows the degrees of freedom of each control and the corresponding devices.

Table 6.1: Degrees of freedom for controls and input devices.

	Move	Turn	Look	Point
	2 translations: - left/right - forward/backward	1 rotation: - left/right	1 rotation: - up/down	2 rotations: - up/down - left/right
Mouse + Keyboard	Keyboard	Mouse	Mouse	Mouse (button pressed)
Trackball + Mouse + Keyboard	Keyboard	Mouse	Mouse	Trackball
Gamepad	D-pad	Left Thumbstick	Left Thumbstick	Right Thumbstick
Joystick	Buttons	Hat	Hat	Main Stick
Wii Controls	Balance Board	Nunchuk's Thumbstick	Nunchuk's Thumbstick	Wiimote

6.5 Properties of Input Devices

There are different properties of input devices that affect how they are used for controlling free pointing (Hinckley, 2003; Jacob, 1996):

Absolute vs. Relative. The mapping of the input space to the virtual space can be either *absolute*, where each point on the input space corresponds to a point in virtual space; or *relative*, which allows more flexible movement but often requires clutching. For example, a mouse or trackball provides control over the position of the controlled object. When the mouse reaches the edge of a mouse pad or the trackball rolls to the edge of the hand, a repositioning of the mouse and the hand is required.

Direct vs. Indirect. The directness of the device is also an issue: *direct* input implies that the input space is the same as the output space (e.g., touch screens or Wii remotes); *indirect* devices use a separate input space (e.g., the mouse and keyboard).

Position vs. Rate control. The translation of raw device input to movement of an object is the device's transfer function, and can be zero-order (i.e., position control), first-order (i.e., rate control), or higher-order (e.g., acceleration control). In general, only position and rate control are used in human motor control tasks. For example, a mouse provides *position* control; moving the mouse changes the position of the cursor. A joystick provides *rate* control; moving the joystick alters the speed of the cursor.

Fixed vs. Variable rate. Devices with rate control can have fixed or variable rate. The input is usually mapped to velocity of the controlled object (e.g., the cursor or the pointing finger). Keyboards or buttons on gamepads have *fixed-rate* control. When a key or button is pressed, it controls an object at a constant rate. Joysticks generally provide *variable-rate* control. How fast an object moves depends on how far the joystick is moved or how much force is applied to the joystick.

The table below provides a summary of the above properties for the five input configurations.

Table 6.2: Properties of input configurations.

			Absolute Relative	Direct Indirect	Position Rate	Fixed Variable
Mouse + Keyboard	Point	Mouse (mode)	Relative	Indirect	Position	
	Move	Keyboard	Relative	Indirect	Rate	Fixed
	Look/Turn	Mouse (mode)	Relative	Indirect	Position	
Trackball + Mouse + Keyboard	Point	Trackball	Relative	Indirect	Position	
	Move	Keyboard	Relative	Indirect	Rate	Fixed
	Look/Turn	Mouse	Relative	Indirect	Position	
Gamepad	Point	Right stick	Relative	Indirect	Rate	Variable
	Move	D-pad	Relative	Indirect	Rate	Fixed
	Look/Turn	Left stick	Relative	Indirect	Rate	Variable
Joystick	Point	Main stick	Relative	Indirect	Rate	Variable
	Move	Buttons	Relative	Indirect	Rate	Fixed
	Look/Turn	Hat	Relative	Indirect	Rate	Fixed
Wii Controls	Point	Wiimote	Absolute	Direct	Position	
	Move	Balance board	Relative	Indirect	Rate	Variable
	Look/Turn	Nunchuk	Relative	Indirect	Rate	Variable

6.6 Method

In this section, I describe the participants, apparatus, conditions, procedure, and tasks of the study.

6.6.1 Participants

Ten university students (5 male and 5 female) were recruited; ages were between 20 and 28 (mean 23.7), and five participants were experienced with video games (more than three hours per week).

6.6.2 Apparatus

The study used a CVE built with XNA and C#, running on a Windows 7 PC with a 22-inch LCD monitor at 1680 x 1050-pixel resolution. The CVE had three versions (one for each task; see Figure 6.11 to Figure 6.13); all contained random targets and an avatar. The avatar was controlled by the input devices as described above (Figure 6.6 to Figure 6.10).

6.6.3 Conditions

The study tested one main factor (input configuration, with five levels as described above) in a within-participants design. Secondary factors were gender and prior gaming experience (gamer or non-gamer). The first two tasks were presented in balanced order, with the third task always last (for additional training time, since it was the most difficult). Differences between tasks were expected and so tasks were analysed separately.

6.6.4 Procedure

At the beginning of the study, the participant was informed of the purpose of the study, signed a consent form, and filled out a demographic survey. Then, the participant did each task with all the five input configurations (balanced with a Latin square design). The participant filled out NASA-TLX worksheets (Hart & Staveland, 1988) to record subjective effort after using each configuration, and stated preferences and gave comments for the devices after finishing all tasks. At the end of the study, the participant was debriefed and was given $15.00 as remuneration. The study took about one hour and thirty minutes to complete.

6.6.5 Tasks

There were three tasks in the study. The tasks involved combinations of pointing, moving, turning, and looking. Participants were asked to point out some objects to a simulated partner in

the CVE. However, in order to find out if people can control free pointing at all, the participants were told to focus on pointing control and did not need to communicate verbally to the simulated partner.

Task 1: Move-and-Point (MP). In the MP task, referents were located on a wall in front of the avatar (Figure 6.11), and participants were asked to point at the referents while moving sideways. This corresponds to many real-world communicative situations (e.g., discussing what items to buy on a grocery store shelf). Avatars were restricted to moving left and right in this task, to prevent participants from simply moving back until they could see all the objects at once.

Participants moved the avatar's arm (using the input device specified by the experimental condition) to point at the referents. A red dot on the CVE's wall indicated where the avatar was pointing. The referent disappeared after being pointed to, and a trial ended once all ten referents were correctly indicated with deictic pointing (i.e., as if the participant is stating "*this* one and *this* one and *this* one…"). The task had eight trials of ten referents each (the first three trials were marked as training and the rest were used for analysis). The dependent measure was completion time. Time was recorded at the end of each trial (i.e., after all ten referents were indicated).

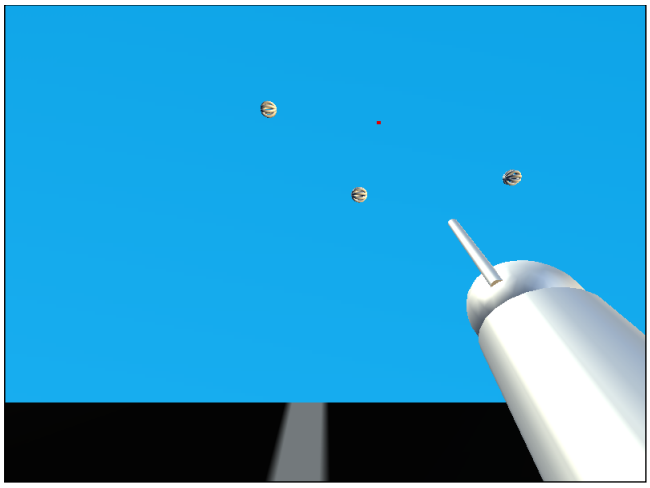


Figure 6.11: Move-and-Point task. Participant moves left or right, pointing at referents along the way.

Task 2: Turn-Look-and-Point (TLP). In the TLP task, referents were placed on the walls of a room in the CVE (Figure 6.12). The participants were asked to turn all the way around in the room, looking up and down to find the objects, and indicating each object to the simulated listener by pointing at it. This task corresponds to real-world communicative scenarios such as

when a realtor indicates various features of a room when showing a house. This task was also involve eight trials (the first three were for training and the rest were for analysis) of ten referents each, with completion time (i.e., the time to point at all ten referents) as the dependent measure.

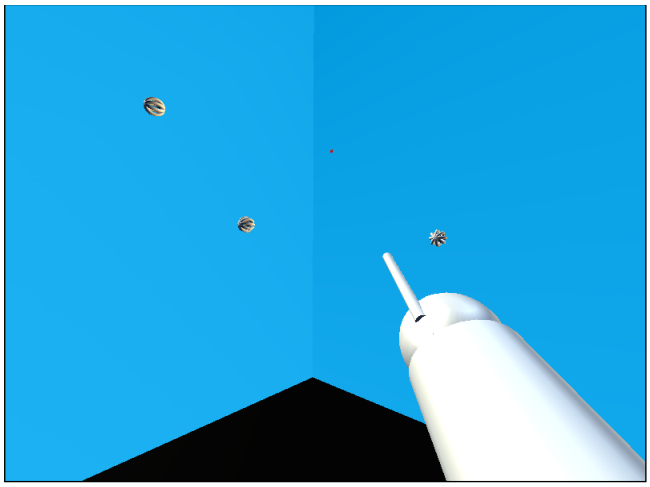


Figure 6.12: Turn-Look-and-Point task. Participant turns around inside the room, finding referents and pointing to them.

In the MP and TLP tasks, referent locations were pre-set with different locations for each trial. Participants saw the same locations for each configuration, but since there were several different sets of referents, and trials were randomly ordered, participants were not expected to learn the locations.

Task 3: Move-Turn-Look-and-Point (MTLP). In the MTLP task, participants pointed at a green spot on the wall while moving to a particular location in the room (marked with a ball on the floor, see Figure 6.13). The green spot traveled in a slow circle on the wall, requiring that participants continually adjust their pointing direction. Since the task also required that the avatar turn en route to the destination (to remain facing the green spot), participants were required to control all four avatar actions simultaneously. This task corresponds to situations where people must point out a moving object to another person, while also moving to a particular location (e.g., pointing out the movements of a bird to a friend, while walking towards and through a gate).

In the MTLP task, each trial had only one spot on a wall and one destination on the floor (there were three spot locations, and three destination locations). Participants carried out 20 trials, with

the first two were marked as training; the remaining trials covered all spot/destination combinations. To force participants to maintain a certain level of accuracy in their pointing, the trial would re-start if the avatar's arm left the green spot for two seconds. The dependent measure was the percentage of total time that participants' gesture is outside the green spot (i.e., error rate). Error rate was used as the dependent measure instead of completion time because I wanted to measure how well participants could adjust pointing directions (for constantly pointing at a moving target) while controlling other avatar actions, but not how fast the participants could reach the destination.

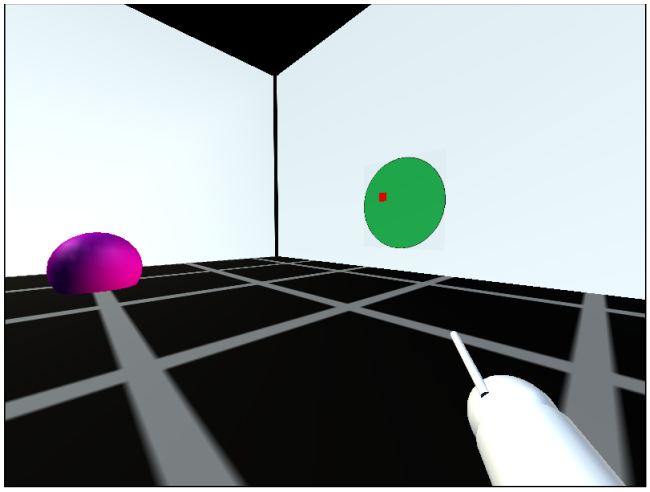


Figure 6.13: Move-Turn-Look-and-Point task. Participant moves to the ball while continuing to point at the green spot on the wall.

6.7 Results

The following sections analyse results from the three tasks, look at the effects of gaming experience and gender, and report perception of effort and preference ratings.

6.7.1 Task 1: Move-and-Point (MP)

The mean time to finish a trial was 16.6s. Analysis of variance (ANOVA) showed a significant main effect of device ($F_{4,36} = 31.79$, $p < .001$). A post-hoc Tukey HSD test showed that the mode-switched mouse was significantly faster than the trackball, gamepad, or joystick; and that the Wii setup was faster than the gamepad or the joystick (all $p < .05$). As seen in Figure 6.14, the differences were substantial: the mouse took half the time of the slower devices.

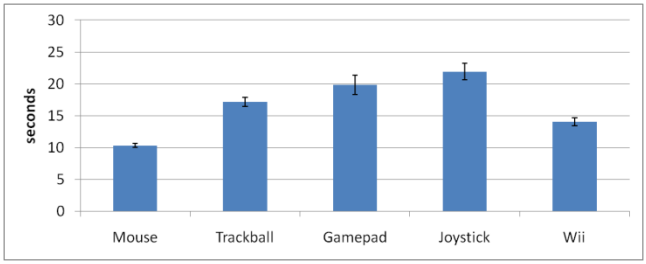


Figure 6.14: Mean completion time, MP task (error bars indicate standard error).

6.7.2 Task 2: Turn-Look-and-Point (TLP)

The mean time to finish a trial was 19.90s. ANOVA again showed a significant main effect of device ($F_{4,36} = 36.09$, $p < .001$). A Tukey HSD test showed that the Wii setup was significantly faster than the gamepad and the joystick; and that the mouse and trackball were also significantly faster than the joystick (all $p < .05$). Again, the differences are large (Figure 6.15): for example, the Wii was almost 15 seconds faster than the joystick and almost eight seconds faster than the gamepad.

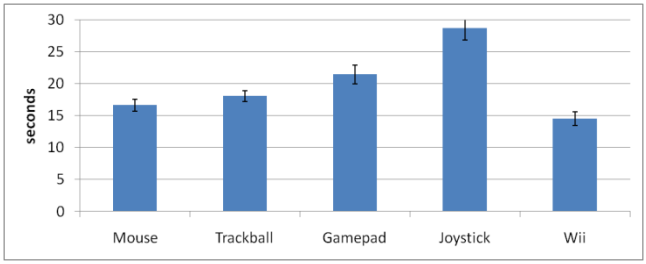


Figure 6.15: Mean completion time, TLP task (error bars indicate standard error).

6.7.3 Task 3: Move-Turn-Look-and-Point (MTLP)

The mean error (percentage of time the avatar was not pointing at the green spot) was 28.1%. ANOVA again showed a significant main effect of device ($F_{4,36} = 18.41$, $p < .001$). A Tukey HSD test showed that the mouse and the trackball were more accurate than the gamepad or joystick, and that the Wii setup was faster than the joystick ($p < .05$). As seen in Figure 6.16, accuracy results are similar to the completion-time results above: the better-performing devices are substantially more accurate (less than half the error rate in some cases) than the poorer devices.

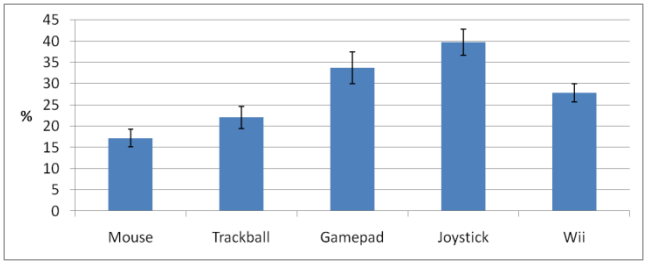


Figure 6.16: Mean error, MTLP task (error bars indicate standard error).

6.7.4 Effects of Video-Game Experience

ANOVA showed no effect of game experience in either the MP task ($F_{1,8} = 3.57$, $p = .095$) or the TLP task ($F_{1,8} = 2.85$, $p = .13$), but a significant effect was found in the MTLP task ($F_{1,8} = 5.34$, $p < .05$) (see Figure 6.17). In this task, gamers were slightly better able to continue pointing at the green spot as they moved (8% less error) than non-gamers. There were no interactions between device and game experience for any task: MP ($F_{4,32} = 0.63$, $p = .65$); TLP ($F_{4,32} = 1.19$, $p = .34$); MTLP($F_{4,32} = 0.27$, $p = .90$).

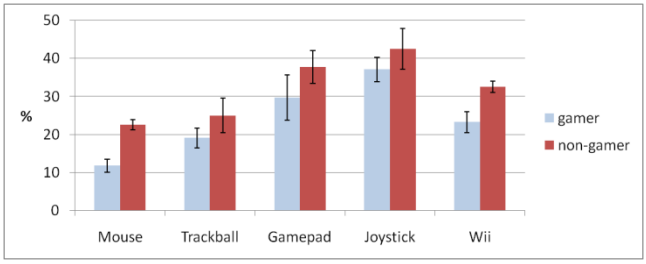


Figure 6.17: Mean error, by game experience in MTLP task (error bars indicate standard error).

6.7.5 Effects of Gender

ANOVA showed no main effect on gender in any of the three tasks: MP ($F_{1,8} = 0.17$, $p = .69$), TLP ($F_{1,8} = 0.00$, $p = .99$), and MTLP ($F_{1,8} = 0.00$, $p = .99$). No significant interaction was found between the devices and gender in the MP ($F_{4,32} = 0.88$, $p = .49$) or TLP tasks ($F_{4,32} = 0.96$, $p = .44$), but there was a device by gender interaction in MTLP ($F_{4,32} = 2.72$, $p < .05$). As shown in Figure 6.18, the interaction likely arises from the fact that women were better with the mouse and Wii setup than men, but worse with the gamepad and the joystick.

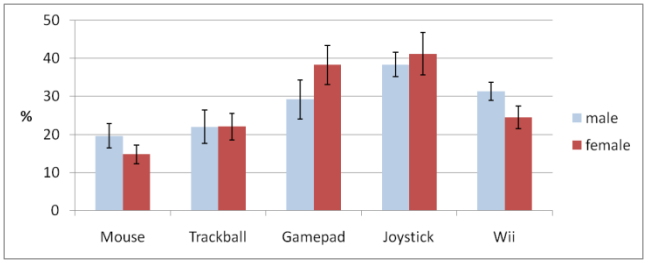


Figure 6.18: Mean error, by gender in MTLP task (error bars indicate standard error).

6.7.6 Perception of Effort and Preferences

The TLX effort questionnaires (taken after each device condition) showed results that are consistent with the performance data (Figure 6.19). In general, people felt that the mouse required the least effort (considering all three tasks), and that the joystick and gamepad required the most. Exceptions did appear, however: for example, the trackball and the gamepad were seen as requiring low physical effort but high mental load.

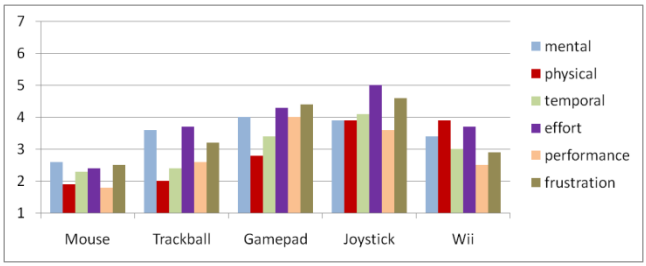


Figure 6.19: Mean scores (1-best, 7-worst) for workload assessment across all devices.

Figure 6.20 shows participants' overall ratings of the devices (1 = best, 10 = worst). There were substantial differences in these ratings: most participants preferred either the Wii configuration or the mouse; the trackball and gamepad were generally next, and the joystick was rated worst.

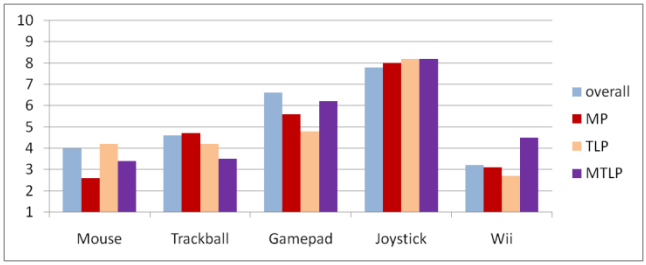


Figure 6.20: Mean rating by control and task (1-best, 10-worst).

6.8 Observations

Several participants found the simultaneous control tasks difficult at first, but they became more comfortable using the devices as the study progressed. Some participants tried to avoid using different controls at the same time in the MP and TLP tasks: for example, they moved (in MP) or turned (in TLP) the avatar, stopped, and pointed at all the objects showing on the screen; then repeated the same action sequence until all objects were indicated. However, in MTLP—where it was easier to complete the task with all actions used together—the participants tried to use multiple controls simultaneously. They manipulated all the controls with the trackball, joystick, and Wii configurations (it was possible to manipulate all controls at the same time with these devices, but not with the mouse or gamepad).

Participants also used a variety of motions to point to objects: in addition to the pointing controls, they also utilized the avatar's movement and rotation. In the process of pointing to an object, people sometimes adjusted the arm to the correct horizontal level (where the object was), and then either moved or turned the avatar so that the arm direction gradually approached the object. This behavior happened with all input devices except the Wii controls; but most often with the trackball setting. This appeared to be because the Wiimote was easier to control for pointing, whereas the trackball was harder for left-right manipulation (see Section 6.8.2 below).

6.8.1 Mouse

Participants were very comfortable using the mouse and keyboard. They were able to switch between modes (pointing and looking/turning) with ease. The mouse's familiarity and simplicity are likely reasons for its top performance.

The mouse was slower in TLP than in MP, and this may have occurred for two reasons. First, the requirement for mode switching may have slowed participants (pointing and viewing controls were both required, but were in different modes). Although participants found the mode switch easy to do, it is impossible to carry out the two actions simultaneously. Second, turning required clutching, whereas there is no clutching when using the keyboard keys to move in the MP task.

6.8.2 Trackball

The trackball showed reasonable performance. While this configuration allows simultaneous control of all actions, horizontal ball manipulation was difficult. Most participants only used the trackball to move the arm vertically with very few horizontal movements. It appeared that left-right manipulation of a mouse-mounted trackball was a somewhat unnatural motion. A typical trackball is controlled with the combination of wrist movement (for left and right) and finger movement (for up and down). However, the trackball in the Mighty Mouse must be manipulated by the finger in both directions, which participants felt was unnatural.

6.8.3 Gamepad

Some participants commented that they wanted to simultaneously manipulate all the avatar's actions with the gamepad but it was impossible to do so in normal use (the left thumb can only control either the left thumbstick or the d-pad). Others said that they liked to use the thumbstick (a variable-rate input) to change the view because they could hold the thumbstick at a fixed angle and the view would change at a steady rate, allowing them to focus on pointing out the referents.

6.8.4 Joystick

The joystick was the worst configuration on all measures (performance, preference, and effort). It caused considerable frustration across all participants, and one main reason is that participants often unintentionally changed their pointing and viewing directions. This happened more often in the MTLP task where the participant tried to use all the actions at the same time. They had substantial difficulty controlling pointing and looking/turning on the same device with one hand. Most participants experienced problems of simultaneous control with the joystick: for example, losing track of the avatar's arm by looking up while pointing down. This occurred because there are two 2D-velocity-based inputs (the main stick and the hat) assigned to one hand. In addition, the two inputs are physically stacked (the hat is built on top of the main stick). Pointing (with the stick) and turning/looking (with the hat) often affect each other—this also appears to be why participants spent more time in the TLP task than the MP task.

6.8.5 Wii controls

Participants liked the Wii configuration, and were enthusiastic about using it. They liked the naturalness of pointing with the Wiimote, and were all able to point easily and comfortably. Participants' comments and preference rankings suggest that they liked using the Wiimote to point and the Nunchuk (thumbstick) to look and turn, but did not like the Balance Board for movement as much. Some participants experienced difficulty with using the Balance Board—for example, moving the avatar past the desired destination in the MTLP task. Some participants also commented that they preferred not to use a foot-control device as it was physically demanding. This difficulty could be improved with different thresholds. Despite problems with the Balance Board, the Wii controls were very good in both performance and preference. The main advantage of this configuration is that people can carry out pointing the same way that they do in the real world.

6.9 Discussion

The study has four main findings:

1. People can successfully control free pointing in tasks that already involve moving, turning, and looking (answer to research question 1);

2. There were significant and substantial differences between the input devices for all three tasks;

3. The mouse and Wii configurations were consistently better (answer to research question 2), and the game controller and joystick were consistently worse;

4. There were minor effects of game experience and gender.

Overall the most important conclusion from these results is that there are avatar control configurations in which adding communicative expressiveness (free pointing) does not unduly burden the individual's control abilities.

6.9.1 Lessons and Design Issues

Here I list eight main lessons learned from the study.

1. ***Adding free pointing to CVEs is feasible.*** The study shows that people can handle the addition of free pointing to existing avatar controls. In all of the tasks, participants were able to complete the tasks successfully and without undue difficulty (although the device matters). This main result suggests that designers can feasibly incorporate this additional capability into CVEs.

2. ***A mode-switch mouse is a usable option.*** The mouse configuration made it impossible to turn and point simultaneously, yet the mouse had the best overall performance and was second best in preference. People were able to perform very well with the device. This configuration also represents the simplest extension of standard controls, and could easily be implemented in CVEs. Although the left button is often already used in some CVEs, a different mode switch could also be equally successful.

3. ***Direct input is good for pointing.*** The Wii controls had strong performance (best in the TLP task, second best in the MP task, and third in the MTLP task) and the best overall preference. Even though the participants did not have much experience with the controls, they got used to this configuration very quickly. Also, the Wii was the only condition where participants did not offload aiming to the avatar's movement and rotation (that is, they always used the pointing controls to move the arm towards the object). Direct-pointing configurations appear to be a useful new direction for avatar control systems. An additional benefit of direct pointing is that it can easily be extended to more complex gestures.

4. ***Effects of previous experience.*** People have more experience with the mouse and less with the Wii, yet these devices both showed better results than others. Also, gamers did not perform significantly better with game controllers. These results suggest that the differences between devices are not solely due to people's experience and that although people can learn to control almost any device, there are configurations that are more natural and simpler for controlling free pointing.

5. ***Controlling two 2D inputs with one hand is difficult.*** The joystick was the worst configuration overall. It required people to control two actions (pointing and turning/looking) with one hand, and was disliked and seen as difficult. Comments (e.g.,

"it's confusing") and observations also showed that people simply had more difficulty controlling the avatar in these configurations: for example, trying to turn one way and point in the opposite direction was problematic. The trackball configuration also had two 2D inputs on one hand, and although it performed well in the MTLP task, people commented (e.g., "harder than I thought") that the combined actions were difficult.

6. ***Variable-rate control is good for panning the view.*** For the gamepad and Wii controls, the TLP task had the best preference ranking compared to overall, MP, and MTLP. This is mainly because of the variable-rate controls that were used for look and turn. With the variable-rate control, the participants were able to control the turning rate by holding the thumbstick at a certain angle, allowing them to focus on other actions. This is interesting because many current CVEs put turning control on the mouse, which is a position-based device.

7. ***Physiological constraints affect device use.*** Participants felt that the Mighty Mouse trackball was unnatural for controlling arm movement because the trackball cannot be used normally (where the wrist is used for left and right control). Input configurations should be designed to fit with an understanding of ergonomic factors such as the range of motion of different limbs

8. ***Input sensitivity should be adjustable.*** Different participants had very different preferences in terms of input sensitivity (most obviously seen on the Wii Balance Board). While having default sensitivity is important, it should be adjustable. Although adjustable settings are common in desktop applications, it is not always common in CVEs to allow full specification of parameters.

6.9.2 Generalization to Other Communication Tasks

The results are likely to translate to more realistic collaborative task situations and other CVEs. First, real communication situations involve the additional task of generating verbal communication along with the avatar's gestures. This is not likely to create problems for free pointing because the control situations in the study likely require more simultaneous activity than what is needed in many collaborative situations, and because people in real tasks will have far

more experience with the controls than the participants. In addition, using natural actions such as direct pointing can greatly simplify the task of generating these gestures. The results can also be useful in non-CVE systems where pointing gestures are important, e.g., rescue operations and equipment maintenance supported by remote experts.

Second, pointing control worked for a broad range of participants: gamers and non-gamers, and men and women. It seems clear that the ability to control free pointing is not limited to only a small group of users. Finally, the devices are all readily available and do not require specialized hardware or software—this means that pointing control could be easily added to a wide range of CVEs.

CHAPTER 7
COMPARING POINTING TECHNIQUES

In previous chapters, I made distant pointing in CVEs more expressive by adding free and natural pointing to basic avatar actions, and showed that people can successfully use and control these pointing methods. However, we do not know whether free and natural pointing are useful for collaboration in CVEs when augmented-pointing techniques are available. To investigate this issue, I conducted a study observing how people communicate using free, natural, and augmented pointing in a CVE with realistic tasks. The primary goal is to determine if free and natural pointing are useful in CVEs even when augmented-pointing techniques are available.

One main reason that distant pointing is challenging in desktop CVEs is that the field of view is too narrow, thus causing the problem of 'fragmentation'—the display cannot show all the relevant information needed for communication, forcing collaborators to change their views to see such information (Fraser et al., 1999; Hindmarsh et al., 1998, 2000). Increasing the field of view width may make collaboration using free and natural pointing easier. Therefore, the secondary goal of the study is to explore how field of view affects the use of distant-pointing techniques.

In this chapter, I answer the following two questions:

1. Are free and natural pointing useful in CVEs even when augmented-pointing techniques are available?

2. How does a wide field of view affect the use of distant-pointing techniques?

7.1 Pointing Techniques

The study compared five pointing techniques: natural pointing (the default), natural pointing with a long arm, a virtual laser beam, a spotlight technique, and an object highlight (See Figure 7.1).

- ***Natural pointing***. The avatar's arm and extended finger pointed as controlled by participants.
- ***Long arm***. Similar to natural pointing, but with a double-length avatar arm, making it easier to see.
- ***Laser beam***. A red line (the 'laser') is drawn from the avatar's finger.
- ***Spotlight***. A small red dot is drawn on the first intersecting object in the arm's pointing direction.
- ***Highlight***. If a selectable object intersects the arm's pointing direction, it is highlighted in red.

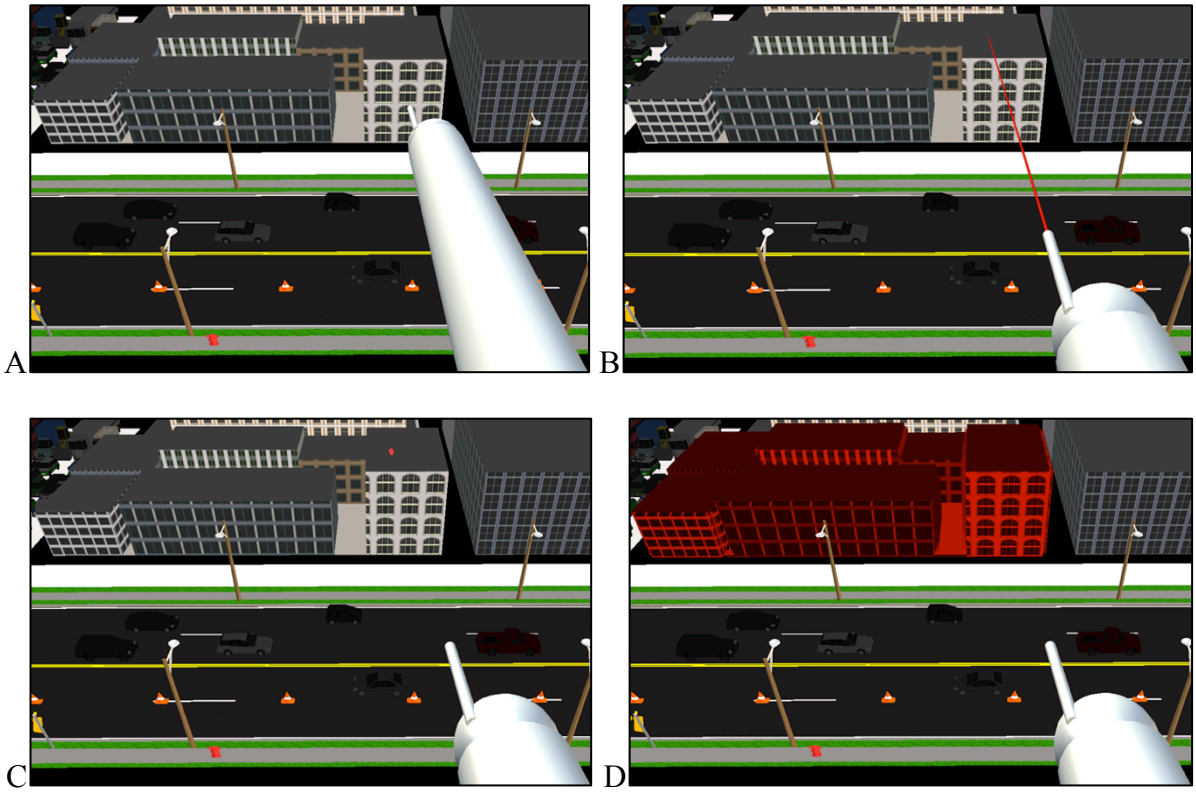


Figure 7.1: Pointing techniques: A) long arm, B) laser beam, C) spotlight, and D) highlight.

To better understand the importance of arm movement and avatar orientation, I also tested the three visual augmentations (laser beam, spotlight, and highlight) with two variations. The first variation is to point without the corresponding avatar arm movement. For example, when a laser beam was used with an immovable arm, the laser would come out from the avatar's shoulder (see Figure 7.2B). The second variation is to point with an invisible avatar; the visual augmentations could still be seen when activated (Figure 7.2C).

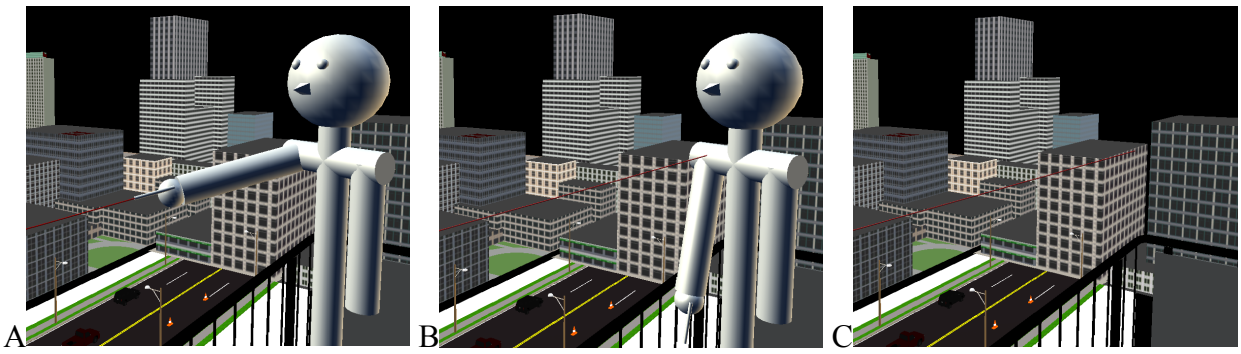


Figure 7.2: Using a laser beam with A) a regular avatar, B) a fixed arm, and C) an invisible avatar.

7.2 Method

In this section, I describe participants, experimental setup, apparatus, procedure, and tasks in the study.

7.2.1 Participants

There were 12 pairs of participants (15 male and 7 female) with the mean age of 26. The participants were university research assistants, and graduate and undergraduate students across nine departments. All had computer experience (10-80 hours/week using computer), and 17 of them had some experience with CVEs via video games.

7.2.2 Experimental Setup and Apparatus

Experimental room. The study was in a quiet room where two desktop workstations were located 10 feet apart at the opposite ends of the room. A pair of participants faced opposite directions so that they could only see their own displays, and could talk to each other freely without using headphones (Figure 7.3).

Collaborative virtual environment. The CVE was built in XNA and C#, running on Windows 7. In the CVE, avatars were set on a balcony looking at the downtown area of a city (Figure 7.4). Users were able to control an avatar to move around the balcony, point anywhere with the avatar's arm and forefinger, and look in any direction by turning the avatar's head.

Figure 7.3: Physical setup of the experimental room.

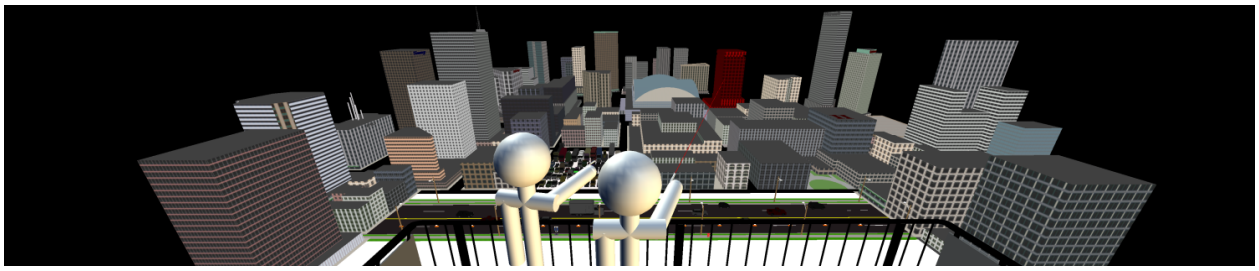


Figure 7.4: A CVE set on a balcony with a downtown view.

Movement and arm control. Participants used the two input methods that were identified as being effective and easy to use from the study in Chapter 6, i.e., mouse and Wii control (with slight modifications). With either control, participants could move their avatars using the W-A-S-D keyboard keys. All of the pointing techniques could be controlled by manipulating the direction of the avatar's arm. For mouse control, the arm could be moved when the left mouse button was pressed; the arm pointed in the direction indicated by the left-right and up-down motion of the mouse. For Wii control, participants could move the arm using real-world pointing movements (pointing left, right, up, or down in the real world changed the avatar's arm in the corresponding directions). Figure 7.5 shows how a participant used the Wii control. The Wiimote was strapped to the back of the hand that held the Nunchuk. Natural pointing was the default behaviour. Other pointing techniques could be activated by pressing number keys on the keyboard.

Figure 7.5: Wii control.

View control. When using mouse control, participants could change the avatar's view direction by moving the mouse with no buttons pressed (this meant that participants could only control either view direction or arm direction at one time). When using the Wii control, participants could change view direction using the thumbstick on the Nunchuk controller. In addition, participants could switch between first-person and third-person views by pressing the Tab key.

CVE view width. One of the workstations had a single 24" monitor with a 1440 x 900-pixel resolution. The other workstation had three 24" monitors with a total resolution of 4320 x 900 pixels, tripling the width of the CVE's view. In both cases, the CVE view width increased when changing to third-person perspective where the camera was located behind and above the avatar.

7.2.3 Procedure

At the beginning of the study, the pair of participants was informed with the purpose of the study. Each signed a consent form, and filled out a demographic survey. Then, the participants performed collaborative tasks with the five pointing techniques (natural, long arm, laser, spotlight, and highlight) in a CVE. The experimenter observed how the tasks were done; the tasks were also video-recorded for analysis. After that, the participants filled out a post-test questionnaire regarding preferences on pointing techniques and view settings. At the end of the study, each participant was given $15.00 as remuneration and was debriefed as to the purpose of the study. The study took about one hour and thirty minutes to complete.

7.2.4 Tasks

Participants carried out a set of referencing tasks and a set of creative tasks. In the referencing tasks, one participant (the gesturer) indicated different referents in the scene to the partner (the observer). The gesturer chose a referent from each of these categories: buildings, rooms in buildings, cars, paths (e.g., the path from a building to a car), areas (e.g., a 2x2 block area), and general directions (e.g., where the downtown area should expand next). The participants could talk to each other and use their avatars to communicate.

In the creative tasks, participants collaboratively constructed stories and made decisions based on scenarios given to them. For example, participants were asked to pretend to be roommates and to look for an apartment together, taking into consideration issues such as type of neighborhood, distance to work sites, and traffic. These tasks involved referencing and pointing from both participants, and were open-ended with no defined solution.

Pairs worked on the tasks with each of the pointing conditions, with both the mouse and Wii controls, and with both the three-monitor and one-monitor setups. Participants could switch between first- and third-person views at any time.

Data were collected via written notes during observation, video recording throughout the study, and a questionnaire at the end. Over 15 hours of video data were collected for interaction analysis.

7.3 Results

The results are organized by differences between pointing techniques, interpretation of natural and augmented pointing, and the effect of different view organizations.

7.3.1 Differences Between Pointing Techniques

Differences between the different pointing techniques can be described in six main areas: the ease of controlling the pointing gesture, the specificity of the technique, the feedback provided to the gesturer, the degree of visual clutter introduced, the ease of connecting the visual effect to an avatar, and participants' overall preference for the techniques.

7.3.1.1 Controllability

All of the techniques were controlled in the same way (through manipulation of the avatar's arm direction), using either mouse or Wiimote control. Participants experienced no difficulties in pointing with either of these methods, and there did not seem to be any interaction between control method and pointing technique.

There were other differences, however, in the way that participants specified referents in the scene. First, all of the augmented techniques had to be switched on and off with a key press, and several participants saw this as a non-trivial effort (since it required moving the left hand on the keyboard). For example, some people preferred the default natural pointing to the long arm technique for this reason.

Second, the object-highlighting technique was seen as more difficult to control because it only worked with pre-defined objects in the scene. In a few cases, when participants did not see the highlight effect, they did not know whether they had missed with their pointing action, or whether the object was simply not selectable. Other participants found that this technique was hard to use because of its discrete movement (i.e., jumping from object to object).

Third, in several cases, the gesturer inadvertently left the avatar's arm in a raised position, even though they were not intending to point at something. There is no natural proprioceptive sensation of the avatar's arm position, and the arm did not automatic lowered, so gesturers sometimes forgot to manually lower the arm to a non-pointing position. The following instance illustrates how this caused distraction. Sean was showing Emily a building (Figure 7.6 shows Sean's view). He kept the city and Emily's avatar in his view. While he was describing the building, Emily's avatar's arm was constantly moving and pointing at random referents. Although Emily knew how to use the controls to lower the arm when not explicitly pointing, she did not do so. As a result, Sean had difficulty interpreting Emily's intentions. Sean commented: "I think the arms can be a little distracting, when trying to figure out what someone else is referring to. A few times my partner wasn't using her arm and just letting it hang in a random direction. I tried to infer meaning from the arm."

Figure 7.6: Emily unintentionally pointing in random directions.

7.3.1.2 Specificity and Perceived Accuracy

There were substantial differences between the techniques in terms of accuracy and specificity. Participants felt that the laser beam and the spotlight were the most specific, followed by the highlight and then the two arm-based techniques. It was clear, however, that there was a strong relationship between the perceived accuracy of the different techniques and the size of the referent in the scene. For example, when participants had to refer to a small object or a precise location, they felt that the task was considerably easier with the laser or spotlight, and required less verbal communication. However, not even the laser was a complete replacement for verbal information (see Section 7.3.2).

For larger targets, participants also perceived other techniques as being accurate—for example, if the referent was a building, then the object-highlight technique was at exactly the correct level of granularity. For indicating general directions, even the arm-based pointing techniques were seen as being sufficiently accurate for the task.

Observations also showed how people used techniques of different specificity for different-sized targets. When participants used a more-specific technique (like the laser beam) with a less-specific referent (e.g., an entire building, or an area of the city, or a general direction), they had to be careful that the other person did not over-interpret their pointing gesture. People used verbal communication to broaden the scope of the referent (e.g., saying "this entire building") and moved the laser around the referent to avoid appearing to point to a more specific location.

When people used a less-specific technique with a more-specific referent (e.g., the arm or the highlight to point to a single window in a building), they also used verbal communication, but now to provide information that would allow the hearer to identify the specific referent. For

example, when referring to a room in an apartment building, participants would highlight the building and said "the leftmost room on the third floor of that building."

With the arm-based techniques, there was more verbal communication overall, and more interaction between the participants as the gesturer sought confirmation that the observer had identified different landmarks at increasing levels of specificity. That is, people often referred to obvious landmarks first (e.g., "OK, you see the stadium?") before moving to smaller objects (e.g., "now look at the tall building to the right of that"). In addition, the arm-based techniques often led to more avatar movement: observers frequently moved their avatars much closer to the gesturer, so that they could see the pointing gesture from the same perspective as that of the other person. However, this occasionally led to confusion as shown in the following conversational fragment (with annotations as shown in Figure 7.7).

G: I think that building down the en:::d there straight in front of me with the (3.5) I intended to say green line,
G: but ah:: I don't know if [( )
D: [( ) in your way here, I tell ya
G: Which arm is mine?

The transcription notations are based on Jefferson's work (Jefferson, 1984).	
(3.5)	time paused in seconds
(.)	a brief pause in less than 0.2 seconds
( )	unintelligible words
:	stretched sound
[ ]	speech overlap

Figure 7.7: Transcription notations.

George and Dan were looking for an apartment to share. George suggested an apartment building to Dan, who was standing to the right of George when Dan said, "I think that building down the en:::d there straight in front of me with the (3.5)". During the 3.5 seconds pause, Dan moved towards George, overlapped his avatar with George's, and pointed at the same building. George could not identify which avatar arm belonged to whom, leading to his question "Which arm is mine?" (Figure 7.8).

Dan wanted to know exactly what George was seeing and pointing at. Although he knew that getting too close could cause problems such as blocking each other's view ("in your way here, I tell ya"), he still moved close and even overlapped the avatars, causing confusion.

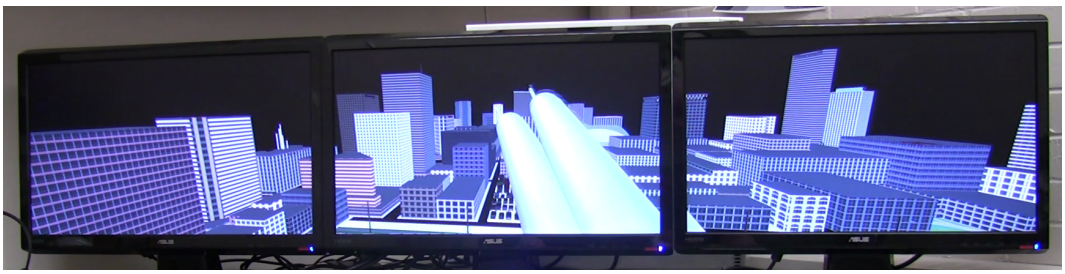


Figure 7.8: George's and Dan's arms were partly overlapped.

7.3.1.3 Feedback

The different pointing techniques were shown to provide different types and amounts of feedback to the gesturer, changing the way that the technique was used. First, the default arms were difficult to see in the first-person view, and most participants switched to the third-person view when using this technique. Second, the spotlight did not provide strong visual feedback—the spot was small, could be hard to see when pointing at an oblique surface, and did not appear at all when pointing at the sky—and this caused some problems for participants (particularly in first-person view where the arm direction was not as visible).

The object highlight and the laser beam provided the most obvious feedback. The highlight was easy to see, although as described above it could be difficult to determine whether objects were selectable or not; in contrast, the laser beam was always visible and was seen as the easiest technique to position.

7.3.1.4 Visual Clutter

There were some discussions about the degree of visual clutter that would be produced if there were several people in the CVE. Several participants mentioned that it would be more difficult to identify referents if there were several lasers visible in the scene—and that the clutter would be worse with the larger-scale highlighting technique.

7.3.1.5 Determining Ownership

A problem related to that of visual clutter is determining who owns the visual effect of an augmented technique. This caused problems in several tasks, particularly when both participants used the spotlight and highlight techniques simultaneously (which occurred frequently during the

open-ended tasks). Determining ownership was seen by participants as a main strength of the laser beam—since the beam always visually connected the gesturer and the referent. Ownership was also easy to determine for the arm-based techniques, since these are connected to the avatar.

7.3.1.6 Preferences

All of the participants stated that they preferred the augmented techniques to the two arm-based techniques, because of the increased specificity provided by the visual effects. However, participants also said that their preference would depend on the specificity requirement of the task—for example, if only general directions were required, then the arm-based techniques would be sufficient.

The laser beam was the most preferred technique. The main reasons given by participants for this preference included its accuracy, its visibility, and the ease of determining ownership when multiple lasers were used at once. The only concerns expressed about the laser involved clutter when there were several people in the scene, and situations where visibility was not desired (e.g., in a game where people would not want to reveal their locations).

In addition, when using augmented pointing, participants preferred pointing with the regular avatar to the avatar with a fixed arm and an invisible avatar. Most participants stated that they liked the regular avatar because it was more natural than the other settings. The invisible avatar was the least preferred overall because it was awkward not to being able to see the avatar. One participant commented: “it’s just strange to see the laser coming out of no where.”

Participants also stated that they preferred to have multiple techniques available, so that they could match the technique to the specificity requirements of the task. As detailed in the next section, it was clear that people did make use of multiple techniques—in particular, they used the avatar’s arm direction (which was always available) even when using an augmented technique. Table 7.1 summarizes the characteristics of the five pointing techniques.

Table 7.1: Characteristics of pointing techniques.

		Characteristics
Pointing Techniques	Natural	- Important at early stages of pointing to provide staging actions - Difficult to see the arm in first-person view - Not specific - Accurate enough for indicating directions - Initiate more interaction and verbal communication between users
	Long Arm	- Not specific - Accurate enough for indicating directions - Initiate more interaction and verbal communication between users - Turning the technique on and off is non-trivial
	Laser Beam	- Overall most preferred - Connect the user to the referent - Specific - Easy to see - Obvious feedback - Turning the technique on and off is non-trivial
	Spotlight	- Specific - Small and difficult to see - Difficult to determine who is using the technique - Turning the technique on and off is non-trivial
	Highlight	- Only work with pre-defined objects - Difficult to determine if objects are selectable - Discrete movement - Easy to see - Obvious feedback - Difficult to determine who is using the technique - Turning the technique on and off is non-trivial

7.3.2 Use of Avatar Position Alongside Augmented Pointing

Augmented pointing techniques are much more visible, and in some cases, much more specific than pointing with just the avatar's arm. These changes suggest the possibility that the communication problems identified in earlier work (e.g., related to mutual orientation or fragmentation (Fraser et al., 1999; Hindmarsh et al., 1998, 2000)) will be fully addressed through the additional visual information.

The observations, however, indicate that although augmented pointing techniques such as the virtual laser beam can improve certain elements of communication, they do not remove the need to support other aspects of the pointing process. In particular, the study showed that even when augmentations were present, people still failed to notice or see some pointing gestures, and

people still looked at the avatar's body and arm as a way to orient themselves to upcoming augmented pointing actions.

7.3.2.1 Failing to See Augmented Pointing Actions

There were several situations where observers failed to notice that a gesturer was making an augmented pointing action. The spotlight technique was particularly difficult to notice, but participants also missed the more obvious augmentations (highlight and laser beam) on several occasions. This occurred when the observer was looking in a different direction (i.e., the visual effect could be out of view, particularly with the single-monitor setup), or when the gesturer was pointing down at the road in front of the balcony (which meant that the laser beam was occluded by the balcony itself). However, some participants missed the augmentations even when they were visible on screen, as illustrated in the following conversational fragment.

J: There's (0.5) a little yellow building (0.8) right in between (.) and behind these two buildings right across the street from us. Next in the little lot next to the parking lot. Have a yellow one (.) in the middle. You see it?
K: Yeah, I can see that.
J: Okay, and::: I guess the (.) one visible, the ground level room on the right that's actually (1.2). ( ) You can actually see more from your angle. Okay, we'll go ar:: ( ) top right. This little room right (.) he::re. This one on top floor. My laser pointing at it.
K: Which one? I actually can't.
J: hey, come over here and have a look.
K: Oh yeah there. I can see it.

Jason wanted to point out a room on the top floor of a yellow building. At the beginning of the fragment, after he said "There's (0.5)", he started to point at the building (on the right monitor of Kate's station; see Figure 7.9) with a laser beam, which was kept on for the whole fragment.

When Jason uttered "a little yellow building (0.8) right in between (.) and behind these two buildings right across the street from us", "in the little lot next to the parking lot", and "Have a yellow one (.) in the middle", Kate thought Jason was referring to another building with similar visual features (as shown in Figure 7.9, both buildings were in line with Jason's arm direction). Kate positioned her avatar and adjusted her view to center the mistaken building. She thought she knew the correct building until Jason said "My laser pointing at it." Kate then realized that she misunderstood ("Which one? I actually can't").

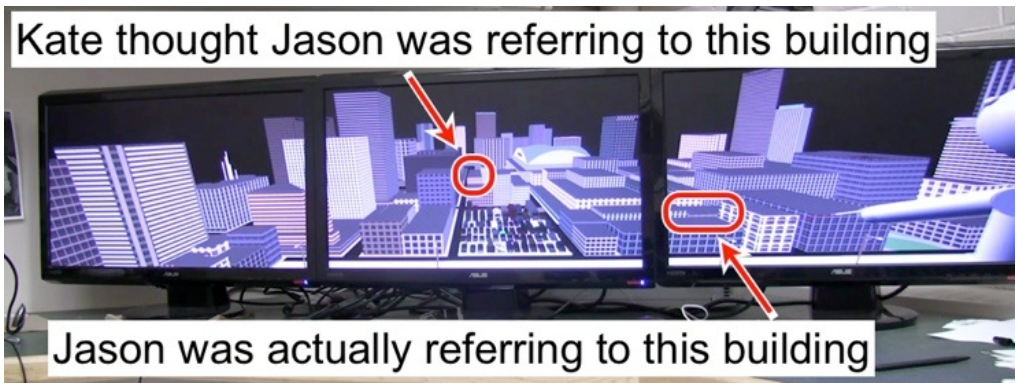


Figure 7.9: Kate did not see Jason's laser beam.

Although Jason was pointing at the building with a laser beam the whole time, Kate did not pay attention to it until he mentioned that his laser was pointing at the building. Kate commented at the end of the study that she was focusing on Jason's verbal descriptions and the mistaken referent; therefore, she did not see the laser beam and misunderstood the referent as a result.

This episode shows that in a busy visual environment, it is not always easy to notice even a substantial augmentation such as the virtual laser. It is also notable that although the laser was only about 25cm from Kate's focus of attention, it was on a different monitor—it is possible that the 4cm visual gap of the monitor bezels made it more difficult to notice activity that was occurring on a different monitor.

7.3.2.2 Watching the Avatar before an Augmented Gesture

There were several episodes where participants appeared to gather information from the other person's avatar as a way to orient themselves to an upcoming pointing action, even when the gesturer was using augmented pointing. For example, explication of the next fragment shows one participant using the gesturer's viewing direction and arm direction to help determine where to look in the scene.

M: That's the room.
K: Okay.
M: Ah:::: the car:: (19) ah::: in the parking lot.
K: Okay let me come over there.

In this fragment, Mark had just finished showing Ken a room in an apartment building with the laser beam. Ken confirmed that he knew which room Mark was pointing at ("okay"). Right after

this confirmation, Mark turned off the laser and moved to the left of the balcony. As he was moving, he started to consider which car he wanted to show Ken next ("Ah:::: the car::"). Mark took several seconds to decide, and during that time, Ken turned his avatar towards Mark's, adjusting his viewing angle so that he could see both Mark's avatar and the cars (Figure 7.10).

Figure 7.10: Ken's view showing Mark (left monitor) and cars.

Once Mark made his decision, he turned the laser beam on and said "ah::: in the parking lot." Ken immediately moved his avatar beside Mark's, such that he only saw the laser beam on his screen, not Mark's avatar (Figure 7.11).

Figure 7.11: Ken focused on Mark's laser beam (left monitor).

Knowing that Mark would use the laser beam to indicate a car, Ken could have oriented his avatar towards only the cars and waited for the laser beam to appear. However, Ken instead began by paying close attention to Mark's avatar: Ken changed his view to keep track of Mark's avatar orientation and arm direction. Before Mark turned the laser beam on and said "ah::: in the parking lot," Ken already knew roughly where the referent would be (because he could see Mark's arm direction) and was able to more quickly respond to the pointing gesture.

Once the area of the referent was established and the laser was switched on, however, it is notable that Ken moved to a location where he could no longer see Mark's avatar or arm. Once the general orientation is determined, the laser beam provides accurate information about the

referent; the observer no longer needs the avatar orientation information, and can focus their view on the target region.

7.3.3 View Changes and Fragmentation

Different displays and view perspectives were used in the study to explore how field of view affects pointing in CVEs. A three-monitor-wide display and third-person perspective were used to provide a wider field of view for improving interpretation of deictic pointing. However, episodes in the sessions suggested that even with these view features, fragmentation was still a problem.

7.3.3.1 Preferences for the Wider Views

All but one of the 24 participants preferred the 3-monitor display. Many participants commented that it gave them a much better ability to see the actions of their partners as well as the referents—for example, one participant said "It was easier to see everything and determine where [the other person] was looking." The three-monitor display also changed the way that people used the CVE. People could look around without changing their view controls, but the more spacious display and the increased number of visible objects make visual attention more of a problem (as in the previous section, it is not enough that objects be visible in the display; the person should also attend to the information).

In addition, most participants preferred the third-person view for two reasons: it showed more of the scene, particularly in the one-monitor display, and it showed more of the avatar's arm for pointing feedback. Some participants used the first-person view for a 'zoom-in' effect; this was sometimes necessary because in the third-person view, the avatar occasionally occluded some objects.

7.3.3.2 Use of the First-Person and Third-Person Views

Participants used different viewing perspectives in different settings and situations. The first-person view was used more often with the three-monitor setup and the third-person view was the primary choice with the one-monitor setup. Also, participants tended to use the first-person view when they were generating gestures. The following fragment shows a typical example.

D: That car::: (4) That car there (3.5) next to the pylon.

I: Okay.

Dora was using the one-monitor setup and the third-person view. She wanted to show Ivan a car on the freeway. After saying "That car:::", Dora tried to point at the car that was partially occluded by her avatar, but she had difficulty doing so (Figure 7.12A). She then changed the perspective to the first-person view and told Ivan "That car there." She then adjusted her view so that the car was in the center of the screen, pointed at it with a laser beam, and said "next to the pylon." (Figure 7.12B)

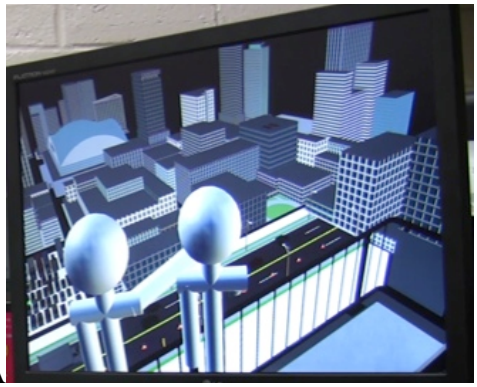
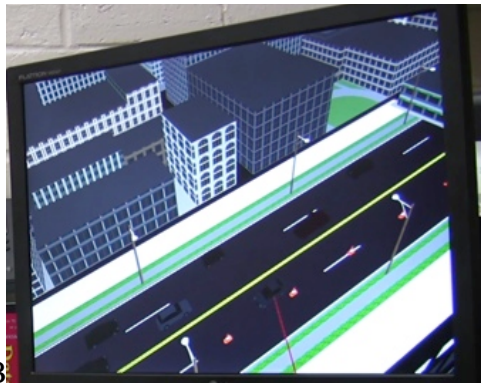
A  B 

Figure 7.12: Dora's one-monitor setup: A) with the third-person view; B) with the first-person view.

Dora used the third-person view as her default viewing perspective allowing her to see more of the scene to compensate for the narrow field of view. However, in third-person view, her avatar blocked the middle area of the scene. She spent about four seconds trying to point at a car that was partially occluded by her avatar, and then realized that the first-person view was the best perspective for the task at hand. She switched to using the first-person view so that she could clearly see and point to the car.

7.3.3.3 Knowing Collaborators' Viewing Perspectives

Participants' actions were highly dependent on what their partners could see. Participants often made assumptions about the viewing perspectives their partners were using. However, wrong assumptions could lead to confusion as shown in the following fragment:

T: Can you mo::ve to my position? And face yourself to the road. (5.5) If you face yourself to the road and (0.8) lower your camera view, you'll see two black cars on the road.

J: Two black cars?

T: The two black cars (0.6) on the left, on the opposite direction, there's a dark blue car

J: um::::::

T: Can you see that?

J: I can only see a black car.

Thomas wanted to show Joyce some cars. He pointed at the cars and helped Joyce finding the cars by giving her instructions ("mo::ve to my position", "face yourself to the road", and "lower your camera view"). Joyce followed his instructions, but still could not see the black cars he mentioned. Thomas then gave a further description, but without success. However, Thomas gave the instructions based on the assumption that Joyce was using the same view (first-person view) as he was, when in fact she was using the third-person view. In third-person view, Joyce's avatar was blocking her view of the objects being discussed (the black cars). Figure 7.13 shows what Joyce saw.

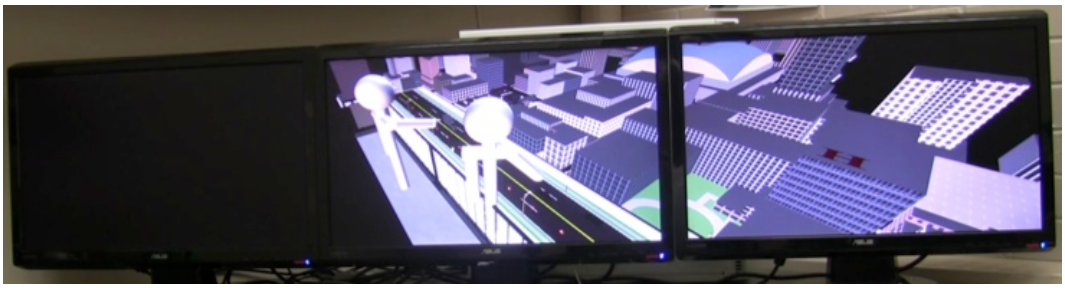


Figure 7.13: Joyce's avatar was blocking her own view.

7.3.3.4 Third-Person Views and Fragmentation

View fragmentation in CVEs is a well-known problem (Fraser et al., 1999; Hindmarsh et al., 1998, 2000). The following examples show how participants recognized the problem used the third-person view to solve it.

M: I just wanna be sure that we're pointing to the same building.

K: yeah yeah understand.

M: ar (.) did you use tab?

K: I'm I'm on the first person right now.

M: first person, so

K: Let me check where you're pointing at. (1.5) Can you point at it?

M: Yeah of course (.) I'm (.) pointing to that building.

Mark wanted to confirm that Ken and him were both pointing at the same referent. However, at the beginning of the fragment, Ken was using the first-person view and facing the buildings, thus he could not see Mark's avatar (Figure 7.14). Mark tried to understand what Ken was able to see, and so he asked, "did you use tab?" This question might have reminded Ken that he could change his viewing perspective. Ken replied that he was in the first-person view, but he switched to the third-person view (Figure 7.15) to see Mark's pointing gesture ("Let me check where you're pointing at.")

Figure 7.14: Ken was facing the downtown and could not see Mark's avatar in the first-person view.

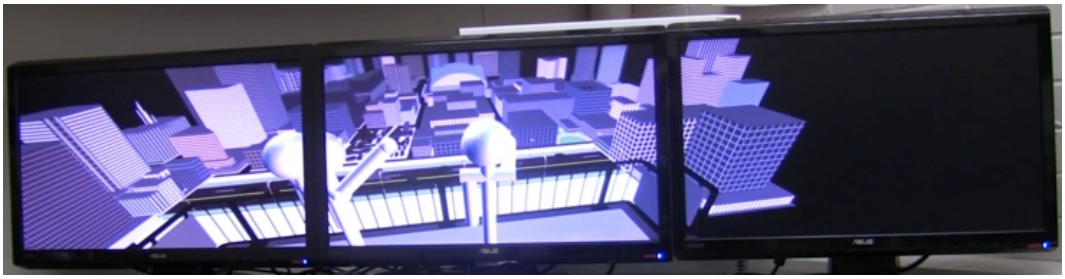


Figure 7.15: Ken was able to see the downtown and Mark's avatar in the third-person view.

Ken was focusing on the buildings, but he also wanted to look at where Mark was pointing to better understand what the referent was. Instead of changing his view back and forth between Mark's avatar and the referent, Ken simply switched to the third-person view and was able to see all the relevant information. This shows that using the third-person view can alleviate the problem of fragmentation. The following table summarizes the characteristics of views and displays.

Table 7.2: Characteristics of views and displays.

		Characteristics
Views	First Person	- Zoomed-in effect - Less occlusion - Tended to use when generating gestures - Mostly used with the three-monitor setup
	Third Person	- Generally preferred over first-person view - May overcome some fragmentation problems - Provide more of the scene - Show more of the avatar's arm - May occlude important information at the middle of the screen - Mostly used with the one-monitor setup
Displays	One Monitor	- Generally not preferred - Easy to miss important information
	Three Monitors	- Strongly preferred to one-monitor setup - Bigger horizontal view

7.4 Discussion

The observational study provides answers to the main research questions:

1. *Are free and natural pointing useful in CVEs even when augmented-pointing techniques are available?*

   Yes, free and natural pointing are useful because they generated arm movements that were regularly used as context even when augmented techniques were available, and provide staging actions that were especially important at early stages of pointing.

2. *How does a wide field of view affect the use of distant-pointing techniques?*

   The three-monitor display was strongly preferred, and people switched between first- and third-person views depending on the needs of the situation.

Other findings about the use of different deictic-pointing techniques in CVEs include:

- Augmented techniques were preferred over natural arm-based pointing;
- The pointing techniques differ substantially in terms of controllability, accuracy, feedback, visual clutter, and ownership;
- The laser beam technique was preferred overall because of its specificity, visibility, and clear connection to the gesturing avatar;

- The increased specificity of augmented pointing was used by participants to simplify their verbal references.

7.4.1 Natural and Augmented Pointing

In the study, I tested five different kinds of pointing techniques—natural, long arm, laser, spotlight, and highlight. People preferred all of the versions that provided extra information, primarily because of the added specificity that this information made possible. However, other features were also important: for example, people also liked the laser because it visually connected the referent to the gesturer.

While augmented pointing techniques helped referencing in many situations, these techniques do not solve the entire problem—they are not a replacement for the context provided by staging actions (such as avatar position, direction, and arm movement), or for the disambiguation capabilities of speech. All these communication channels remained critical to the success of collaboration. Collaborators relied on different resources at different stages of referencing conduct. Before the onset of pointing actions, observers paid close attention to the locations and orientations of gesturers. When the gesturer started to point, the arm direction gave additional information about the direction of the referent, and observers often adjusted their views so that they could see both the arm and the potential referent. Once the gesturer activated a technique, observers would often change their view to focus more closely on the visual effects.

Augmented pointing could improve the referencing process by providing additional information. Gesturers used the techniques to precisely indicate referents, and the visual effects helped observers locate what gesturers wanted to show. However, most of the benefit of using augmented pointing came at the later stages of referencing. Observers still needed to pay attention to the avatar's arm direction at the earlier stages of pointing, suggesting that augmented pointing cannot entirely replace free arm movements. Furthermore, collaborators still needed to talk to each other even when they were using augmented pointing: the techniques helped simplify verbal descriptions, but collaborators still relied heavily on speech.

7.4.2 View Extents, Field of View, and Viewing Perspectives

Knowing what collaborators could see was often critically important. Several episodes showed that collaborators wanted to know their partner's view: observers moved their avatars close to the gesturer, even to the point of overlapping. By knowing what was in the gesturer's view, the observer could more easily understand both the pointing gesture and the verbal conversation.

The study explored the value of a wide display, and showed that the three-monitor view improved people's ability to coordinate activities in pointing-based communication. Furthermore, participants made frequent use of the third-person view, and with it were able to overcome some of the fragmentation problems that have been seen in earlier research.

However, the third-person view is not a complete solution. Although collaborators can see more objects, those objects are smaller and harder to see. Also, collaborators' own avatars occlude the middle of the screen, which is the most important area because collaborators often adjust the views to put the referent at the center. Last, there is no visual feedback to indicate to others which view is being used; misunderstanding about what the other person could see led to several problems in the tasks.

7.5 Lessons

There are five main lessons learned from the study. These can help designers support distant pointing in CVEs.

1. ***Free arm movement.*** People paid a large amount of attention to avatar arm movements, even when augmented pointing techniques were used. Being able to move the arm provided a staging action that helped collaborators predict the general direction of a referent. Having the arm indicate direction as well as the augmentation simplified verbal descriptions of referents in some episodes.

2. ***Large field of view.*** All but one participant preferred using the three-monitor to the one-monitor setup, and all made frequent use of the third-person view. A larger field of view and more screen real estate greatly helped collaborators stay aware of each other's actions, knowledge that was crucial to smooth communication.

3. ***Clear connection between the gesturer and the referent.*** A main reason for the laser's success was that it linked the gesturer and referent. The explicit connection made referents easier to identify, and reduced the problem of determining which beam belonged to which participant.

4. ***View awareness***. There were several episodes where confusion occurred because collaborators did not know what the other person could see. Providing an indication of the other person's view (first- or third-person) could avoid some common communication errors.

5. ***Clear pointing and non-pointing states.*** People need to know when others are pointing or not; adding a simple method for lowering an avatar's arm would reduce confusion about inadvertent and unintended pointing.

Overall, the study suggests that if designers want to support deictic pointing in CVEs, the best set of techniques will be a combination of free-arm pointing for context and staging, and the laser beam for added specificity. In addition, wide displays and third-person views should be provided to improve the visibility and interpretability of staging actions.

CHAPTER 8
GENERAL DISCUSSION

In this chapter, I provide a summary of the main findings, and discuss overall progress on the main research problem and the importance of different pointing techniques. I then provide a set of design guidelines for distant pointing in CVEs. Finally, I discuss limitations and generalizability of the findings and the design recommendations.

8.1 Summary of Main Findings

In the first study, I observed how people point at distant referents and interpret others' pointing gestures in the real world. From the observations, I identified five important aspects of distant pointing (i.e., accuracy requirements, types of pointing, speech, field of view, and avatar actions) that can be applied in CVEs. I then addressed these issues in my subsequent studies.

To examine if natural pointing has sufficient accuracy to be used in CVEs, I conducted the second study to compare pointing accuracy in the real world and a CVE. The main finding was that we can interpret pointing direction in CVEs almost as well as we can in the real world, suggesting that natural pointing can be successful especially in situations where pointing does not need high accuracy.

The third study was conducted to determine whether people can control free pointing together with other avatar actions. I compared five input configurations based on commonly-available input devices. The main finding was that people are able to control free pointing while controlling the movement, orientation, and view direction of an avatar, and that the mouse and the Wii configurations are the best overall.

After determining that natural pointing is accurate enough to be used in CVEs and that free pointing can be controlled along with basic avatar actions, I conducted the fourth study to determine if free and natural pointing are useful in realistic collaborative settings. I observed how collaborators communicate using free, natural, and augmented pointing in a CVE with

realistic collaborative tasks. There were two main findings. First, the laser beam was the most preferred pointing technique because of its specificity and visual connection between referents and gesturers. Second, free and natural pointing are useful even when augmented pointing techniques are available, and that they are particularly important at early stages of pointing.

8.2 Progress on the Original Research Problem

The research problem I addressed in this dissertation is that *pointing in CVEs is limited in comparison with pointing in the real world*. In particular, pointing in CVEs is limited in terms of generation, control, and observation.

Pointing in current CVEs is generally created with fixed movement, i.e., instead of continuous and gradual generation of pointing gestures that we use in the real world, pointing is often created immediately with discrete movement in CVEs (e.g., command-based pointing). After a pointing command is executed, the avatar generates a pointing gesture. The users, however, generally have no control over the speed and direction of the gesture. They cannot pause the gesture, change its speed, or adjust the pointing direction.

Controlling pointing gestures is another limitation in current CVEs. In the real world, controlling an arm and index finger to point is generally easy. However, manipulating these movements in CVEs is much more difficult because it requires controlling at least two more degrees of freedom for the arms and hands. Existing input devices would not work because of the extra controls. In addition, users are already busy with other avatar actions (e.g., movement and view controls).

The third limitation is that pointing in CVEs is harder to observe than in the real world. In CVEs—especially desktop CVEs—the field of view is much smaller that that in the real world, making pointing gesture much harder to see. Observing a pointing gesture via a monitor (where the gesture generally only appears in a small portion of the screen) is undoubtedly more difficult than observing someone points in real life. It is even more difficult when the observer needs to pay attention to the pointing gesture as well as the referent.

To solve the research problem, I developed pointing techniques for improving the expressiveness of pointing gestures in CVEs. The pointing techniques are *free*, *natural*, and *augmented* pointing.

Users are able to move the avatar arm freely without restrictions from other actions of the avatar (free pointing), to point using avatar arm movements without extra visual effects (natural pointing), and to use pointing techniques that have extra visual aids (augmented pointing). With these three kinds of pointing, users are no longer limited to fixed pointing gestures and are able to control pointing speed and directions.

In addition, I reconfigured five commonly-available input devices (mouse, trackball, gamepad, joystick, and Wii controls) to control an avatar's pointing direction. With these input devices, users can control pointing together with avatar's movement, orientation, and view direction. While the mouse and the Wii configurations are consistently better than other settings, there may be other input methods (e.g., using motion sensors such as the Kinect) that could outperform the mouse and Wii controls. Although the configurations tested in this research may not be the perfect solution to the problem of control limitation in CVEs, they provide extra control—arm movement—to widely-available devices for pointing.

For the issue of gesture visibility in CVEs, I showed that pointing gestures can be easier to observe by increasing field of view with the multi-monitor setup and the third-person view. Multiple monitors provide a wider field of view making it easier for the observer to see pointing gestures and the referents without changing view direction. The third-person view also let users see more of the scene including the users' avatar, making it easier to see others' and their own pointing gestures. While these methods are not a perfect solution (e.g., the user's avatar located in the centre of the screen in third-person view and may occlude important information), they make observing pointing gestures much easier.

8.3 Importance of FN and FA Pointing

Two main kinds of pointing were used in my studies: free-and-natural (FN), and free-and-augmented (FA). They were shown to be useful in supporting collaborations in CVEs. The two kinds of pointing have different properties, and so one is better the other in different situations. In this section, I discuss the importance of FN and FA pointing.

8.3.1 Free Arm Movements

Traditionally pointing in CVEs can only be performed with limited arm movement. For example, pointing in most first-person shooter (FPS) games has no arm movement and avatars can only point at the centre of the view—i.e., they point by adjusting the location and view orientation of an avatar who is holding a weapon that always points at the centre of the screen. There are other CVEs that support pointing gestures with only fixed arm movements (e.g., pointing by typing a command such as '/point' in World of Warcraft) that raise the arm straight to the front of the avatar.

Unlike traditional pointing methods, both FN and FA pointing allow users to move the avatar arm freely without these restrictions. Having free arm movement and being able to use it in concert with moving, turning, and looking are critical. Here, I provide three typical situations in CVEs to illustrate why they are important. The situations are oriented around a typical game scenario called a 'collaborative escort mission' (Figure 8.1). The objective of the mission is for two players to escort an in-game character (the VIP) from one location to another. There are several requirements to an escort mission: the players need to move from the starting location to the destination, defend against enemies on the way, and ensure the VIP's safety. The following situations describe episodes where one player must communicate referents to their partner. In current versions of FPS games, this communication needs to take place through a combination of weapon-based pointing and verbal descriptions, neither of which is likely to be as successful as FN and FA pointing.

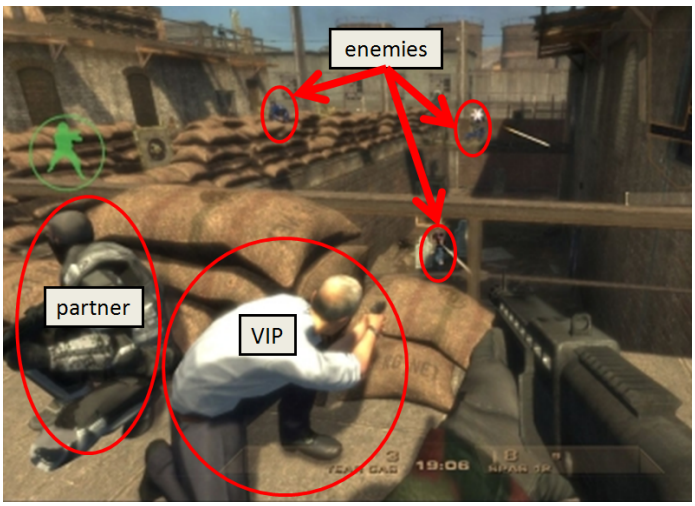


Figure 8.1: First-person view in an escort mission (Rainbow Six).

The first situation is about pointing at different targets in the same view. Multiple enemies are often located at different places on the screen simultaneously. While keeping an eye on the VIP, the player must both point out targets to the partner, and shoot at some of the enemies (Figure 8.1). It is important to maintain a certain viewing angle to keep all of the targets, the VIP, and the partner in view at the same time. With FN and FA pointing, the player is able to move the avatar's arm freely and point at different targets within the view. This type of action is not possible with common pointing methods where weapon-based pointing is tied to the center of the view.

Pointing at targets while moving is also crucial during an escort mission. Sometimes it is dangerous to stop or even slow down. Players must keep moving while at the same time keeping track of the VIP and pointing out enemies. This situation requires that the player be able to point freely with the avatar while also looking, moving, and turning to follow a path or track the VIP.

Another situation is that the player needs to point at targets while changing the view. After arriving at a relatively safe location during the mission, the player may need to communicate the locations of multiple enemies that are spread out over a wide area. This situation requires pointing with the avatar while also moving the view (both horizontally and vertically) to cover the entire range where enemies are located.

8.3.2 Pointing without Additional Visual Effects

The primary difference between FN and FA pointing is that the gesture of FN pointing has no visual effect other than the movement of the gesture. It was shown to be good for indicating referents that do not require high accuracy (e.g., distinct or far-away referents, and general directions) and is particularly important at early stages of pointing. Also, FN pointing may be more preferable depending on the purpose of CVEs where naturalness is more important than accuracy.

In both the real world and CVEs, observers shift their attention during the four stages of distant pointing (i.e., orientation, preparation, production, and holding) from gesturers to referents. During orientation, the gesturer orients themselves so that the observer is able to see the referent and the pointing gesture that is about to be made. Knowing that the gesturer is going to generate

a pointing gesture, the observer pays attention to the gesturer's arm and hand to look for preparatory motions such as taking the hand out of the pocket. Then the focus is shifted from the gesturer's arm to the referent as the arm is pointing towards the referent during the production stage. Finally, the observer mainly focuses on the referent when the gesturer is holding the gesture.

One reason why FN pointing is good for early stages of pointing is that it has no extra visual effects to avatar arm movements. This lets observers focus on the avatar and the arm movements, which is where the attention should be during orientation and preparation. In addition, gesturers only need to let observers know the general direction of the referents during these two stages. FN pointing provides the right amount of information to the observers without being too specific. Also, without extra visual effects, it is clear to observers that pointing gestures are still at the orientation and preparation stage.

Furthermore, naturalness of pointing gestures may be more important than pointing accuracy in some CVEs. For example, adventure games that emphasise atmosphere and storytelling may require more naturalness but less accuracy than FPS games that focus on shooting at targets. FN pointing would be more suitable for the former than the latter.

8.3.3 Pointing with Additional Visual Effects

With the addition of visual effects, FA pointing can provide more specific indication of referents than FN pointing, and is more suitable to be used at the later stages of pointing (i.e., production and holding).

As discussed earlier, observers' attention switches from gesturers to referents during production of a gesture and mainly stays on the referents during the holding stage. Most augmented effects (e.g., laser beam and object highlighting) provide visual effects that make referents more obvious and easier to see. Therefore, FA pointing is more suitable for the later pointing stages where the observers need to know exactly where referents are. In addition, because referents are more prominent, much time can be saved from typing or speaking detailed descriptions of referents, making collaboration more effective.

Some visual effects convey more information than just indicating referents. For example, a laser beam connects the referent to the gesturer. The observer not only knows where the referent is, but also knows who is pointing at the referent. However, this kind of FA pointing is not suitable for orientation and preparation stages because the visual effects can mislead the observer into thinking that the gesturer is pointing at the referent before the gesture is produced.

8.4 Design Guidelines

In Chapter 3, I presented design guidelines based on problems of distant pointing in existing CVEs. Here, I provide a more comprehensive list of design guidelines that incorporate the lessons learned from the four studies I conducted. The guidelines have four main parts: arm movement, controls, augmented techniques, and other supports.

Arm Movement

1. ***Free arm movement.*** Like pointing in the real world, avatars should be able to point in any direction in a CVE. As discussed earlier, pointing only at the centre of the screen or where an avatar is facing is inconvenient and can cause problems in many situations. CVE designers should provide mechanisms to support free arm movement. (Based on Sections 3.6, 6.9.1, and 7.5)

2. ***Avoid unintentional pointing.*** Whether someone is pointing or not is generally obvious in the real world. It takes effort to hold the arm to point but almost no effort to lower it after pointing. It can be the opposite in some CVEs (i.e., no user interaction is needed for holding the arm, but requires user input to lower the arm). For example, deselecting a highlighted object in Second Life can only be done by clicking somewhere else on the screen, and lowering the arm with free pointing can only be done by moving the mouse down. Gesturers may forget to retract the pointing gesture and cause confusion. To avoid this, effort could be added for holding the arm. For example, constantly pressing a button to raise the avatar arm. When the button is not pressed, the arm would automatically lower. (Based on Section 7.5)

3. ***Incorporating pointing with other avatar actions.*** While pointing is important for referential communication, other avatar actions cannot be ignored. Actions such as

changing avatar's location, orientation, and view direction help convey referential information and smooth collaboration. For example, gesturers are more likely to point at referents in front of them and in their field of view. Observers can then get a general idea of where referents are by looking at the location, orientation, and view direction of the gesturer before pointing gestures are produced. (Based on Sections 4.4 and 6.9.1)

Controls

4. ***Speed and direction.*** Gesturers should have continuous control of the speed and direction of pointing gestures. It is critical to synchronize pointing gestures with other communicational conduct such as speech and view direction. Without continuous control over pointing speed and direction, it is difficult to have coherent verbal and gestural communication. (Based on Sections 3.6 and 4.4)

5. ***Easy to generate.*** Generating and controlling pointing gestures should be intuitive and easy to use. Common input devices (e.g., a mouse) and direct input devices (e.g., a Wiimote) can be good for generating pointing gestures with intuitive mapping: for example, pointing up by moving the mouse forward or the Wiimote up. Gesturers should not need to memorize pointing commands or navigate through menus to create pointing gestures. (Based on Sections 3.6 and 6.9.1)

Augmented techniques

6. ***Varying accuracy.*** Because pointing accuracy requirements vary depending on the situation (e.g., showing distinct referents does not need high pointing accuracy, but identifying a referent within a group of similar objects does), an avatar should be able to point with the appropriate accuracy level. Augmented pointing can provide such flexibility. For example, a laser beam is more accurate than a spotlight, which is more accurate than an elongated arm. Designers should provide pointing techniques with different accuracy. (Based on Sections 4.4, 5.5, and 7.5)

7. ***Ownership.*** When using augmented techniques, it is important to know who generates the pointing effects. The ability to identify the ownership of augmented effects becomes increasingly critical as more collaborators use pointing gestures in a CVE. Without proper identification, it is difficult to know which object is being referred to when

multiple augmented effects appear together. Linking referents to gesturers with lines (e.g., a laser beam), colours, or patterns (e.g., effects with unique shapes) can help alleviate the ownership problem. (Based on Section 7.5)

Other supports

8. ***Speech.*** It is natural and useful to provide verbal description while pointing in the real world, and this is the same in CVEs. In addition to providing a basic communication channel such as text chat, CVEs should support verbal communication for more effective referential activities. (Based on Section 4.4)

9. ***Wide field of view.*** Much research has shown that a narrow field of view leads to communication problems in CVEs (Fraser et al., 1999; Hindmarsh et al., 1998, 2000). Providing a wide field of view allows collaborators to see both gestures and referents, to establish mutual orientation, and to be aware of everything that happens in the environment. CVEs should provide wide fields of view by, e.g., supporting multiple-monitor setups or third-person views. (Based on Sections 4.4 and 7.5)

10. ***View awareness.*** CVEs should be able to show what collaborators see. As discussed earlier, gesturers are more likely to point at referents that are within their view, so knowing what they see makes it easier to identify referents. Also, view awareness is particularly important when collaborators can switch between first- and third-person views. Not knowing which views collaborators are using can cause much confusion, as seen in Chapter 7. CVE designers should provide view awareness mechanisms, such as a wire frame of collaborators' view frustum (Fraser et al., 1999; Hindmarsh et al., 1998, 2000), a popup window showing what others can see, or a function to temporarily switch views between collaborators. (Based on Section 7.5)

8.5 Limitations and Generalizability

This research focuses on distant pointing in desktop CVEs. In the studies, there were one or two collaborators using the CVE simultaneously; the scenes were a room with artificial targets or cityscape from a balcony; and input devices were commonly available with new mappings. The research focus and study designs have direct influences on the findings and design

recommendations. Here, I discuss limitations based on the types of pointing, CVEs, group size, tasks, and input devices; then discuss generalizability.

8.5.1 Pointing at Nearby Referents

Distant pointing is the primary focus of this research. Arm movements and augmented techniques used in the studies were designed for distant pointing. The findings may not apply to non-distant pointing (i.e., pointing at referents within reach). For example, the laser beam may not be the most useful. One main reason why it was considered the best technique is that it creates a visual link between a referent and the gesturer. However, this link is unnecessary if the gesturer is close enough to touch the referent. Perhaps natural pointing would be the best for pointing at reachable referents because gesturers no longer need to activate and deactivate augmented techniques.

In addition, pointing with a straight arm is insufficient for close-up pointing. With only a straight arm, gesturers cannot point at themselves. A bendable elbow (and maybe wrist) becomes necessary.

8.5.2 Other CVEs

A desktop CVE was used in all the studies. When other types of CVEs (e.g., immersive CVEs) are used, some findings could be affected. In particular, findings related to field of view are likely to be different. Users of immersive CVEs usually either wear head-mounted displays or are surrounded by walls of projected images (Figure 8.2). Generally, this type of CVE gives users much bigger fields of view than what desktop CVEs can provide. The main benefit of using a third-person view in a desktop CVE is to have a larger field of view. This benefit, however, is nullified by the large field of view in immersive CVEs.

A  B 

Figure 8.2: Immersive CVE settings: A) with a head-mounted display; B) surrounded by walls of displays.

8.5.3 Group Size

Some augmented techniques and design recommendations may not be optimal for CVEs with more than two people. When multiple augmented effects (e.g., highlight, spotlight, and laser) appear, it is critical to be able to identify which effect belongs to whom. Using unique colours and patterns can help in identifying ownership. However, as the number of people in a CVE rises, these identification methods become less effective—it is more difficult for CVE designers to assign unique colours and patterns, and for CVE users to remember the colours and patterns of their collaborators.

Also, problems that do not exist in two-person CVEs with augmented pointing may arise in CVEs with more people. Visual clutter is one example. In two-person CVEs, there are at most two augmented effects appearing at the same time. If more people use a CVE and more visual effects in the scene, the augmented techniques not only become less effective, but also create distractions that can hinder collaboration.

8.5.4 Tasks

The tasks used in the studies did not put people in situations where the visual effects of augmented pointing could be a detriment (e.g., a shooter game). Because the augmented techniques used in the studies publicize the fact that a pointing gesture is being made, using these techniques maybe improper in those situations. People may also develop different strategies for using different techniques.

Also, the realistic scene used in the study was on a balcony facing a cityscape. There was limited space for avatar movement and most referents were on one side of the environment. If tasks were set in an open field where people are surrounded by objects and can go anywhere in the field, different dynamics may develop between collaborators. The fragmentation problem (Fraser et al., 1999; Hindmarsh et al., 1998, 2000) may also be more severe because referents can be in any location.

8.5.5 Input Devices

The input devices used in this research were all widely available. As technology improves, more advanced input devices will become available and common. These devices could possibly be more suitable for controlling distant pointing. For example, the Kinect sensor was not available at the time when I was working on the study that compared input devices, but this sensor is now commonly available and could possibly be better than the mouse and the Wiimote because of its ability to track natural movements.

8.5.6 Summary

While there are some findings that are more applicable to the specific settings of the studies and the focus of this research, most of the findings can be generalized to other settings. For example, the important aspects of distant pointing, the accuracy of different pointing techniques, the ability to control pointing with other avatar actions, the characteristics of input configurations for distant pointing, and the importance of free, natural and augmented pointing should be applicable to other CVEs with different group size and tasks.

CHAPTER 9
CONCLUSIONS AND FUTURE WORK

Deictic pointing—pointing at referents during conversations—is ubiquitous in daily activities and is important in real-world communication; it is also important CVEs because CVEs are three-dimensional virtual worlds that resemble the real world. However, pointing in CVEs is very limited compared to how we point in the real world. To address this problem, I designed, developed, and evaluated new pointing techniques for CVEs, and provided design guidelines for improving the expressiveness of pointing gestures in CVEs.

In this dissertation, I focused on distant pointing—where referents are out of reach—in desktop CVEs. I first presented a framework that describes the stages and enactment of distant pointing based on previous work. The framework helps guide the design and development of pointing techniques used in this research. I observed how people point at distant referents in the real world and identified important aspects of distant pointing. I then designed and developed distant-pointing techniques based on the insights gained from the observational study, and built a CVE that incorporates the pointing techniques as the bases for the subsequent studies. I verified that natural pointing has sufficient accuracy to be used in desktop CVEs by comparing pointing accuracy in the real world and a CVE. I then reconfigured commonly-available input devices for controlling free pointing, determined the best input devices for free pointing, and verified that people can control free pointing together with other avatar actions. I conducted an observational study comparing different distant pointing techniques, and verified that free and natural pointing are useful and important in CVEs even when augmented techniques are available. Finally, I provided a set of guidelines for designing distant pointing in desktop CVEs.

9.1 Contributions

The main contribution of this dissertation is the design and evaluation of distant-pointing techniques that improve the expressiveness of pointing gestures in desktop CVEs. I developed free pointing that allows avatars to have free pointing movement that is independent from other

avatar actions. Combining free with natural and augmented pointing, avatars can generate pointing gestures that were not previously possible.

This work also has the following minor contributions:

- A framework that can be used to guide the design and development of distant pointing. (Chapter 3)
- Identification of important aspects of distant pointing from observational work. (Chapter 4)
- Verification that natural pointing is accurate enough to be used in desktop CVEs. (Chapter 5)
- Determination of the best way to control free pointing among five commonly-available input devices that have new input mappings. (Chapter 6)
- Verification that free pointing can be controlled together with other avatar actions. (Chapter 6)
- Verification that free and natural pointing are useful and important for distributed collaboration even when augmented techniques are available. (Chapter 7)
- A set of guidelines for designing distant pointing in desktop CVEs. (Chapter 8)

9.2 Future Work

To further improve pointing expressiveness in CVEs and to expand the scope of the audience who can benefit from this work, the limitations listed in the previous chapter must be addressed. Below, I list six future research directions that can extend this work.

Pointing at nearby referents. Pointing at referents within reach is extremely common. Thus, extending this research to include nearby referents is valuable. To do that, more flexible gestures (e.g. with a bendable elbow, a rotatable wrist, and movable fingers) need to be explored.

Immersive CVEs. As technology advances, immersive CVEs can become more affordable and common. Exploring different pointing techniques that can take advantage of the large field of view of immersive CVEs is an important future direction.

CVEs with many users. Common CVEs allow many users to interact with each other at the same time. Pointing techniques used in this research may no longer be suitable for such CVEs. For example, ownership of augmented effects will be increasingly difficult to determine as the number of users increases; visual clutter caused by the effects will also be amplified. New pointing techniques need to be designed for CVEs with many users.

Visibility control. Pointing techniques used in this research can be seen by all users in the CVE. However, this may not be desirable in all situations. For example, publicizing a pointing gesture in first-person shooter games can put the gesturer in danger. Controlling who can see a pointing gesture can be critical and is worth exploring.

Input devices. This research uses commonly-available input devices for controlling pointing gestures with the purpose of benefiting a large number of people. Advanced, expensive, and uncommon devices will gradually become more affordable and widely available. Using different input devices, such as motion sensors, is valuable to explore.

Other gestures. With more advanced input devices, more flexible gestures (with movable fingers and wrists) can be generated. This can help exploring complex gestures, such as iconic gestures, that were shown to be useful in referential communication (as discussed in Chapter 4).

By following these future research directions, pointing gestures can become more expressive, referential communication in CVEs can be much easier, and more people can benefit from the work in this dissertation.

GLOSSARY

Absolute input: each point on the input space corresponds to a point in output space.

Adapters: objects that change one's focus and nimbus (Benford et al., 1994; Benford & Fahlen, 1993) (e.g., a loudspeaker increases one's nimbus).

Augmented pointing: a type of pointing that has additional visual effects.

Aura: boundary of the presence of an object (Fahlén & Brown, 1992) (a concept of a spatial model of awareness).

Avatars: three-dimensional representations of the users and are commonly shown as human-like shapes.

Awareness: "an understanding of the activities of others, which provides a context for your own activity" (Dourish & Bellotti, 1992, p. 107).

Beats: hand movements that move along with the rhythm of the speech and have only two movement phases (McNeill's classification (1992)).

Boundaries: dividers of areas that affect the properties of aura, focus, nimbus, and interactions between objects (Benford et al., 1994) (e.g., walls and windows).

Cohesive gestures: gestures that are used to link together temporarily separated parts of a discourse that are within the same theme (McNeill's classification (1992)).

Collaborative virtual environments (CVEs): computer generated three-dimensional worlds that resemble the real world, and allow people to interact with one another and objects in the environment via their avatars.

Computer supported cooperative work (CSCW): a field that has a multidisciplinary nature and "covers anything to do with computer support for activities in which more than one person is involved." (Bannon & Schmidt, 1989, p. 359)

Degree of freedom: the number of independent ways that can change the space configuration of a mechanical system.

Deictic gestures: pointing gestures for indicating objects or events either concrete or abstract (McNeill's classification (1992)).

Deictic pointing: a pointing gesture that provides a deictic reference.

Deixis: a reference to a thing that is relevant to the context of an utterance.

Desktop CVE: a CVE that is setup in a desktop environment.

Direct input: implies that the input space is the same as the output space (e.g., touch screens or Wii remotes).

Distant pointing: a type of deictic pointing where the referent is out of reach.

Emblems: well-formed gestures that need to be performed in some specific ways (part of Kendon's continuum (1988)).

Enactment of distant pointing: describes how distant pointing occurs. The enactment is characterized in terms of movement and visual effect in this dissertation.

Expressive phase: a phase of a pointing gesture that contains either a stroke or a stroke-less hold, also called independent hold (Kita et al., 1998).

Focus: the attention of observers (Benford et al., 1994; Benford & Fahlen, 1993) (a concept of a spatial model of awareness).

Fragmentation: describes the screen cannot display all the relevant things needed for communication (Hindmarsh et al., 1998).

Free pointing: a type of pointing that is independent from other avatar actions.

Gesticular phrase (G-Phrase): a part of G-Unit that is composed by different phases (i.e., preparation, stroke, and recovery) (Kendon, 1980).

Gesticular unit (G-Unit): a part of pointing gestures that starts when a limb moves from a rest position (e.g., one's lap or the arm rest of a chair) and ends when the limb moves back to another rest position. A G-Unit contains one or more gesticular phrases (Kendon, 1980).

Gesticulation: idiosyncratic spontaneous hand and arm movements during speech (part of Kendon's continuum (1988)).

Hand internal preparation phase: a phase of a pointing gesture that is for shaping and orienting the hand for an expressive phase (Kita et al., 1998).

Head-mounted display (HMD): a helmet-like device with a display in front of each eye.

Iconic gestures: gestures that depict the appearances of objects or actions of events, and have a close formal relationship to the semantic content of speech (McNeill's classification (1992)).

Immersive CVEs: CVEs that give users a feeling of being in a virtual environment. Users of immersive CVE usually use motion-tracking devices along with a head-mounted display (HMD) or a spatially immersive display (SID).

Index: one of Peirce's classes of signs (Buchler, 1955) that has a physical connection to the object of interest.

Indicating: the method of signaling for an index.

Indirect input: the input and output spaces are separated (e.g., the mouse and keyboard).

Kendon's continuum: an ordering of hand gestures (i.e., gesticulation, language-like gestures, pantomimes, emblems, and sign languages).

Language-like gestures: gestures that are grammatically integrated in the utterance (part of Kendon's continuum (1988)).

Liberating movement phase: a phase of a pointing gesture that is for freeing the hands from some constrained locations (Kita et al., 1998).

Location preparation phase: a phase of a pointing gesture that moves the arm to the starting position of an expressive phase (Kita et al., 1998).

Metaphoric gestures: pictorial gestures that show abstract ideas (McNeill's classification (1992)).

Mutual orientation: the orientations of people in a conversation. With mutual orientation, people can see each other and the referents.

Natural pointing: a type of pointing that does not have any visual effect other than the movement of the gesture.

Nimbus: the projection of the information from the person being observed (Benford et al., 1994; Benford & Fahlen, 1993) (a concept of a spatial model of awareness).

Non-immersive CVEs: CVEs that primarily use a desktop display with a mouse and a keyboard.

Omni-directional locomotion system (ODLS): an input device that allows users to control avatar's movement by walking in any direction in-place.

Pantomimes: gestures without speech (part of Kendon's continuum (1988)).

Position control devices: input devices that control the position of an object (e.g., a cursor).

Post-stroke hold: a period to extend the period of a stroke (Kita, 1990; McNeill, 1992).

Preparation: a part of G-Phrase that is the movement of a limb from a rest position to the beginning of a stroke (Kendon, 1980).

Pre-stroke hold: a period where the gesture waits for the speech to occur (Kita, 1990; McNeill, 1992).

Rate control devices: input devices that control the speed of an object (e.g., a cursor).

Ray casting: a visual technique of projecting a ray from the user's hand or input devices.

Recovery: a part of G-Phrase that describes the limb moving back to a rest position or becomes ready for another stroke (Kendon, 1980).

Referent: the thing being referred to.

Relative input: the input and output space are offset with variable mapping.

Restricted pointing: a type of pointing that is restricted by other avatar actions.

Sign language: a set of gestures that has a full linguistic system (part of Kendon's continuum (1988)).

Signal: "the presentation of a sign by one person to mean something for another." (Clark, 1996, p. 160)

Situation awareness: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." (Endsley, 1988, p. 792)

Spatially immersive display (SID): a type of display that usually has four to six walls arranged like a cube. Images are projected to the walls.

Stages of distant pointing: orientation, preparation, production, and holding.

Stroke: a part of G-Phrase that is an accented movement with a distinct peak of effort in the sense of dance movements (Dell, 1977; Kendon, 1980).

Telepointers: replicated pointers that track the locations of other users' mouse cursors.

Virtual cursor: a 3D cursor in a virtual environment that allows users to control its location in 3D (Hinckley et al., 1994).

Virtual environments: computer generated three-dimensional environments that resemble the real world.

Workspace awareness: "the up-to-the-moment understanding of another person's interaction with a shared workspace" (Gutwin & Greenberg, 2002, p. 412).

REFERENCES

Bangerter, A. (2004). Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue. *Psychological Science*, *15*(6), 415–419.

Bannon, L. J., & Schmidt, K. (1989). CSCW: Four Characters in Search of a Context. In *Proceedings of the First European Conference on Computer Supported Cooperative Work* (pp. 358–372).

Bateman, S., Doucette, A., Xiao, R., Gutwin, C., Mandryk, R., & Cockburn, A. (2011). Effects of View, Input Device, and Track Width on Video Game Driving. In *Graphics Interface 2011* (pp. 207–214). St. John's, Canada.

Bekker, M. M., Olson, J. S., & Olson, G. M. (1995). Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. In *Proceedings of the 1st Conference on Designing Interactive Systems: Processes, Practices, Methods, & Techniques* (pp. 157–166). Ann Arbor, Michigan, United States: ACM.

Benford, S., Bowers, J., Fahlen, L. E., & Greenhalgh. (1994). Managing Mutual Awareness In Collaborative Virtual Environments (pp. 223–236).

Benford, S., Bowers, J., Fahlén, L. E., Greenhalgh, C., & Snowdon, D. (1995). User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '95* (pp. 242–249). Presented at the the SIGCHI conference, Denver, Colorado, United States.

Benford, S., & Fahlen, L. (1993). A Spatial Model of Interaction in Large Virtual Environments (pp. 109–124). Kluwer Academic Publishers.

Benford, S., Greenhalgh, C., & Lloyd, D. (1997). Crowded collaborative virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '97* (pp. 59–66). Presented at the the SIGCHI conference, Atlanta, Georgia, United States.

Benford, S., Greenhalgh, C., Rodden, T., & Pycock, J. (2001). Collaborative virtual environments. *Communications of the ACM*, *44*(7), 79–85.

Bessière, A. F., Seay, A. F., & Kiesler, S. (2007). The ideal elf: identity exploration in World of Warcraft. *Cyberpsychology and Behavior*, *10*(4), 530–535.

Biocca, F., & Delaney, B. (1995). Immersive Virtual Reality Technology. In *Communication in the Age of Virtual Reality* (pp. 57–124). Routledge.

Bouguila, L., Ishii, M., & Sato, M. (2002). Realizing a new step-in-place locomotion interface for virtual environment with large display system. In *Proceedings of the Workshop on Virtual Environments 2002* (pp. 197–207). Aire-la-Ville, Switzerland, Switzerland: Eurographics Association.

Bowers, J., Pycock, J., & O'Brien, J. (1996). Talk and embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Common Ground - CHI '96* (pp. 58–65). Presented at the the SIGCHI conference, Vancouver, British Columbia, Canada.

Bowman, D. A., Datey, A., Ryu, Y. S., Farooq, U., & Vasnaik, O. (2002). Empirical Comparison of Human Behavior and Performance with Different Display Devices for Virtual Environments (pp. 2134–2138).

Buchler, J. (1955). *Philosophical Writings of Peirce*. Dover Publications.

Clark, H. H. (1996). Signaling. In *Using Language* (pp. 155–188). Cambridge University Press.

Clark, H. H. (2003). Pointing and placing. In *Pointing: Where Language, Culture, and Cognition Meet* (pp. 243–268). Hillsdale, NJ: Erlbaum.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 245–258.

Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 135–142). New York, NY, USA: ACM.

Darken, R. P., Cockayne, W. R., & Carmein, D. (1997). The omni-directional treadmill: a locomotion device for virtual worlds. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology* (pp. 213–221). New York, NY, USA: ACM.

Dell, C. (1977). *A primer for movement description using effort-shape and supplementary concepts*. (New York): Dance Notation Bureau, Center for Movement Research and Analysis, Bureau Press.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM Conference on Computer-supported Cooperative Work* (pp. 107–114). New York, NY, USA: ACM.

Draper, J. V., Kaber, D. B., & Usher, J. M. (1998). Telepresence. *Human Factors*, *40*(3), 354–375.

Ducheneaut, N., Wen, M.-H., Yee, N., & Wadley, G. (2009). Body and mind: a study of avatar personalization in three virtual worlds. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI '09* (pp. 1151–1160). Presented at the the 27th international conference, Boston, MA, USA.

Duchowski, A. T., Cournia, N., Cumming, B., McCallum, D., Gramopadhye, A., Greenstein, J., … Tyrrell, R. A. (2004). Visual deictic reference in a collaborative virtual environment. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications* (pp. 35–40). San Antonio, Texas: ACM.

Dyck, J., Gutwin, C., Subramanian, S., & Fedak, C. (2004). High-performance telepointers. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work - CSCW '04* (p. 172). Chicago, Illinois, USA.

Ekman, P. (1992). Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, *3*(1), 34–38.

Ellis, S. R. (1994). What are virtual environments? *Computer Graphics and Applications, IEEE*, *14*(1), 17–22.

Ellis, S. R. (1995). Virtual Environments and Environmental Instruments. In *Simulated and Virtual Realities - Elements of Perception* (pp. 85–101). Taylor & Francis.

Endsley. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32–64.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National* (pp. 789–795). Presented at the Aerospace and Electronics Conference, 1988. NAECON 1988, IEEE.

Etiquette in Asia. (2012, December). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Etiquette_in_Asia

Fabri, M., & Moore, D. (2005). The use of emotionally expressive avatars in collaborative virtual environments. In *Proceeding of Symposium on Empathic Interaction with Synthetic Characters, at Artificial Intelligence and Social Behaviour Convention 2005* (pp. 88–94).

Fabri, M., Moore, D., & Hobbs, D. (2004). Mediating the expression of emotion in educational collaborative virtual environments: an experimental study. *Virtual Reality*, *7*(2), 66–81.

Fahlén, L. E., & Brown, C. (1992). The Use of a 3D Aura Metaphor for Computer Based Conferencing and Teleworking. In *In Proc. 4th Multi-G Workshop* (pp. 69–74).

Fraser, M., & Benford, S. (2002). Interaction effects of virtual structures. In *Proceedings of the 4th International Conference on Collaborative Virtual Environments* (pp. 128–134). Bonn, Germany: ACM.

Fraser, M., Benford, S., Hindmarsh, J., & Heath, C. (1999). Supporting awareness and interaction through collaborative virtual interfaces. In *Proceedings of the 12th Annual*

ACM Symposium on User Interface Software and Technology (pp. 27–36). Asheville, North Carolina, United States: ACM.

Fraser, M., Hindmarsh, J., Benford, S., & Heath, C. (2004). Getting the picture: Enhancing avatar representations in collaborative virtual environments, *29*(4), 133–150.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., & Kramer, A. D. I. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Hum.-Comput. Interact.*, *19*(3), 273–309.

Garau, M., Slater, M., Bee, S., & Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 309–316). New York, NY, USA: ACM.

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., & Sasse, M. A. (2003). The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 529–536). New York, NY, USA: ACM.

Goodwin, C. (1981). *Conversational Organization: interaction between speakers and hearers*. Academic Press.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*(10), 1489–1522.

Goodwin, C., & Goodwin, M. (1996). Seeing as a situated activity: Formulating planes. In Y. Engeström & D. Middleton (Eds.), *Cognition and Communication at Work* (pp. 61–95). Cambridge University Press.

Greenberg, S., Gutwin, C., & Roseman, M. (1996). Semantic telepointers for groupware. In *Proceedings of the Australian Conference on Computer-Human Interaction* (pp. 54–61).

Greenhalgh, C., & Benford, S. (1995). MASSIVE: a collaborative virtual environment for teleconferencing. *ACM Transactions on Computer-Human Interaction*, *2*(3), 239–261.

Grossman, T., & Balakrishnan, R. (2005). The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 281–290). New York, NY, USA: ACM.

Grudin, J. (1994). Computer-supported cooperative work: history and focus. *Computer*, *27*(5), 19–26.

Gutwin, C., & Greenberg, S. (2002). A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work*, *11*(3), 411–446.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati (Ed.), *Advances in Psychology* (Vol. Volume 52, pp. 139–183). North-Holland.

Heath, C. (1986). *Body movement and speech in medical interaction*. (K. Nicholls, Ed.). Cambridge: Cambridge University Press.

Heath, C., & Hindmarsh, J. (2000). Configuring Action in Objects: From Mutual Space to Media Space. *Mind, Culture, and Activity*, *7*(1), 81–104.

Heath, C., & Luff, P. (1991). Collaborative activity and technological design: task coordination in London underground control rooms. In *Proceedings of the Second Conference on European Conference on Computer-Supported Cooperative Work* (pp. 65–80). Amsterdam, The Netherlands: Kluwer Academic Publishers.

Heath, C., Luff, P., Kuzuoka, H., Yamazaki, K., & Oyama, S. (2001). Creating coherent environments for collaboration. In *Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work* (pp. 119–138). Bonn, Germany: Kluwer Academic Publishers.

Hinckley, K. (2003). Input Technologies and Techniques. In J. A. Jacko & A. Sears (Eds.), *The Human-computer Interaction Handbook* (pp. 151–168). Hillsdale, NJ, USA: L. Erlbaum Associates Inc.

Hinckley, K., Pausch, R., Goble, J. C., & Kassell, N. F. (1994). A survey of design issues in spatial input. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology - UIST '94* (pp. 213–222). Presented at the 7th annual ACM symposium, Marina del Rey, California, United States.

Hindmarsh, J., Fraser, M., Heath, C., & Benford, S. (2001). Virtually Missing the Point: Configuring CVEs for Object-Focused Interaction. In *Collaborative Virtual Environments* (pp. 115–133).

Hindmarsh, J., Fraser, M., Heath, C., Benford, S., & Greenhalgh, C. (1998). Fragmented interaction: establishing mutual orientation in virtual environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 217–226). Seattle, Washington, United States: ACM.

Hindmarsh, J., Fraser, M., Heath, C., Benford, S., & Greenhalgh, C. (2000). Object-focused interaction in collaborative virtual environments. *ACM Trans. Comput.-Hum. Interact.*, *7*(4), 477–509.

Hindmarsh, J., & Heath, C. (2000). Embodied reference: A study of deixis in workplace interaction. *Journal of Pragmatics*, *32*(12), 1855–1878.

Iwata, H., & Fujii, T. (1996). VIRTUAL PERAMBULATOR: a novel interface device for locomotion in virtual environment. In *Proceedings of the IEEE Virtual Reality Annual International Symposium* (pp. 60–65).

Jacob, R. J. K. (1996). Human-computer interaction: input devices. *ACM Comput. Surv.*, *28*(1), 177–179.

Jefferson, G. (1984). Transcript Notation. In J. M. Atkinson & J. Heritage (Eds.), *Structures of Social Interaction* (pp. ix–xvi). New York: Cambridge University Press.

Kendon. (1980). Gesticulation and speech: Two aspects of the process of utterance. In *The Relationship Between Verbal and Nonverbal Communication* (pp. 207–227). Mouton Publishers.

Kendon. (1988). How gestures can become like words. In F. Poyatos (Ed.), *Cross-cultural Perspectives in Nonverbal Communication* (pp. 131–141). C. J. Hogrefe.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(1), 22–63.

Kendon, & Versante, L. (2003). Pointing by Hand in "Neapolitan". In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet* (pp. 109–137).

Kirch, M. S. (1979). Non-Verbal Communication Across Cultures. *The Modern Language Journal*, *63*(8), 416–423.

Kirk, D., Crabtree, A., & Rodden, T. (2005). Ways of the hands. In *Proceedings of the Ninth Conference on European Conference on Computer Supported Cooperative Work* (pp. 1–21). Paris, France: Springer-Verlag New York, Inc.

Kirk, D., Rodden, T., & Fraser, D. S. (2007). Turn it this way: grounding collaborative action with remote gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1039–1048). San Jose, California, USA: ACM.

Kita. (2003). Interplay of gaze, hand, torso orientation and language in pointing. In *Pointing: Where Language, Culture, and Cognition Meet* (pp. 307–328). Erlbaum.

Kita, & Essegbey, J. (2001). Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. *Gesture*, *1*(1), 73–95. doi:10.1075/gest.1.1.06kit

Kita, Gijn, I. van, & Hulst, H. van der. (1998). Movement Phase in Signs and Co-Speech Gestures, and Their Transcriptions by Human Coders. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction* (pp. 23–35). Springer-Verlag.

Kita, S. (1990). *The temporal relationship between gesture and speech: A study of Japanese-English bilinguals*.

Kjeldskov, J. (2001). Interaction: Full and Partial Immersive Virtual Reality Displays. In *Proceedings of IRIS24* (pp. 587–600).

Krauss, R., Chen, Y., & Gottesman, R. (2000). Lexical Gestures and Lexical Access: A Process Model. In *Language and Gesture* (pp. 261–283).

Labarre, W. (1947). The Cultural Basis of Emotions and Gestures. *Journal of Personality*, *16*(1), 49–68.

Lee, C., SangWon, G., Park, C., & Wohn, K. (1998). The control of avatar motion using hand gesture. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (pp. 59–65). New York, NY, USA: ACM.

Liang, J., & Green, M. (1994). JDCAD: A Highly Interactive 3D Modeling System. In *Computer and Graphics* (Vol. 18(4), pp. 499–506).

Linebarger, J. M., Janneck, C. D., & Kessler, G. D. (2003). Shared simple virtual environment: an object-oriented framework for highly interactive group collaboration. In *7th IEEE International Symposium on Distributed Simulation and Real-Time Applications* (pp. 170–180).

Luff, P., Yamashita, N., Kuzuoka, H., & Heath, C. (2011). Hands on hitchcock: embodied reference to a moving scene. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems* (pp. 43–52). New York, NY, USA: ACM.

Lyons, J. (1977). Deixis, space and time. In *Semantics* (Vol. 2, pp. 636–724).

Mackinlay, J., Card, S. K., & Robertson, G. G. (1990). A semantic analysis of the design space of input devices. *Hum.-Comput. Interact.*, *5*(2), 145–190.

McNeill, D. (1992). *Hand and mind*. The University of Chicago Press.

Meadows, M. (2007). *I, avatar: the culture and consequences of having a second life* (First.). Thousand Oaks, CA, USA: New Riders Publishing.

Mine, M. (1995). Virtual Environment Interaction Techniques. UNC Chapel Hill Computer Science Technical Report TR95-018.

Moore, R. J., Ducheneaut, N., & Nickell, E. (2007). Doing Virtually Nothing: Awareness and Accountability in Massively Multiplayer Online Worlds. *Comput. Supported Coop. Work*, *16*(3), 265–305.

Moore, R. J., Gathman, E. C. H., Ducheneaut, N., & Nickell, E. (2007). Coordinating joint activity in avatar-mediated interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 21–30). New York, NY, USA: ACM.

Olwal, A., & Feiner, S. (2003). The Flexible Pointer: An Interaction Technique for Selection in Augmented and Virtual Reality. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (pp. 81–82). Presented at the UIST.

Otto, O., Roberts, D., & Wolff, R. (2005). A Study of Influential Factors on Effective Closely-Coupled Collaboration based on Single User Perceptions. In *Proceedings of The 8th Annual International Workshop on Presence* (pp. 181–188).

Otto, O., Roberts, D., & Wolff, R. (2006). A review on effective closely-coupled collaboration using immersive CVE's. In *Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications - VRCIA '06* (pp. 145–154). Hong Kong, China.

Pace, T. (2008). Can an orc catch a cab in stormwind?: cybertype preference in the world of warcraft character creation interface. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (pp. 2493–2502). New York, NY, USA: ACM.

Pausch, R., Proffitt, D., & Williams, G. (1997). Quantifying immersion in virtual reality. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '97* (pp. 13–18).

Peinado, M., Meziat, D., Maupu, D., Raunhardt, D., Thalmann, D., & Boulic, R. (2009). Full-Body Avatar Control with Environment Awareness. *Computer Graphics and Applications, IEEE*, *29*(3), 62–75.

Pettenati, P., Sekine, K., Congestrì, E., & Volterra, V. (2012). A Comparative Study on Representational Gestures in Italian and Japanese Children. *Journal of Nonverbal Behavior*, *36*(2), 149–164.

Poupyrev, I., Billinghurst, M., Weghorst, S., & Ichikawa, T. (1996). The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (pp. 79–80). New York, NY, USA: ACM.

Robertson, G., Czerwinski, M., & Van Dantzich, M. (1997). Immersion in desktop virtual reality. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology* (pp. 11–19). New York, NY, USA: ACM.

Rouse, R. I. (1999). What's your perspective? *ACM SIGGRAPH Computer Graphics*, *33*, 9–12.

Salamin, P., Thalmann, D., & Vexo, F. (2006). The benefits of third-person perspective in virtual and augmented reality? In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (pp. 27–30). New York, NY, USA: ACM.

Salem, B., & Earle, N. (2000). Designing a non-verbal language for expressive avatars. In *Proceedings of the Third International Conference on Collaborative Virtual Environments - CVE '00* (pp. 93–101). Presented at the the third international conference, San Francisco, California, United States.

Salomo, D., & Liszkowski, U. (2012). Socio-cultural settings influence the emergence of prelinguistic deictic gestures. In *Child Development. Advance Online Publication.* doi:10.1111/cdev.12026

Sherzer, J. (1973). Verbal and Nonverbal Deixis: The Pointed Lip Gesture Among the San Blas Cuna. *Language in Society*, *2*(1), 117–131.

Shneiderman, B. (1998). Virtual Environments. In *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (pp. 221–228).

Slater, M., & Steed, A. (2000). A Virtual Presence Counter. *Presence: Teleoperators and Virtual Environments*, *9*(5), 413–434.

Steptoe, W., Oyekoya, O., Murgia, A., Wolff, R., Rae, J., Guimaraes, E., … Steed, A. (2009). Eye Tracking for Avatar Eye Gaze Control During Object-Focused Multiparty Interaction in Immersive Collaborative Virtual Environments. In *Proceedings of the 2009 IEEE Virtual Reality Conference* (pp. 83–90). Washington, DC, USA: IEEE Computer Society.

Steptoe, W., Wolff, R., Murgia, A., Guimaraes, E., Rae, J., Sharkey, P., … Steed, A. (2008). Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (pp. 197–200). New York, NY, USA: ACM.

Tang, J. C. (1991). Findings from observational studies of collaborative work. In *International Journal of Man Machine Studies* (pp. 11–28). Academic Press Ltd.

Tang, & Minneman, S. (1991a). VideoWhiteboard: video shadows to support remote collaboration. In *CHI'91* (pp. 315–322). ACM Press.

Tang, & Minneman, S. L. (1991b). VideoDraw: a video interface for collaborative drawing. *ACM Transactions on Information Systems*, *9*, 170–184.

Tang, Neustaedter, C., & Greenberg, S. (2006). VideoArms: Embodiments for Mixed Presence Groupware. In *Proceedings of the 20th British HCI Group Annual Conference* (pp. 85–102). London: Springer London.

Templeman, J. N., Sibert, L. E., Page, R. C., & Denbrook, P. S. (2007). Pointman - A Device-Based Control for Realistic Tactical Movement. In *IEEE Symposium on 3D User Interfaces, 2007. 3DUI '07.*

Turkay, S., & Adinolf, S. (2010). Free to be me: a survey study on customization with World of Warcraft and City Of Heroes/Villains players. *Procedia - Social and Behavioral Sciences*, *2*(2), 1840–1845.

Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster Trade.

Vanacken, L., Grossman, T., & Coninx, K. (2007). Exploring the Effects of Environment Density and Target Visibility on Object Selection in 3D Virtual Environments. In *IEEE Symposium on 3D User Interfaces, 2007. 3DUI '07* (pp. 117–124).

Virtusphere. (n.d.). Retrieved from www.virtusphere.com

Wadley, G., & Ducheneaut, N. (2009). The "out-of-avatar experience": object-focused collaboration in Second Life (pp. 323–342). Springer London.

Wang, Y., MacKenzie, C., Summers, V., & Booth, K. (1998). The Structure of Object Transportation and Orientation in Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (pp. 312–319).

Wilkins, D. (2003). Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In S. Kita (Ed.), *Pointing: Where Language, Culture, and Cognition Meet* (pp. 171–215).

Wyss, H. P., Blach, R., & Bues, M. (2006). iSith - Intersection-based Spatial Interaction for Two Hands. In *IEEE Symposium on 3D User Interfaces, 2006. 3DUI 2006* (pp. 59– 61).

Zhai, S., Buxton, W., & Milgram, P. (1994). The "Silk Cursor": investigating transparency for 3D target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence* (pp. 459–464). New York, NY, USA: ACM.

Zlatev, J., & Andrén, M. (2009). Stages and transitions in children's semiotic development. In *Studies in Language and Cognition* (pp. 380–401). Studies in Language and Cognition London: Cambridge Scholars.

APPENDIX A: STUDY MATERIALS

This appendix contains materials for the four studies in this dissertation.

Study 1:
- consent form
- photographs used in task 1

Study 2:
- consent form
- demographic survey
- post-experiment questionnaire

Study 3:
- consent form
- demographic survey
- NASA-TLX Form
- post-experiment questionnaire

Study 4:
- consent form
- demographic survey
- post-experiment questionnaire

Study 1 (in Chapter 4): Consent Form

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF SASKATCHEWAN
INFORMED CONSENT FORM

Research Project: Pointing at Distant Referents in the Real World
Investigators: *Carl Gutwin*, Department of Computer Science (966-8646)
Nelson Wong, Department of Computer Science (966-2327)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

In this study you will be asked to indicate different distant targets to your partner and identify targets indicated by your partner. You will also be asked to communicate with your partner via different communication methods (e.g., gesture only, gesture and written notes, and gestures and speech). With your permission, we may also video record parts of the study. The recording will be used for analysis, and your name would not be associated with it.

The session will take up to 1 hour to complete. You will receive a $10.00 payment for your participation.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (they will be made available on the HCI web site, hci.usask.ca). This summary will outline the research and discuss our findings and recommendations.

All of the information we collect from you (data logged by the computer, observations made by the experimenters, and your questionnaire responses) will be stored so that your name is not associated with it (using an arbitrary participant number). Any write-ups of the data will not include any information that can be linked directly to you. The research materials will be stored with complete security throughout the entire investigation. Do you have any questions about this aspect of the study?

You are free to withdraw from the study at any time without penalty and without losing any advertised benefits. Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. In addition, you are free to not answer specific items or questions on questionnaires.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

- Dr. Carl Gutwin, Professor — Dept. Computer Science — (306) 966-8646 — gutwin@cs.usask.ca
- Office of Research Services — University of Saskatchewan — (306) 966-4053

Participant's signature: _____
Date: _____
Investigator's signature: _____
Date: _____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Office of Research Services at the University of Saskatchewan.

Study 1 (in Chapter 4): Photographs Used in Task 1

Study 2 (in Chapter 5): Consent Form

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF SASKATCHEWAN
INFORMED CONSENT FORM

Research Project: Pointing in Collaborative Virtual Environment and Real World Environment
Investigators: *Carl Gutwin*, Department of Computer Science (966-8646)
Nelson Wong, Department of Computer Science (966-2327)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

In this study you will be asked to conduct pointing tasks in both collaborative virtual environment and real world environment. In both environments, you will be asked to point at targets projected on a wall and observe pointing by an avatar and a real person. You will be asked to fill out a questionnaire at the end the study.

The session will take up to 1 hour to complete. You will receive a $10.00 payment for your participation.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (they will be made available on the HCI web site, hci.usask.ca). This summary will outline the research and discuss our findings and recommendations.

All of the information we collect from you (data logged by the computer, observations made by the experimenters, and your questionnaire responses) will be stored so that your name is not associated with it (using an arbitrary participant number). Any write-ups of the data will not include any information that can be linked directly to you. The research materials will be stored with complete security throughout the entire investigation. Do you have any questions about this aspect of the study?

You are free to withdraw from the study at any time without penalty and without losing any advertised benefits. Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. In addition, you are free to not answer specific items or questions on questionnaires.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

- Dr. Carl Gutwin, Professor — Dept. Computer Science — (306) 966-8646 — gutwin@cs.usask.ca
- Office of Research Services — University of Saskatchewan — (306) 966-4053

Participant's signature: _____
Date: _____
Investigator's signature: _____
Date: _____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Office of Research Services at the University of Saskatchewan.

Study 2 (in Chapter 5): Demographic Survey

Demographic survey

1. Age: _____

2. Sex: **M** **F** (circle one)

3. University major and year: _____

4. Handedness: □ Right □ Left □ Ambidextrous

5. How many hours a week, on average, do you spend working with computers?

   □ 0-4 □ 4-8 □ 8-16 □ 12-16 □ 16-20 □ 20+

6. How many hours a week, on average, do you spend playing video games (including PC, console, and arcade)?

   □ 0-4 □ 4-8 □ 8-16 □ 12-16 □ 16-20 □ 20+

7. If you play video games, what kind of video games do you usually play? E.g. first person shooter, role playing, etc…

_____

Study 2 (in Chapter 5): Post-Experiment Questionnaire

Post-Experiment Questionnaire

1) How *confident* were you in the following cases:
a. pointing at the targets accurately (*pointing* task) in the **virtual** environment?

not confident at all not confident neutral confident very confident

b. pointing at the targets accurately (*pointing* task) in the **real world** environment?

not confident at all not confident neutral confident very confident

c. knowing where the avatar was pointing (*watching* task) in the **virtual** environment?

not confident at all not confident neutral confident very confident

d. knowing where the experimenter was pointing (*watching* task) in the **real world** environment?

not confident at all not confident neutral confident very confident

2) How *difficult* did you feel in the following cases:
a. pointing (*pointing* task) in the **virtual** environment?

not difficult at all not difficult neutral difficult very difficult

b. pointing (*pointing* task) in the **real world** environment?

not difficult at all not difficult neutral difficult very difficult

c. looking at the avatar pointing (*watching* task) in the **virtual** environment?

not difficult at all not difficult neutral difficult very difficult

d. looking at other people pointing (*watching* task) in the **real world** environment?

not difficult at all not difficult neutral difficult very difficult

3) In the *pointing* task in the virtual world, you used two different field of views:
- small field of view (you see a relatively shorter arm, but bigger targets)
- large field of view (you see a relatively longer arm, but smaller targets)

Which field of view do you prefer? Why?

4) In the *watching* task you observed from two different locations:
- observed from the side
- observed from behind

a. Which location gave you more confidence in knowing where the target is?

b. In both locations, did you feel any differences between the virtual world and the real world? What are the differences?

5) In the whole study (the virtual world and the real world), you were at two different distances from the wall:
- far
- near

Any comment on the two settings?

6) Overall which environment, *virtual* or *real*, do you have more confidence in doing the tasks? Why?

Study 3 (in Chapter 6): Consent Form

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF SASKATCHEWAN
INFORMED CONSENT FORM

Research Project:	Controlling Avatar Pointing in a Collaborative Virtual Environment
Investigators:	Nelson Wong, Department of Computer Science (966-2327)
	Carl Gutwin, Department of Computer Science (966-8646)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

In this study you will be asked to conduct pointing tasks in a collaborative virtual environment. You will be using different input devices to control an avatar. You will be asked to fill out a questionnaire at the end the study.

The session will take 1 hour to 1 hour and 30 minutes to complete. You will receive a $15.00 payment for your participation.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (they will be made available on the HCI web site, hci.usask.ca). This summary will outline the research and discuss our findings and recommendations.

All of the information we collect from you (data logged by the computer, observations made by the experimenters, and your questionnaire responses) will be stored so that your name is not associated with it (using an arbitrary participant number). Any write-ups of the data will not include any information that can be linked directly to you. The research materials will be stored with complete security throughout the entire investigation. Do you have any questions about this aspect of the study?

You are free to withdraw from the study at any time without penalty and without losing any advertised benefits. Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. In addition, you are free to not answer specific items or questions on questionnaires.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

• Nelson Wong	Dept. Computer Science	(306) 966-2327	nelson.wong@usask.ca

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

• Dr. Carl Gutwin, Professor	Dept. Computer Science	(306) 966-8646	gutwin@cs.usask.ca
• Office of Research Services	University of Saskatchewan	(306) 966-4053	

Participant's signature: _____
Date: _____
Investigator's signature: _____
Date: _____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Office of Research Services at the University of Saskatchewan.

Study 3 (in Chapter 6): Demographic Survey

Demographic survey

1. Personal Information:
   Gender □ Male □ Female
   Age □ _____

   University major or occupation_____

2. How many hours a week, on average, do you spend working with computers?

_____

3. How many hours a week, on average, do you spend playing video games (including PC, console, and arcade)?

_____

4. What kind of video games do you usually play? E.g. first person shooter, role playing, etc…

_____

5. Please rate your gaming experience:

1	2	3	4	5	6	7
Low						High

Study 3 (in Chapter 6): NASA-TLX Form

Workload Assessment

Mental Demand: How mentally demanding was the task?

Physical Demand: How physically demanding was the task?

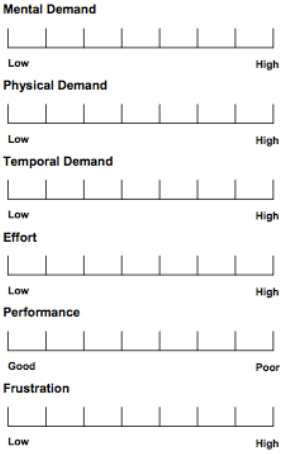
Temporal Demand: How hurried or rushed was the pace of the task?

Effort: How hard did you have to work to accomplish your level of performance?

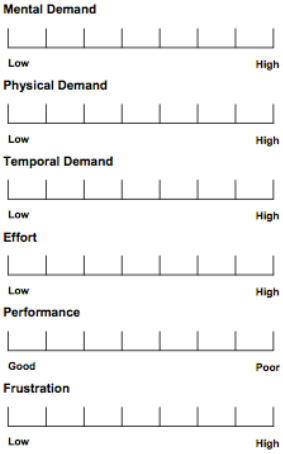
Performance: How successful were you in accomplishing what you were asked to do?

Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

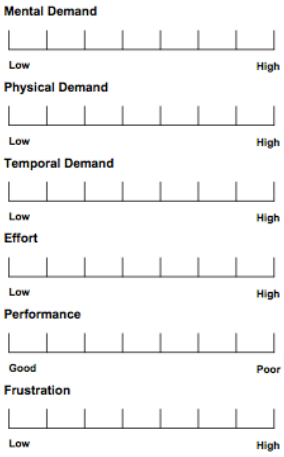
Trackball



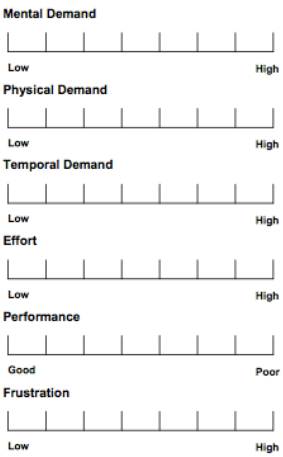
Joystick



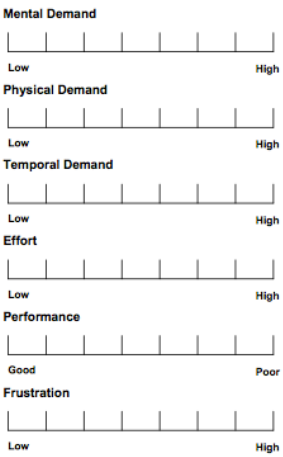
Mouse



Gamepad



Wii



Study 3 (in Chapter 6): Post-Experiment Questionnaire

Post-Experiment Questionnaire

1) Please rate your *overall* preference (each device must have a different rating)

Mouse (mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Trackball (trackball + mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Gamepad

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Joystick

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Wii Controls

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

---

2) Please rate your preference for the *moving + pointing* task (each device must have a different rating)

Mouse (mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Trackball (trackball + mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Gamepad (right thumbstick + d-pad)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Joystick (main stick + buttons)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Wii Controls (wiimote + balance board)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

---

3) Please rate your preference for the *turning + pointing* task (each device must have a different rating)

Mouse (mouse only)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Trackball (trackball + mouse)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Gamepad (right thumbstick + left thumbstick)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Joystick (main stick + hat)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Wii Controls (wiimote + numchuk)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

---

4) Please rate your preference for the *moving target* task (each device must have a different rating)

Mouse (mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Trackball (trackball + mouse + keyboard)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Gamepad (right thumbstick + right thumbstick + d-pad)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Joystick (main stick + hat + buttons)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

Wii Controls (wiimote + numchuk + balance board)

1	2	3	4	5	6	7	8	9	10
least preferred									most preferred

***** Comments**:*

Study 4 (in Chapter 7): Consent Form

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF SASKATCHEWAN
INFORMED CONSENT FORM

Research Project:	Pointing in a Collaborative Virtual Environment
Investigators:	Nelson Wong, Department of Computer Science (966-2327)
	Carl Gutwin, Department of Computer Science (966-8646)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.

In this study you will be asked to collaborate with someone to work on various tasks in a collaborative virtual environment (CVE). These tasks will involve gestural communication via avatars in the CVE. With your permission, we may also video record parts of the study. The recording will be used for analysis, and your name would not be associated with it. You will be asked to fill out a questionnaire at the beginning and the end of the study.

The session will take 1 hour to 1 hour and 30 minutes to complete.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

The data collected from this study will be used in articles for publication in journals and conference proceedings.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (they will be made available on the HCI web site, hci.usask.ca). This summary will outline the research and discuss our findings and recommendations.

All of the information we collect from you (data logged by the computer, observations made by the experimenters, and your questionnaire responses) will be stored so that your name is not associated with it (using an arbitrary participant number). Any write-ups of the data will not include any information that can be linked directly to you. The research materials will be stored with complete security throughout the entire investigation. Do you have any questions about this aspect of the study?

You are free to withdraw from the study at any time without penalty and without losing any advertised benefits. Withdrawal from the study will not affect your academic status or your access to services at the university. If you withdraw, your data will be deleted from the study and destroyed. In addition, you are free to not answer specific items or questions on questionnaires.

Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact:

• Nelson Wong	Dept. Computer Science	(306) 966-2327	nelson.wong@usask.ca

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. If you have further questions about this study or your rights as a participant, please contact:

• Dr. Carl Gutwin, Professor	Dept. Computer Science	(306) 966-8646	gutwin@cs.usask.ca
• Office of Research Services	University of Saskatchewan	(306) 966-4053	

Participant's signature: _____
Date: _____
Investigator's signature: _____
Date: _____

A copy of this consent form has been given to you to keep for your records and reference. This research has the ethical approval of the Office of Research Services at the University of Saskatchewan.

Study 4 (in Chapter 7): Demographic Survey

Balcony Study Demographics

* Required

Participant ID *

Group ID *

Gender *

- Male
- Female

Age *

University major or occupation *

How many hours a week, on average, do you spend working with computers? *

How many hours a week, on average, do you spend playing video games (including PC, console, and arcade)? *

What kind of video games do you usually play? E.g. first person shooter, role playing, etc...

What gaming system do you usually use? E.g. Wii, Xbox, PlayStation, PC, arcade system, etc...

Please rate your gaming experience. *

	1	2	3	4	5	6	7	
Low	○	○	○	○	○	○	○	High

Submit

Study 4 (in Chapter 7): Post-Experiment Questionnaire

Balcony Study Post-test Questionnaire

* Required

Participant ID *

Group ID *

Pointing Method Preferences

Please uniquely rank the pointing methods in each situation based on your preferences.
1 = most preferred
5 = least preferred

Indicating a Large Object *
Please give each pointing method with a different rating. ** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the above preferences? (for indicating a Large Object)

Indicating a Small Object *

Please give each pointing method with a different rating. ** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the above preferences? (for indicating a Small Object)

Indicating a Path *

Please give each pointing method with a different rating. ** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the above preferences? (for indicating a Path)

Indicating an Area *

Please give each pointing method with a different rating. ** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the above preferences? (for indicating an Area)

Indicating a Direction *

Please give each pointing method with a different rating. ** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the above preferences? (for indicating a Direction)

Please rank the pointing methods based on your OVERALL preferences. *

** 1 = most preferred; 5 = least preferred **

	1	2	3	4	5
no assistant	○	○	○	○	○
long arm	○	○	○	○	○
laser	○	○	○	○	○
spotlight	○	○	○	○	○
highlighting	○	○	○	○	○

Why do you have the overall preferences?

Avatar Preferences

Please uniquely rank the avatars based on your preferences. *

** 1 = most preferred; 3 = least preferred **

	1	2	3
normal avatar	○	○	○
avatar with fixed arms	○	○	○
invisible avatar	○	○	○

Why do you have the avatar preferences?

Did you pay attention to the arm of your partner's avatar? In what situation? Why? *

Did you pay attention to your partner's avatar? In what situation? Why? *

View Preferences

Overall which view do you prefer more? *

- ○ 1st person view
- ○ 3rd person view

Please comment on your view preferences.
Why do you prefer one over the other in general? Are there any situations you prefer the other view? Why?

Monitor Preferences

Overall which monitor setup do you prefer more? *

- ○ one monitor
- ○ three monitors

Why do you prefer one setup over the other?

Overall Comments

Do you have any other comments?
Please leave your comments about the study. They can be anything.

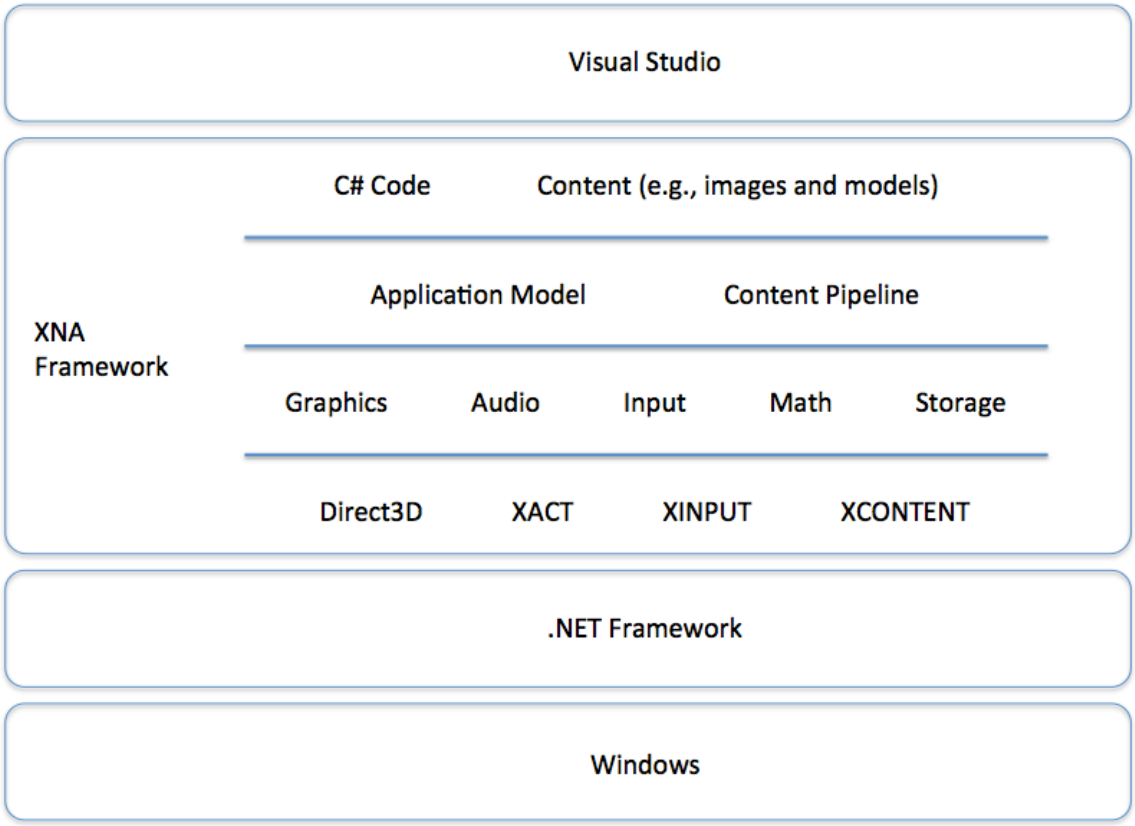
Submit

APPENDIX B: CVE

This appendix contains information of the CVE used in studies 2, 3, and 4 in this dissertation.

XNA Platform

The CVE was built using Visual Studio 2008 with C# that uses features in the XNA Framework. The XNA Framework is based on the .NET Framework that has important libraries for running applications on Windows. The diagram below shows the architecture of the XNA platform.



Models

The avatars were built using SketchUp. The buildings and the balcony used in study 4 were downloaded from Trimble 3D Warehouse (http://www.sketchup.com/product/3dwh.html).

Input Mappings

Different input devices were used in the studies: a keyboard, a mouse, an Apple Mighty Mouse (with a trackball in the normal mousewheel location), a gamepad, a joystick, a Wiimote, a Nunchuk, and a Wii Balance Board. All these inputs were mapped to control an avatar via transformation functions. For example,

$$x' = vx$$

where

x' is the degree of the avatar's arm rotates,
v is a variable controlling how much the arm rotates, and
x is the distance of the mouse cursor travels.

The larger the value of v is, the more the arm rotates with the same mouse movement, and vice versa.

External Libraries

- WiimoteLib v1.7 (http://wiimotelib.codeplex.com) was used to control Wii devices.
- GT#—Groupware Toolkit for C# (http://hci.usask.ca/research/view.php?id=34) was used to handle networking.