8-8-2007

On Inter-referential Awareness in Collaborative Augmented Reality

Jeffrey William Chastine

ON INTER-REFERENTIAL AWARENESS IN COLLABORATIVE AUGMENTED REALITY

by

JEFFREY W CHASTINE

Under the Direction of Ying Zhu

ABSTRACT

For successful collaboration to occur, a workspace must support *inter-referential awareness* – or the ability for one participant to refer to a set of artifacts in the environment, and for that reference to be correctly interpreted by others. While referring to objects in our everyday environment is a straight-forward task, the non-tangible nature of digital artifacts presents us with new interaction challenges. Augmented reality (AR) is inextricably linked to the physical world, and it is natural to believe that the re-integration of physical artifacts into the workspace makes referencing tasks easier; however, we find that these environments combine the referencing challenges from several computing disciplines, which compound across scenarios.

This dissertation presents our studies of this form of awareness in collaborative AR environments. It stems from our research in developing mixed reality environments for molecular modeling, where we explored spatial and multi-modal referencing techniques. To encapsulate the myriad of factors found in collaborative AR, we present a generic, theoretical framework and apply it to analyze this domain. Because referencing is a very human-centric activity, we present the results of an exploratory study which

examines the behaviors of participants and how they generate references to physical and virtual content in co-located and remote scenarios; we found that participants refer to content using physical and virtual techniques, and that shared video is highly effective in disambiguating references in remote environments. By implementing user feedback from this study, a follow-up study explores how the environment can passively support referencing, where we discovered the role that virtual referencing plays during collaboration. A third study was conducted in order to better understand the effectiveness of giving and interpreting references using a virtual pointer; the results suggest the need for participants to be parallel with the arrow vector (strengthening the argument for shared viewpoints), as well as the importance of shadows in non-stereoscopic environments. Our contributions include a framework for analyzing the domain of inter-referential awareness, the development of novel referencing techniques, the presentation and analysis of our findings from multiple user studies, and a set of guidelines to help designers support this form of awareness.

INDEX WORDS: Referencing, Collaboration, Augmented Reality

ON INTER-REFERENTIAL AWARENESS IN COLLABORATIVE AUGMENTED REALITY

by

JEFFREY W CHASTINE

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2007

ON INTER-REFERENTIAL AWARENESS IN COLLABORATIVE AUGMENTED REALITY

by

JEFFREY W CHASTINE

Major Professor:	Ying Zhu
Committee:	G. Scott Owen
	Sushil K Prasad
	Michael Weeks
	Blair MacIntyre

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2007

"*Daddy? Are you done working? Can you come play now?*"

To Laurie, Jack and Pierce, Mom and Dad – thank you for the sacrifices you've made.

Acknowledgements

I realize that all the things we accomplish in life cannot be viewed as wholly our own; instead, it is through the support and interaction with others that good things happen. First, I'd like to thank those who have made an academic impact in my life, including previous professors, my committee and my advisor. Ying, thank you for allowing me to be your student; I greatly appreciate your insight, support and patience. Blair, thank you for your expertise. I'd also like to thank those at Georgia Tech who persuaded me to pursue graduate studies, including Russ Shackleford, Mike McCracken, and Kurt Eiselt. There were numerous friends who helped me find my way through my undergraduate degree, including The Tomlinsons and The Holtons. I've been fortunate enough to be surrounded helpful colleagues, including Kristine Nagel, Mary Hudachek-Buswell, Byron Jeff, and William Hong, as well as my friends, Jeff Wilson, Jeremy Brooks, and Maribeth Gandy; Maribeth, without you I would probably still be coding. Of all my colleagues, however, my sincere thanks go to Jon Preston, who served as a study partner, moral compass and friend; Jon, you made this experience more enjoyable, and have influenced me both spiritually and academically.

My most profound thanks are reserved for my family, whom I thank God for on a daily basis. Mom and Dad, you have always been supportive of my endeavors, and words cannot express how important that has been to me. To my sons, thank you for your sacrifice by in allowing me time away from you to pursue this degree. Most importantly, I must thank my wife; Laurie, you are the centerpiece of my life, without whom none of this would have come to fruition.

Table of Contents

List of Figures

1.1. List of Tables

List of Abbreviations

AR	Augmented Reality
VR	Virtual Reality
CVE	Collaborative Virtual Environment
HMD	Head-Mounted Display
CSCW	Computer Supported Cooperative Work
VRPN	Virtual Reality Peripheral Network
DART	Designer's Augmented Reality Toolkit
TAR	Tangible Augmented Reality
FoV	Field of View
DoF	Degrees of Freedom
FPS	Frames Per Second
WIM	World In Miniature
WYSIWIS	What You See Is What I See
AMMP	"Another Molecular Modeling Program"
VoIP	Voice over IP
MVC	Model-View Controller
HCI	Human-Computer Interaction
HMPD	Head-Mounted Projective Display

CHAPTER 1

INTRODUCTION

The nascent field of augmented reality (AR) is highly multi-disciplinary, integrating knowledge from computer graphics, computer vision, system development, and Human-Computer Interaction (HCI). Whereas virtual reality (VR) attempts to create a completely immersive environment, synthesizing most, if not all aspects of a user's experience, augmented reality integrates virtual objects seamlessly into the physical world in real time[1]. Unlike VR, AR participants are able to interact with objects that exist in their everyday environment; there is no need to render existing physical objects, virtual likenesses for each participant, or the complexity of the environment that surrounds them. This fact is especially beneficial in the domain of *co-located collaborative augmented reality*, in which multiple participants occupy the same physical space; by preserving many of the important non-verbal communicative cues - such as gesturing, facial expression, hand, lip and eye movements – users can maintain work context and interaction can occur in a natural manner.

As application domains emerge, they bring with them a variety of new interaction techniques; the emergence of 3D applications brought with it spatial techniques to manipulate artifacts within the environment. Similarly, as applications become more collaborative, interface designers must find ways to gracefully support awareness between participants that are appropriate to the domain. Unlike other forms of computer-supported collaborative work (CSCW), augmented reality is inextricably linked with the

[1] The term *real time* here is loosely defined as "interactive rates", and does not require the system to meet hard deadlines.

physical world, and challenges many of the mental models we have developed for interacting within the environment. The non-tangible nature of virtual objects can be unnatural, as ideally these objects would give tactile feedback as well as provide proper occlusion and depth cues. These problems are exacerbated when participants are geographically separated.

A critical component of successful collaboration is the ability for participants to generate and interpret effective reference cues; more specifically, the environment must support *inter-referential awareness* - or the ability for one participant to refer to a set of objects and for that reference to be understood. It is challenging to support interactions where virtual content is present, but even more so in a collaborative setting. Often these collaborations occur across distance, relying on computer-mediated communication and interactions. Thus, in addition to providing a set of techniques that are flexible enough to work with multi-modal content, designers of collaborative augmented reality systems must support communication and awareness while maintaining the contextual properties of the environment.

Though a significant amount of research has been performed in collaborative AR, very little has addressed the fundamental task of ensuring that collaborators share a mutual understanding of an object of reference; while techniques from purely physical and purely virtual environments appear at first to be applicable, we demonstrate that there are unique and significant challenges to achieving inter-referential awareness in collaborative AR. This research intends to address these issues by 1) providing a solid theoretical background for inter-referential awareness, 2) acquiring an understanding of how users generate references and the kinds of support they desire, 3) evaluating a

subset of current referencing techniques, 4) developing new, multi-modal referencing techniques, 5) architecting, implementing and evaluating methods of environmental support for inter-referential awareness, and 6) proffering a set of guidelines for designers of collaborative AR systems.

1.2. Motivation

Augmented reality generates a unique set of referencing scenarios not possible in purely physical or virtual realities. Users can refer to objects of differing modalities which exist in either local or remote workspaces. A simple cross between these two dimensions forces one to consider appropriate techniques for generating references to remote, physical objects. While virtual artifacts normally augment a physical one (e.g. providing meta-data about the object), the literature suggests scenarios where virtual objects are embedded *within* physical ones; thus the proximity from which a physical reference is made is restricted and consequently has a higher probability of becoming ambiguous (if the referencing technique is susceptible to distance). Objects maintain many of their spatial properties that influence referencing, including distance, scale, proximity to other objects – which can potentially occlude the views of one or more participants. In AR however, physical objects do not naturally occlude virtual ones, so additional steps must be taken to ensure that physical reference techniques such as pointing do not become ambiguous. Further, collaborative AR systems are often comprised of heterogeneous hardware configurations, including Head-Mounted Displays (HMDs) with varying display capabilities, cameras with various Fields of View (FoV),

and a mix of tracking technologies. Many of these HMDs provide a bioscopic[2] view of the world, limiting the depth cues that users receive from the environment. While this list is not exhaustive, it demonstrates the complexity of an ostensibly straight-forward task. We continue by examining a few of the common scenarios found in collaborative AR literature.

Local Expert/Remote Technician: Technicians frequently travel to a remote jobsite to maintain or repair products. When technicians lack the expertise necessary to complete their task, they may contact a remote expert for assistance, but must often rely strictly on verbal communication. One method of increasing workspace awareness requires the technician to be equipped with a head-mounted camera - providing the expert with a view of the remote environment [1]. During this collaboration, if confusion occurs, it is necessary for the expert to clarify instructions by establishing a common point of reference in the remote environment. However, deictic speech (e.g. "*this*", "*that*", "*those*") alone may be insufficient in environments where there are few discernable features available, such as when re-wiring a network panel or examining a series of pipes. Using AR, references can be *spatially* registered in the remote world – augmenting the environment for the technician. Additionally, it should be possible for either party to clearly refer to physical objects regardless of distance; for example, a technician may look up to examine a series of pipes located tens of feet above.

Medical Scenario: A medical staff gains important insight into the status of a patient by acquiring 2D and 3D images, such as X-rays, CAT scans, or ultrasounds. Surgeons often view this data on a lightboard or computer monitor before and *during* a

[2] A single camera feed which is replicated for both eyes.

medical procedure, requiring them to physically turn their head (between the patient and lightboard) and to mentally map information between two disparate coordinate systems. In an ideal AR medical scenario, this data can be projected *into* the patient, allowing the staff to see, for example, a fetus within a mother's womb or a tumor embedded deep within the brain [2-4]. In describing such a system, Johnson et al. write "*The challenge is to make it available for routine use by surgeons at all hospitals in the country. This will require significant further development of visualization techniques, better augmented reality systems, and the successful integration of these systems in the operating room.*" [5]. During the course of surgery, the medical team needs the ability to refer to non-contiguous regions of tissue, such as potentially cancerous areas. These regions can lie on or below the surface of the skin, requiring sub-dermal referencing. A misunderstanding of the reference can result in undesirable, if not fatal, consequences. In such scenarios, it is important that that doctors and nurses be able to convey their ideas effectively through referencing virtual and physical artifacts at arbitrary depths; further, it may be necessary to do this in a hands-free manner, as the surgeons are most likely occupied with the surgical task.

Scientific Visualization: Computers have long been used in the domains of visualization, molecular structure modification, crystallography and molecular docking [6-12]. Similarly, AR has been used to collaboratively visualize molecules as well as internal vectors within fluids. Co-located environments allow participants to spatially visualize 3D information and interact within the environment in natural ways. However, when discussing attributes of a series of data, scientists must have flexibility in the granularity with which they generate references. For example, scientists may refer to a

molecule in its entirety, to large sub-structures, to small clusters, or to individual atoms. Further, referencing small portions of a larger data set can be quite difficult given that there are often few discernable features within the information. For example, organic molecules often contain a large number of atoms from a limited set of elements. Here, a significant portion of each molecule will be comprised of hydrocarbons, such as methane (CH_4), ethane (C_2H_6), propane (C_3H_8), butane (C_4H_{10}) or even decane ($C_{10}H_{22}$). Further, referencing can be challenging given that 1) parts of the visualization (such as atoms) can occlude the views of one or more participants, 2) data may be tightly clustered, and 3) data with similar traits are hard to distinguish through speech alone. How to interact with clustered sets of objects is still an active area of research [13].

1.3. Current State of the Art and Limitations

Referencing techniques found in AR environments are ad-hoc or are derived from VR selection techniques; alternatively, developers may assume that natural gesturing and deictic speech are sufficient. One naïve approach is to implement selection techniques from the domain of VR; however, many of these rely on mathematical intersection, requiring geometric knowledge of physical objects for them to function correctly. Consequently, a majority of these techniques work exclusively with virtual artifacts. Another option is to "ignore" support for referencing, and rely solely on the natural referencing capabilities of the users. However, this approach fails for remote scenarios - where it has been shown that when proper referencing techniques are not supported (e.g. audio only conditions), "*lengthy descriptive sequences were typically required*" [14]; thus, reducing support comes at the cost of efficiency. We describe the state-of-the-art in the context of the scenarios presented in section 1.1.

Given the "granularity of referencing" in scientific visualization, techniques such as ray-casting[3] may be inappropriate – as they produce a single point of reference; thus, when referring to multiple objects, such as regions of a molecule, this technique is not scalable. Physically pointing behaves much like a virtual arrow - inferring a broad set of objects; this is especially pronounced if the distance exceeds a threshold or the objects are clustered. Thus, pointing to an atom in a molecule is analogous to pointing to a tree in a forest: very little information is available to verbally distinguish between them. Physically pointing can also be disconcerting to users unless the hand is tracked to produce occlusion cues.

In the expert/technician scenario, we cannot rely on traditional tracking methods, as realistically, the technician will *not* be located in an environment conducive to tracking. Further, little to no geometric data is typically available about the objects in the remote environment. Until this data is made accessible, it is doubtful that traditional virtual selection techniques can be applied to real-world objects. Recent research has investigated annotating the technician's view two-dimensionally to provide a time-sensitive reference, but is obviously no longer relevant - even inferring incorrect objects - if the user changes their viewpoint [15].

In the medical scenario described above, the areas of reference may be located on the surface of the skin, be sub-dermal, on the surface of the virtual object, or may be embedded within the virtual object itself; thus, it is necessary to refer to both physical and virtual objects at arbitrary depths. An occluding physical barrier (e.g. the skull or chest cavity) prevents the embedded virtual objects from being physically "touched" –

[3] An intersection-based selection technique found in VR environments

forcing references to occur from a distance. Objects may be visible only from a particular viewpoint (for example, when the object is within a cavity), and while referenced correctly, may not be visible for others. Image plane techniques (see Chapter 2) may infer a cross-section of the data. Another approach is to use a variable-length virtual ray, which requires specialized hardware. Further, using an external device to make references, such as a wand, requires the *use of their hands* – which are occupied in the process of surgery. The medical scenario represents a tight coupling between the virtual and physical, emphasizes the importance of context, and highlights ARs unique ability to embed virtual objects within physical ones.

1.4. Problem Statement and Research Goals

As suggested by the literature, it is crucial that referencing techniques be present in collaborative scenarios [16]. However, no comprehensive research has been performed that analyses how referencing occurs across various scenarios or methods that can support it. As we transition into future computing domains, there is a need for a generic framework that describes referencing and encapsulates discipline-specific factors. Beyond this, we need to understand the limitations of current referencing techniques, how participants use them, and in what context their inclusion is appropriate.

Though some research has investigated the communication behaviors of co-located pairs, we need more insight into how groups make references across a variety of scenarios, as well as the technologies that they prefer. Further, referencing can be negatively influenced by environmental factors. Thus, there is a need to identify these factors as well as explore new techniques that might help alleviate them. Conversely, we

need to explore ways in which the environment can passively *support* referencing, which can then be incorporated into future systems.

Finally, we must consider new techniques for referencing multi-modal content. The medical scenario exposes the need for referring to content at arbitrary depths, as well as in a hands-free manner. The molecular modeling scenario demonstrates the need for generating references at varying granularity and overcoming occluded views; thus, spatial referencing techniques must be investigated as well as techniques that may alleviate occlusion. Further, while distributed AR systems focus on sharing virtual artifacts, the expert/technician scenario indicates a strong need for referring to physical content in remote environments.

1.5. Contributions and Significance

Inter-referential awareness is viewed as a fundamental, yet critical, component in collaborative augmented reality [16]. Billinghurst et al. describe the importance of (general) gesturing to facilitate communication [17]. Agrawala describes how gestures are used to refer to objects, and how they are important in "*establishing a shared context for the group… When someone refers to an object by pointing to it, the object becomes the focus of the group*" [18]. It is apparent that for *any* collaboration to occur, an intrinsic requirement is the ability for participants to refer to the physical and virtual objects that surround them in a consistent manner, and be confident that these references are understood. Our work intends to examine this form of awareness holistically while addressing contextual problems in depth. Specifically, the contributions of this dissertation include:

1) providing a flexible framework for conceptualizing the complexity of inter-referential awareness across a variety of computing domains,
2) analyzing the design space of referencing by applying the framework to collaborative AR; this includes observation of user behaviors, the presentation of user feedback and the examination of techniques that support inter-referential awareness,
3) developing a novel, hands-free, multi-modal technique that is flexible enough to be used in a variety of scenarios,
4) analyzing the properties of a common referencing technique (as it applies to the selection and representation phases of the framework) through independent studies of giving and interpreting references,
5) proffering a set of guidelines to help system designers support inter-referential awareness, and
6) discussing the underlying architectural issues that support this form of awareness.

Throughout this dissertation, an overarching theme is that there is a cost that must be incurred when supporting this form of awareness. This cost is attributed to several inter-related factors, many of which are identified in the framework and taxonomy of Chapter 4. To better understand how they are related, we refine our understanding of it through user studies, where we find that these costs may be relaxed under certain contexts, or may be reduced through alternative support. We view this discussion as a significant contribution; it is subsequently re-examined in Chapter 8, where it is presented in the context of our findings.

1.6. Organization of the Dissertation

This dissertation is presented chronologically with the exception of Chapter 2 (background and related work). Chapter 3 describes our initial work in virtual environments for molecular modeling, which exposed the basic problem of inter-referential awareness between collaborators; there, we demonstrate our approach to spatial referencing as well as a new, multi-modal technique that can refer to local and remote content at arbitrary depths. Later, we transitioned the VR environment into AR with the intent of collaborating remotely, which exposed additional scenarios and complexity of referencing in AR. Thus, to enumerate and encapsulate referential concepts and factors, Chapter 4 proposes a generic CSCW framework and applies it to the field of AR. This framework informed the design of an exploratory study, which examined user behavior across the most common scenarios found in the literature; the intent behind this study was to cull the design space of referencing, the results of which can be found in Chapter 5. A common, multi-modal referencing technique that works across a wide variety of scenarios is the virtual pointer; Chapter 6 presents the properties of this technique, including the effectiveness of how users give and interpret references using it and environmental factors that may influence its accuracy. Based on the feedback from the exploratory study, Chapter 7 presents a follow-up study which examines how the environment can passively support referencing, and describes the underlying architecture that supports it. Chapter 8 summarizes our work by presenting the overall knowledge gained, including design guidelines, a discussion on the cost of referencing, as well as future work.

CHAPTER 2

BACKGROUND

Collaborative AR is highly multi-disciplinary, drawing knowledge from a diverse set of computing domains, including Human-Computer Interaction (HCI), computer graphics, computer vision, and CSCW. Section 2.1 gives a brief overview of AR environments – summarizing general ideas, requirements, hardware and system architecture. For the efficient reader, the remaining sections of this chapter are summarized in section 2.5. Section 2.2 examines related work from the field of CSCW, stressing the field of awareness and effects of shared video. We next examine relevant work in VR, including selection techniques, pertinent studies on awareness and referencing techniques. Finally, we return to AR to examine prior work in collaborative systems and techniques that facilitate communication.

2.1. Overview of Augmented Reality Systems

Augmented reality environments super-impose virtual objects within the physical environment, allowing participants to perform tasks previously not possible. In defining a reality-virtuality continuum, Milgram describes AR as a subset of the broader domain of *mixed reality* (see Figure 1) [19]. Located at the extremes of this spectrum are real-world environments and purely virtual environments. Augmented reality lies between these two, but is logically offset closer to the real environment given the amount of information synthesized for the user. The continuum also includes the lesser-known field of *augmented virtuality*, which includes techniques for compositing live video of a user's face to augment a virtual avatar.

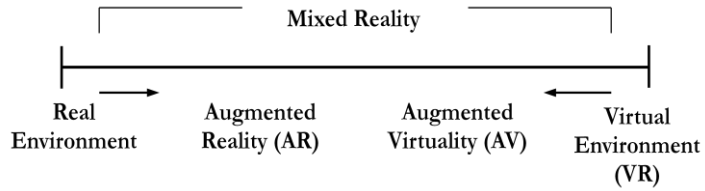

Figure 1 - Milgram's Reality-Virtuality Continuum [19]

In 1965, Ivan Sutherland suggested the notion of mixed reality by describing a tethered *head-mounted display* using half-silvered mirrors – allowing the user to view real and virtual objects simultaneously [20]. Beyond being a visual display, HMDs must also inform the system of the user's viewpoint - tracking with three or six degrees of freedom (DoF)[4]. As with most technology, HMDs have evolved, becoming lighter (with some weighing less than 8 ounces), containing much higher resolutions, providing more accurate tracking, and costing significantly less than previous generations. Similar to Sutherland's display, modern HMDs can be see-through, in which a semi-transparent display (or an *optical combiner*) lies between the eyes and the real world, displaying only the augmentation[5]; this technique affords a large field of view of the physical world, but can suffer from latency issues (i.e. *dynamic errors*) when registering virtual objects into physical space – causing virtual objects to lag behind physical ones. Though predictive algorithms can minimize this effect, it can become especially pronounced with rapid changes in the user's viewpoint or when rendering geometrically complex virtual objects. To overcome this error, an alternative approach is to provide a mediated view of the world through a *video-based* HMD - in which a camera is mounted on an

[4] 3 DoF tracks orientation while 6 DoF tracks location as well. AR requires 6 DoF.
[5] The user is provided with an unmediated view the physical world even when no power is applied.

opaque display as close to the eyes as possible; this provides the system a continuous view of the world from the user's perspective (see Figure 2 and Figure 3). The video stream is then sent to a computer, which is responsible for augmenting it with virtual objects and feeding the stream back to the display. This genre of HMDs is restricted to the parameters of the camera, including resolution, frame rate, color depth, focal length, field of view and distortion; while eliminating dynamic delays between virtual objects and the physical world, lag can still occur between head movement and what is displayed to the user.

Figure 2 - A Head-Mounted Display [21]

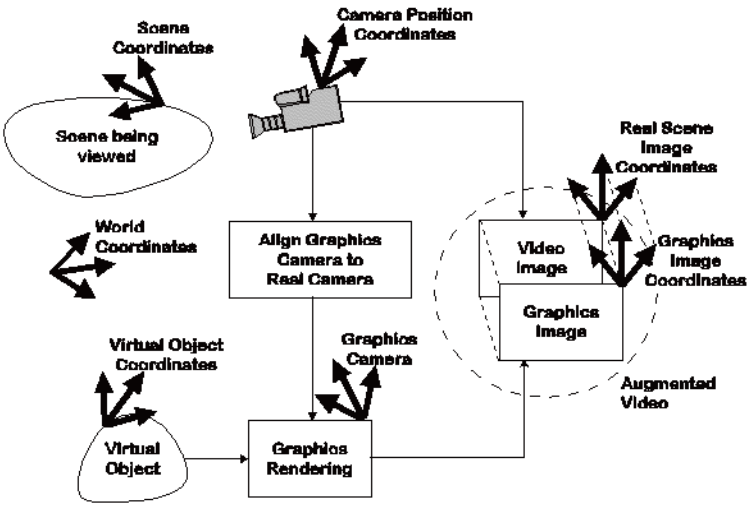

Figure 3 - AR system architecture [22]

For a system to qualify as an augmented reality system, it must 1) combine virtual information into the real world 2) be interactive and in real time and 3) register virtual objects in physical 3-D space [23]; for the system to be *useful*, it must allow participants to "*go beyond*" their normal experience, providing methods of interaction not possible in other environments [24]. In addition to tracking the user's viewpoint, the system is often required to be acutely aware of the pose of physical objects within the workspace. When geometric data of these objects is known a priori, tracking them allows for realistic interaction between real and virtual artifacts[6].

Tracking technologies are typically electromagnetic, acoustic (using triangulation and time-of-flight principles), gyroscopic, inertial, or vision-based - in which the system is trained to recognize *fiducials* (or *markers*); the ARToolkit is an example of a widely-adopted, marker-based tracker [25]. The vision system first identifies the corners of a series of black squares to determine the *pose* (i.e. orientation and position) of the fiducial. The system then recognizes the inner pattern to identify which fiducial is present. A vision-based approach requires that the environment be modified to accommodate the fiducials, cameras to be properly calibrated (to accommodate for lens distortions and focal points) and the environment to be adequately illuminated. Other vision techniques that are being researched do not require fiducials, but instead track features within the environment [26]. Unlike VR, augmented reality suffers from *static registration problems* – or the misalignment of virtual objects within the physical world caused by inaccurate tracking; the amount of registration error a user is willing to tolerate is still an open problem. In general, it can be said that there will always be a

[6] Such as providing occlusive cues.

need for a low-latency, low-cost, wide-area, accurate tracker; therefore, this field is still an active area of research.

Augmented reality has undergone strong growth in past few years, in part due to the recent emergence of consumer-priced hardware as well as a "curiosity shift" from the virtual environments community. AR systems are also rapidly moving into the wearable domain - creating the field of mobile AR [27]. While still an emerging technology, wearable computing is a next generation, portable, often less-powerful system similar in features to a standard laptop; given its portability, however, alternative input and output devices are required, such as chorded keyboards, audio interfaces and monocular displays [28]. These systems rely on wireless technology to keep the system aware, and experience difficulties when tracking outside. Tangible AR (TAR) has also been explored, where a physical object serves as a handle for the virtual objects [29, 30]. Weghorst et al. have explored augmenting physical, molecular objects with a virtual electrostatic field - providing users with a tangible interface [29]. However, the object must still be tracked, requiring a device to be mounted in or on it - hindering interaction.

When developing AR technology, HMDs are not the only choice of display. By mounting a camera onto the back of tablet-like PCs, users are provided a windowed view into the augmented world [31]. While this technology allows participants to see one another's eye gaze, they require two hands, leverage alternative methods of interaction with virtual objects and suffer from user fatigue. Tabletop devices are common as well, where a physical object (most often a table) is augmented through projection; examples can be seen in [32-34]. Another rather unusual alternative is using HMPD (Head-Mounted Projective Display) technology, in which lightweight projectors

are mounted on the user's head, projecting images *into* the environment [35]. Such a system is unlikely to be widely adopted, as a majority of the environment must be covered in retroreflective material. However, this system has the benefit that projections can occur on arbitrarily-shaped surfaces due to the properties of the material.

2.2. Related Work from CSCW

The importance of non-verbal communication is undeniable. Anthropologist Edward T. Hall claims that approximately 55 percent of our communication is non-verbal [36]. Albert Mehrabian identified that words account for only 7 percent of the overall message, with 38 percent vocal (such as tone, inflection, and other sounds) and 55 percent non-verbal (corroborating Hall's claim) [37]. These studies suggest the importance of supporting non-verbal communication during collaborative tasks; often, others are implicitly aware of our actions, or can be made so explicitly; in essence, we get this form of communication "for free". Moving into a digital medium, however, changes the nature of the communication and can introduce the asymmetries described in section 2.4 [1, 38]. Robinson researched the importance of embodied actions in collaborative workspaces, and discussed the relationships that exist between the physical environment and the participants, noting that "*Pointing is the classic example of an action used to maintain indexicality*" and that "*The interpretation of what is being pointed at is dependent not just on the act of pointing but on other people being able to perceive what is being pointed at"* [39].

Awareness is an important topic in CSCW, and is still an active area of research. Bafoutsou et al. conducted a review in the area of "*creating collaborative application taxonomies*" noting that many taxonomies begin by classifying environments by *when*

and *where* the collaboration occurs [40]. To better understand how to support referencing in collaborative applications, Hindmarsh studied users in a real world setting [41]. Gutwin et al. have researched *workspace awareness* – or an up-to-the-minute knowledge of the activities occurring in the workspace – and provided a framework for describing it [42]. Fundamental to communication is establishment of *common ground,* which is defined as "*mutual knowledge, mutual beliefs, and mutual assumptions*" (i.e. a shared set of information between participants) [43]. In early work, Dourish et al. argue that awareness of individual and group activities is critical to successful collaboration, and notes that information sharing and coordination are central to it [44]. To demonstrate the importance of referencing, users in this environment experienced difficulties when their activities could not be observed by others – as referencing techniques were not explicitly supported. To overcome this limitation, subjects asked where others were editing by verbally referencing a location, though one group created an ad hoc visual indexing and indentation scheme to give location references to others. Communication and mutual awareness are key factors in establishing and maintaining collaborative work [45, 46].

Based on informal discussion, groupware (i.e. software that allows users to collaborate) participants felt that distributed groupware makes discussion and *group focus* more difficult, and that it can confuse users unless verbal explanations accompany actions [47]. However, participants felt that distributed sessions provide better access to information and facilitates parallel work. WYSIWIS ("What You See Is What I See") interfaces allow users to share viewports and undoubtedly play an important role in inter-referential awareness [48].

Findings indicate that speech is the most important medium in teleconferencing [49][7]. With the availability of more network bandwidth, the research community shifted their investigation into the effects of remote video feeds on collaborative work. Fussell et al. demonstrated that shared video helps in establishing common ground [50]. However, mediated communication is not the same as unmediated (i.e. the fact that it has been transmitted by technology affects the communication); Sellen studied the effects of communication in a mediated environment, showing that sharing the same space is important for communication [38]. Similarly, Gaver notes that, depending on the media, the transmission of communication cues may not occur between the collaborators [51]. For example, the amount of communicative information that is transferred in video conferencing is dependent on the field of view and resolution of the camera; positioning the camera to close to the user creates a "floating torso" - failing to capture hand gestures - while a camera positioned further away is incapable of accurately capturing subtle gestures. In their analysis of camera configuration and placement, Ranjan et al. noted the head-mounted camera of the technician "*works well for establishing a joint focus of attention*" [52].

In order to better facilitate communication between a remote expert and local technician, researchers have studied uni-directional video. Kraut et al. suggest that providing a view of the technician's workspace increases task efficiency and that references form a critical part of collaboration [14]. Ou et al. extend this idea by describing a method which allows experts to refer to remote objects through annotation

[7] It comes as no surprise that speech was also shown to be the strongest indicator of presence in Collaborative Virtual Environments (CVEs).

of the technician's view [53]; however, these references quickly become stale if the technician changes his viewpoint. A further variation of this is to allow the expert to freeze the current video frame and then annotate over top of it, at the cost of not seeing a current view of the world [15]; this research further recognized the importance of speech when referencing an object, including absolute referencing, relative referencing and deictic referencing. Using a more complex approach, Cheng et al. propose stitching together a mosaic of viewpoints from the technician over time, allowing the expert to view the remote environment independent from the technician's viewpoint [54]; however, if the position of the technician changes, the mosaic becomes outdated. An alternative, more-physical approach was taken by Sakata et al., who suggest mounting an active camera - equipped with a laser pointer - on the shoulder of the technician; this technique allows the expert to view the remote space independently as well as generate point references within the environment [55].

2.3. Related Work from VR

The human being is well-equipped to interact with physical objects in the real world. However, interacting within virtual environments is considerably more challenging for us - as virtual objects do not provide tactile feedback[8]. Because AR derives several interaction patterns from VR, it is important to better understand their affordances and limitations. This section briefly addresses theoretical concepts and interaction techniques that relate to our work.

[8] Though they can in a limited capacity, using devices such as the PHANToM

Benford et al. introduced a general, theoretical framework for mutual awareness in collaborative virtual reality called the *spatial model of interaction* [56]. The key components to the spatial model are *space* and the *objects* that inhabit that space, *aura* – which is a sub-space that surrounds an object and acts as an enabler for interaction, *focus* – which is the direction of attention (increasing awareness of the object), and *nimbus* – which is the capability for objects to make themselves available to others (i.e. increasing the nimbus makes an object more noticeable). The work of Curry in tele-immersion extended Benford's model, identifying major types of awareness including presence, attention, environmental and action awareness [57]. To maintain gaze awareness, Hindmarsh recommends rendering the view frustum[9] of participants during referential tasks [58].

It is possible to divide selection techniques into the more primitive components of 1) an indication of the object of interest, 2) a confirmation the selection by the user, and 3) feedback that the selection was completed [59, 60]. Variables that affect selection performance include target distance, size, number, density of surrounding objects and target occlusion [61]. We can classify interaction techniques as either *isomorphic* - in which there is a direct mapping between the real and virtual hand position - or *non-isomorphic* - which use linear or exponential scaling to perform selections at a distance. Poupyrev further categorized interaction techniques as either ego-centric (i.e. from the perspective of the avatar – first person) or exo-centric (i.e. any other viewpoint) [62]. Exo-centric metaphors include World-In-Miniature (WIM – where the world is scaled to fit in the palm of your hand in order to overcome distance) or automatic scaling

[9] A truncated pyramid which defines which virtual objects are rendered; the viewing space.

techniques. Ego-centric metaphors are further decomposed into virtual hand techniques (such as the "classical hand", "Go-Go" [10] and "Indirect Go-Go") and virtual pointer metaphors (like ray-casting, aperture, flashlight and image plane techniques) [61, 63].

A primary benefit to using the pointing metaphor is that it naturally affords selection of objects at a distance, allowing the user to remain stationary. The most common method is ray-casting, in which an infinitely-long line emanates from the virtual hand to intersect with a virtual object. However, this technique suffers from the magnification of small errors (from tracking or user ability) over distance, its inability to reference general space, and the possibility for the ray to intersect with more than one object. Several theme and variations are derived from this, including two-handed pointing, where one hand controls the direction vector and the relative distance between the two hands controls its length [64]. Using two hands, another technique is the curved virtual pointer, where a twisting of the hands bends the pointer to select occluded objects [65]. Still another variation is the flashlight technique where a conical volume is projected from the hand [66]. This alleviates some of the distance-related inaccuracy problems of ray casting, but may be too course of a technique in clustered scenarios. To reduce this limitation, the aperture technique allows the user to control the radius of the conic volume [67]. The "fishing reel" method allows users to control the length of the virtual ray through the physical manipulation of a specialized device, such as a slider. Image plane techniques allow users to select objects by their projection onto a 2D virtual image plane - essentially framing the object. For example, the "head-crusher" technique

[10] Named after the children's TV show *Inspector Gadget*, in which the appendages of the main character could extend to arbitrary lengths.

allows a user to select an object by placing their hand between the eye and the object, and positioning their index finger above and their thumb below the object (allowing someone to virtually "crush the head" of another person) [68]. In general, Bowman et al. recommend pointing metaphors for selection in virtual environments [61].

The virtual hand metaphor includes the linear hand, the Go-Go technique, and the World-In-Miniature [69]. Simple hands are isomorphic – directly mapping to movement in the real world and serving as a 3D cursor in the virtual world. Its graphical representation can take the form of an actual hand, or can be a semi-transparent volume, such as the "Silk Cursor" - which uses a semi-transparent bounding box which allows the user to view occluded virtual objects and yield important depth cues [70]. The Go-Go technique is similar in behavior to the classic hand - where movements within a fixed distance from the user's origin are isomorphic; however, once the hand extends beyond a pre-determined distance, the hand becomes non-isomorphic, scaling exponentially. A final variation on the virtual hand is the world-in-miniature technique, in which the entire virtual world is scaled to fit within the user's virtual hand. Users can then select virtual objects in the shrunken world, and have those selections reflected in the fully-scaled environment. A drawback of this technique is its inability to work in crowded or very large environments, or when selecting small objects.

2.4. Collaborative AR

Many techniques for selecting virtual objects in AR are derived from those found in virtual environments and rely on similar hardware, including data gloves, haptic feedback equipment and pointing devices. However, AR poses unique interaction challenges – as participants do not work exclusively with virtual content. Because AR

includes both the physical world *and* the way in which we interact with it, environments must inherently support multi-modal interaction methods. Bowman et. al describe the general problem:

> "*… the user is required to use* different input modalities *for physical and virtual objects: real hands for physical objects and special-purpose input devices for virtual objects. This introduces an* interaction seam *into the natural workflow."* [61]

These challenges compound when extended to include multiple participants, in which support must be provided for real-world collaboration including gesturing and deictic references[11]. In this section, we summarize a representative cross-section from the literature, highlighting common scenarios, requirements and awareness methods that have transferred from collaborative virtual environments.

Collaborative AR

Collaborative AR allows for multiple participants to interact with a shared set of virtual and physical objects; these participants may be *co-located* (i.e. in the same physical space) or *remotely located* (i.e. geographically separated). When designing collaborative AR applications, Billinghurst warns that "*mechanisms which may be effective in face to face interactions may be awkward if they are replicated in an electronic medium, making users reluctant to use the new medium*" [1]. While collaborative AR can certainly allow users to "*go beyond*" the normal group experience, they pose unique problems, several of which are summarized in this section.

[11] In linguistics, these are context-dependent words for who or what is being referenced (i.e. "this", "that", "I", etc.)

Early research in collaborative AR includes Billinghurst et al.'s *Shared Space*, Rekimoto's *Transvision*, and Schmalsteig et al.'s *Studierstube* (German for "study room") [31, 71, 72]. These authors argue that, by preserving many non-verbal cues, such as gestures, eye gaze and lip movements, co-located environments naturally facilitate collaboration among the participants; because of this, interaction can rely heavily on social protocols, such as turn taking, and can preserve the context of their everyday surroundings [73]. They further note that because the coordinate system of each participant is the same, *deictic referencing* (verbally referring to something, usually in conjunction with gestures) is meaningful.

By extending the general AR requirements, the developers of *Studierstube* identified five key characteristics for collaborative AR:

- *Virtuality*: virtual objects can be viewed and examined.
- *Augmentation*: Real objects can be augmented by virtual annotations, and changes to the virtual model will be disseminated to all participants.
- *Cooperation*: Multiple users can see each other and cooperate in a natural way.
- *Independence*: Each user has an independent viewpoint.
- *Individuality*: The data displayed can be different for each viewer. This includes customized view as well as the viewing of private information.

Renevier and Nigay define a collaborative AR system as one "*in which augmentation of the real environment of one user occurs through the actions of other users and no longer relies on information pre-stored by the computer*" [74]. They classify collaborative AR into three categories:

1. *Remote collaboration in one augmented reality*: in these systems, at least one user is physically next to the object of the task, while some users are distant, similar to the remote/expert scenario in Chapter 1. Other work in this scenario examined communication patterns and user interaction in spatial workspaces [75].

2. *Remote collaboration in augmented realities*: there are several objects involved in the task, say, one for each user, which are remotely linked yet *physically* present at different places. An example is a shared whiteboard.

3. *Local collaboration in one augmented reality*: all users are located in a physically shared environment next to the object of the task; this is the most common scenario, and includes the medical scenario described in Chapter 1 (for another example, see Figure 4).

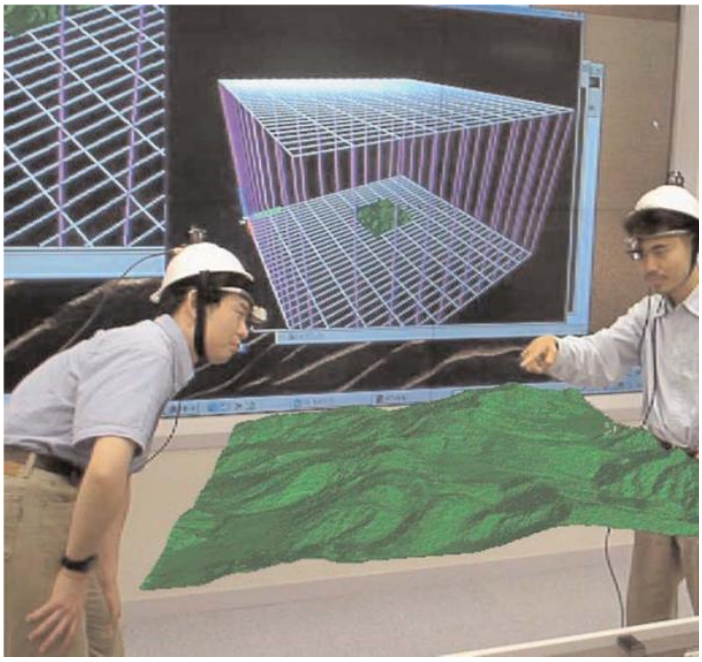

Figure 4 – Co-located users examining virtual terrain [73]

Symmetry and Seams

The concept of *symmetry* permeates much of the research in collaborative AR. Billinghurst et al. explored asymmetries that occur between participants using wearable computing and those using a desktop, creating disparities in communication [17]. Asymmetries exist when the technological capabilities differ for users on both ends. For example, asymmetries are present when a remote expert can view only the task space of a technician, and the technician can view only the face of the remote expert. The authors define communication asymmetries as "*an imbalance in communication introduced by the interface user from communication, the expertise or roles of the people communicating, or the task undertaken*". When two systems present the same abilities to users, they are *functionally symmetric*. For example, two identical systems may have the ability to share audio, video and documents. If the quality of the representation differs (such as the resolution of the monitor), the overall system has *implementation asymmetries*. Even though two users may share video, if one displays his face while the other displays his viewpoint, the system is *socially asymmetric*. If the collaborators share the same task and have access to the same information, they are *task* and *information symmetric*. In general, they found the following: "*A wearable user will be able to collaborate effectively with a remote expert provided the functional and implementation asymmetries match the task and information asymmetries"*. Other work corroborates that functionally symmetric interfaces improve collaboration [75].

Another important concept in the design of collaborative media is that of *seams*. Ishii defines them as "*spatial, temporal and functional constraints that force shifting among a variety of spaces or modes of operation*", and can be either functional or cognitive [76].

A cognitive seam occurs when a different technology changes the way people work. Functional seams occur between two functional workspaces, and are most exposed when the interpersonal and shared workspaces do not overlap (such as what happens in remote CSCW). These kinds of seams are found in the expert/technician scenario.

Co-located Communication Patterns

Kiyokawa et al. compared virtual environments with AR implementations in mono, stereo and see-through HMDs, and found that real world visibility positively affects communication behaviors - requiring less amount of verbal communication in a collaborative, referencing task (see Figure 5) [77]. Additionally, they found that redundant pointing occurred more often in mono environment than stereo environments. They state that "*Generally, the more difficult it was to use non-verbal cues, the more people resorted to speech cues to compensate*". Further, they found that the percentage of perceived pointing gestures is dependent on workspace - determining that positioning the task between the subjects produced the most active behaviors and reduced miscommunications; however, participants felt a shared virtual whiteboard was the easiest for them to work on, because they could view the task space from the same perspective.

Studies have compared the differences in the effectiveness of collaboration in virtual and augmented realities [78]. This research confirms that rendering an avatar as well as the view vector in virtual environments enhances the collaborative efficiency between users for a virtual environment. It also suggests that collaborating in AR is more efficient than in VR, as key factors in recognizing the will of others are present, such as head direction, body language, gestures and facial expressions. Unlike virtual

environments, however, users performed *worse* in AR when the viewing vector was rendered. They believe this is attributed to the inaccuracies of the line, giving only a rough approximation to where the user is looking, and suggest that rendering a conical frustum (the volume of space that is visible to a user) may be better.

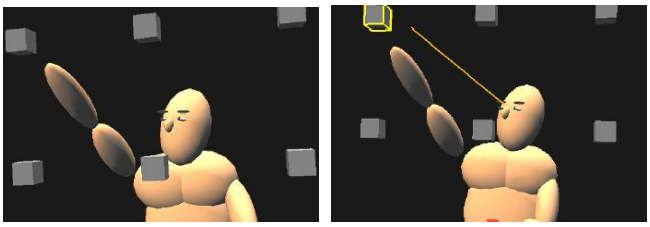

Figure 5 - Kiyokawa's experiment [78]

It has been argued that wearable computing is an ideal interface for 3D CSCW, given its mobility and how it frees the hands to interact with objects [1]. In exploring wearable interfaces for communicative purposes, it was suggested that the use of video to enhance communication may not provide as much aid in communication as once thought since users must exaggerate gestures for them to be clearly seen [79]. Other forms of systems, such as CAVE and tabletop environments typically render from one participant's point of view – distorting virtual objects for others, and making it difficult (if not impossible) to refer to the same 3D location[12][18, 80]. Wall-mounted displays for augmented surfaces, such as AR Groove - restrict interaction, as collaborators can miss many of the non-verbal cues because their focus is most often on the display and not on the physical world [81].

[12] Though a two-person workbench has been created by interlacing the stereoscopic views of each user.

Remote Communication

Remote communication removes many of the benefits of co-located collaboration, but attempts to overcome these shortcomings in a variety of ways. Early AR environments were teleconferencing-centric in design; the system provided a 2D icon registered in 3D space for each participant and leveraged on spatialized audio to maintain location awareness (yet lacked support for shared virtual objects - see Figure 6) [79]. Later live video streams were bound to fiducials, allowing participants to position participants within the workspace, and provided a virtual whiteboard [82]. This research matured into live 3D avatars, though the visual communication was uni-directional [83]. Zhong et al. describe an industrial training system in which technicians can collaborate while constructing virtual circuit boards [84]. However, remote participants interacted via a mouse and keyboard to position circuits - making interaction awkward.

Meeting Spaces

Several collaborative applications exist in city planning and architecture - where multiple participants are situated around a table, sharing 3D applications [21, 85]. Users interact within the environment using tangible objects (phicons) which are bound to virtual artifacts. Participants can interact with the environment using only their hands, yet referencing virtual objects results in occlusion problems. Participants in ARTHUR could additionally interact via a wand (for objects out of reach) or through gesturing (see Figure 7).

Figure 6 - The evolution of remote communication in AR [73, 79, 83]

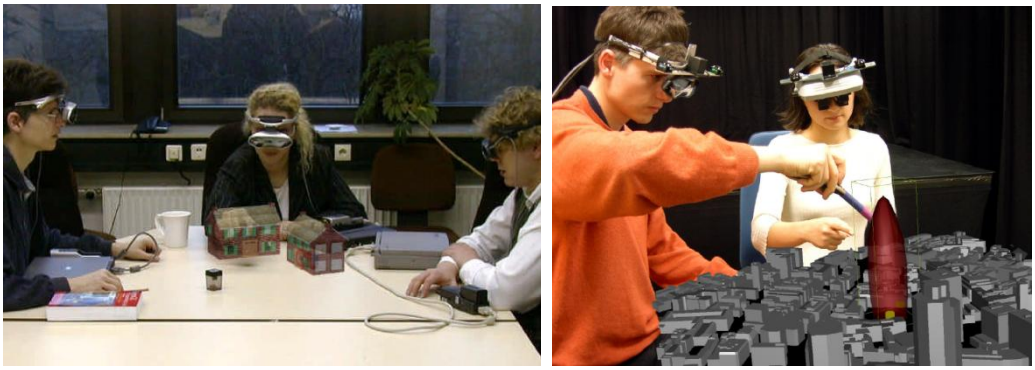

Figure 7 - The Virtual Round Table and ARTHUR [21, 85]

An early collaborative system, EMMIE, allowed participants to bring heterogeneous devices to a meeting space. [86]. When the object is identified in this environment, virtual *leader lines* connect the user's devices with the objects that meet the search criteria[13]. Regenbrecht explores the possibility of using a physical cake platter mounted with a fiducial as a tangible interaction device, allowing interaction to occur through social protocols [87]. Of interest are the multi-modal interaction techniques that

[13] EMMIE was also unique in the way it could *reduce* awareness through the use of Privacy Lamps[77].

are provided in this environment, such as the ability to hold up either a real or virtual clipping plane, or use a physical flashlight which emits both physical and virtual light for examining model. In addition, users can refer to virtual objects by coloring their parts and leaving annotations for other users – a useful asynchronous feature.

Mobile AR Interaction

AR is suited to help guide users in navigation tasks, and was first researched by Rekimoto as well as Feiner et al. [88, 89]. Höllerer et al. extended the work into collaborative AR, exploring interfaces for both indoor and outdoor users (see Figure 8) [27]. The interface consisted of world and screen-stabilized virtual objects[14], as well as a menu bar and a cone-shaped pointer. Indoor participants interacted via a desktop, and could draw the attention of outdoor users by inserting and moving virtual objects in the real world. Thus, an indoor user can give guidance to the outdoor user by drawing a path to the desired location. Reitmayr et al. describe similar work to aid in navigation to other participants [90]. Stafford advanced this idea by implementing a God-like interaction technique allowing an indoor user to physically point to a tabletop building - creating larger-than-life scale references for the outdoor participant [91].

Heterogeneous gaming has occurred across mobile augmented reality and Web technologies [92]. Based on the ARVIKA system, co-located and remote participants played a multi-user 3D Tetris [93]. Web users receive an augmented reality 2D video feed of the environment, and the selection contention between multiple participants for the virtual object is left to social protocols. Similarly, Brown studied the effect of

[14] This defines the base coordinate system. World-stabilized implies a world coordinate system.

location awareness in mixed reality environments, combining web, virtual reality and physical visitors together in real time [94].

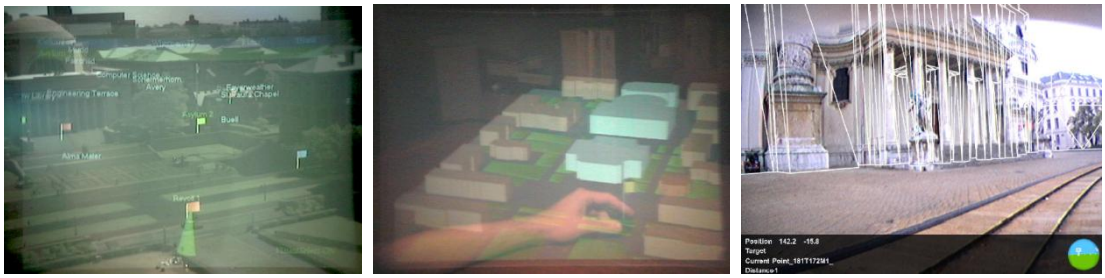

Figure 8 - MARS and Reitmayr's work [27, 90]

Generic Interaction Techniques

Though VR and AR share similar traits, there *are* critical differences in implementation. VR is "omnipotent" in that, to render the environment, all geometric information must be available, as well as the pose of the user. While a primary benefit of AR is in *not* having to render the complexity of the world, it comes at the cost of knowing very little about it. In addition, it is very difficult to capture and represent the small, non-verbal communication found in co-located environments. Nonetheless, many VR techniques have transitioned into AR. For example, Bell et al. transferred the WIM into AR, while both Slay and Rekimoto used a virtual ray of fixed length [16, 95] [31]. Piekarski has developed ray-based AR glove techniques to interact with virtual objects [96]. Later, Renevier et al. defined a set of generic interaction techniques for mobile collaborative mixed reality [97]. According to this work, interaction can cross real and virtual boundaries, creating mixed interaction techniques, such as the interaction with a physical object in the digital world and vice-versa.

The primary interaction device in *Studierstube* is the Personal Interaction Panel (PIP – see Figure 9) [98, 99]. The PIP is an augmented panel containing virtual widgets, such as buttons and sliders, and presents a contextual interface. Applications in this environment display virtual content in 3D windows, and can be referenced using the virtual pen. Of interest, this group, as well as Furmanski et al., have explored techniques for introducing occlusion between the body and virtual objects [100, 101].

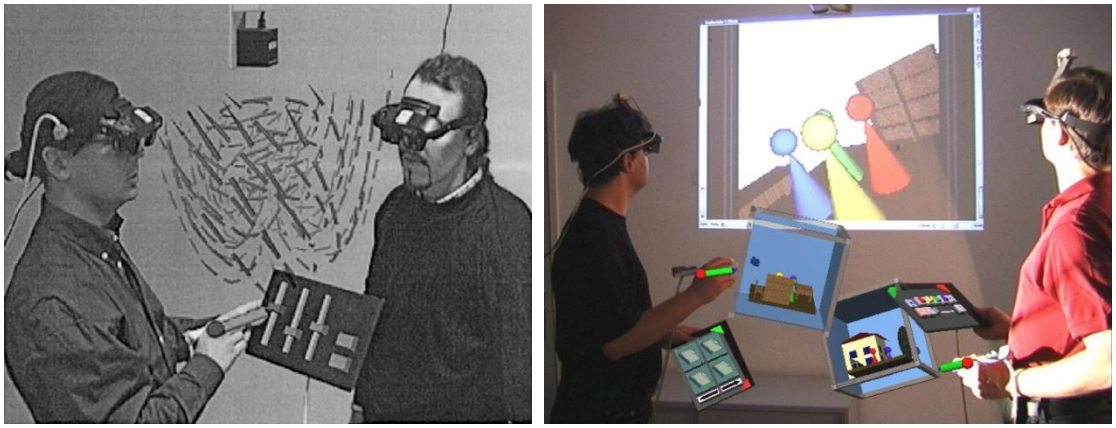

Figure 9 - The PIP [102, 103]

In VOMAR, Kato et al. have attempted to create a universal tool that could pick up both (assumedly light) physical and virtual pieces as well as push them or drop them [104]. However, this device does not transfer to life-size-scaled environments

To a lesser extent, audio has been used to augment an environment with auditory cues, with the intent of passively increasing awareness and summarize activities relevant to the user [105]. Other interaction techniques include agents who act on the behalf of the user to move something, such as Anabuki's Welbo [106]. However, the virtual agent cannot interact with the physical world.

Tabletops accommodate collaborative AR as well, but are less common (and consequently, are only briefly mentioned). Tabletops allow for participants to interact without the need for HMDs, typically contain a stationary camera and wall-mounted display, and leverage tangible interaction techniques. Examples include a science fair exhibit and the Augmented Groove [107] [81]. However, the display diverts attention away from the physical space, making the ability to refer to virtual objects quite difficult. More traditional tabletop systems include Ullmer's metaDesk and Wellner's DigitalDesk [108, 109]. These interfaces are tangible as well, providing tactile feedback for virtual objects (such as the phicons previously mentioned). However, the augmentation is projected on a 2D surface, and full 3D interaction is often limited.

Reference Representations

Every reference must have a representation, which most often manifest visually. To increase awareness of selected objects that fall outside the view of the user, Biocca and Tang introduced the "*attention funnel*" [110]. By creating a tunnel from a series of concentric squares, the user can be directed to the object. While this technique is especially useful in increasing the user's awareness of the surrounding environment, this representation occludes a significant amount of screen real estate. Alternatively, Tönnis describes how to increase location and presence awareness of other drivers [111]. Similar in nature to other head-up, in-car displays, the system directs the gaze of the user (using an arrow) to the location of other cars that do not appear in the field of view of the driver; however, this technique suffers from scalability issues as well.

2.5. Summary of Background

Prior work in CSCW has demonstrated that when referencing is not supported, there is a hindrance to collaboration. It has also shown the importance of non-verbal communication – which is negatively impacted when digitally replicated. To raise *awareness*, there have been several attempts at incorporating a shared visual channel within the environment, and to provide remote participants the ability to annotate the visual field of participants; however, many of these techniques are 2D (and thus *projected* into the environment) and are temporal (becoming stale if the pose of the remote participant changes).

Because many interaction techniques in AR parallel those in VR, we have summarized a variety of methods for selecting and referring to virtual content; however, many of these techniques rely on mathematical algorithms to function, and are not applicable to physical objects unless their pose and geometry are known in advanced. We have summarized related theoretical concepts from this field as well.

A majority of this chapter naturally focuses on the field of collaborative AR. We summarized theoretical concepts (such as requirements), and described the asymmetries that are found in collaborative AR, including *information*, *implementation* and *social* asymmetries – which ultimately affect referencing techniques. We saw that workspace placement affects communication, and that when non-verbal cues are not present, users compensate through additional speech. We have presented several application domains of collaborative AR, including remote communication, meeting spaces, and mobile systems. Finally, we summarized the generic interaction techniques in AR, as well as context-sensitive methods of reference representation.

CHAPTER 3

SPATIAL AND MULTI-MODAL REFERENCING IN MIXED REALITIES

In this chapter, we present our early work to address some of the referencing difficulties described in Chapter 1. We begin by describing the development of a collaborative molecular modeling environment, including our goals, referencing problems we encountered, client interaction (including referencing techniques), and supporting architecture. We then discuss how real-time shaders can be used to represent the area of interest. Finally, we present our work in developing a multi-modal, hands-free technique that can refer to content at arbitrary depths.

3.1. AMMP-VIS

The impetus behind studying inter-referential awareness stems from our development of an affordable, immersive system that allows biologists and chemists to manipulate molecular models via natural gestures [112, 113]. Molecular modeling is an important research area, helping scientists to develop new drugs against diseases such as AIDS and cancer. Prior studies have demonstrated that immersive virtual environments have unique advantages over desktop systems in visualizing molecular models. The system allows participants to receive and visualize real-time feedback from a molecular dynamics simulator as well as share customized views, and provides support for local and remote collaborative research. Our virtual environment is developed around a centralized molecular dynamics simulator, AMMP, which provides the application with molecular mechanics and dynamics functionality [114-116]; therefore, manipulations of virtual molecular models are governed by the molecular potential field.

User Interaction

The usability goals of the system were to 1) have natural interaction between the users and the models they were manipulating and 2) create an arena in which biologists could communicate and collaborate with one another. We felt that interaction techniques should be as intuitive as possible, leveraging from pre-existing mental models, and should avoid the awkward mapping of 2D input devices to 3D space where possible. For collaboration to occur, we further believed (as does the literature) that each participant must be aware of other users in the system (*presence awareness*), their current area of focus (*attention awareness*), as well as any manipulations they are performing (*action awareness*) [117].

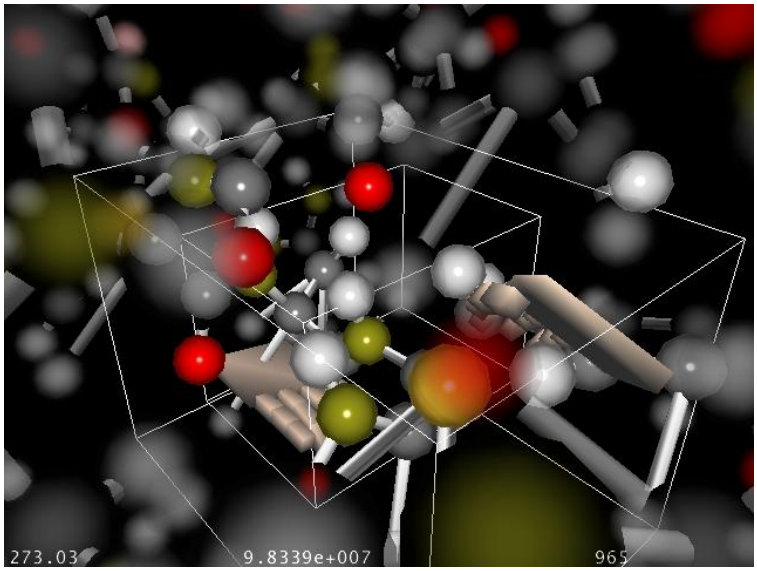

Figure 10 - Example of attention awareness

Scientists interact with molecules primarily through the natural hand gestures of pinching, grabbing and moving - all of which are tracked via a data glove. Each user is

represented by an isometric hand avatar which behaves according to the user's gestures. Real-time updates of hand location, orientation and finger bends are continuously being disseminated to other clients. The inclusion of hand avatars heightens the awareness of both presence and actions of other users (see Figure 10). Perhaps the strongest indicator of presence in our system is the incorporation of a shared audio channel. Speech is an effective (and rapid) method of communication, and in combination with the virtual hand, allows participants to refer to objects through deixis[15].

Though the first implementation served primarily as a visualization tool, the importance of performing tasks at varying granularities soon surfaced. When working with large molecular structures[16], specifying sub-regions for an action (e.g. translation) is both tedious and inefficient if atoms are selected individually. Further, even though support for referencing was present through virtual pointing (i.e. using the hand avatar), generating a reference to an individual atom was difficult given the lack of visual features in the data. The problem was one of how to effectively represent the area of focus for individual users in a collaborative environment - where several areas of interest may exist. Given the complexity and density of large models, selected areas could be occluded by atoms that fell within the line of sight – something especially pronounced for those located near the viewpoint. To overcome this, it was decided that, at a minimum, portions of the model that were *not* of interest should be semi-transparent – essentially, *culling out information* to draw the attention of others. This approach allows the model to be seen as a whole, while keeping important contextual and

[15] The act of pointing which is concomitantly supported by deictic speech (e.g. "*this*", "*that*", "*those*"…)
[16] Around 12,000 atoms - but much larger molecules exist.

spatialization information. However, a stronger contrast is possible through the use of programmable shaders (see section 3.2).

To allow participants to select and reference at varying granularities, the environment allows participants to generate a 3-dimensional bounding box; this technique served the dual role heightening attention awareness by dynamically defining an area of interest, as well as selecting *sets* of objects (atoms). The design of this technique leverages an existing mental model: scientists pinch (index to thumb) and drag in 3D space. This action is analogous to selecting multiple objects in 2D desktop environments (a.k.a. rubberbanding). During its creation, the system renders the 3D bounding box, which can be seen by other participants. The bounding box remains active until the next bounding box is defined and is completed at the release of the pinching motion. Atoms whose position lies within the defined region are highlighted – and represent the area of interest for that user. Further, once an area of focus has been determined, the user can manipulate that region - performing tasks such as translating or rotation. To accomplish this, the user makes a *grabbing* gesture within the bounds of the box, and then moves and/or rotates the data glove for direct manipulation of the atoms. Any new atoms that may fall within the bounding box (due to the actions of other users) after this process are not considered to be in the area of focus for that user. Grabbing gestures outside of the box have no effect in the system. The manipulation ends with a release from the grabbing state. Any updates to the molecule are sent to AMMP, which returns the molecule's energy (in KJ/mol) and propagates atom positions to all clients.

There were several other HCI factors to consider with this technique. While users see the bounding boxes of others, the atoms selected by them were intentionally *not*

highlighted to provide a sense of ownership to which area one is currently referring; it was thought that when multiple participants (and bounding boxes) are present, it would be difficult to determine which user owns a region unless the reference is put into context *while* it is being made; in other words, the visibility of the bounding box alone may be ambiguous when multiple boxes are present. In retrospect, a more-appropriate choice is to highlight *any* referenced region and use a simple color-coding scheme to denote ownership. Also, because bounding boxes remain visible until they are redefined, they may ambiguously refer to content that is no longer being considered. Thus, it is left as a social protocol for participants to remove outdated references.

There are other factors that may positively influence referencing as well. First, real-time hand updates are continuously being sent, including location, orientation and finger bend information; this allows users to see the actions of others, including pinching (when defining a bounding box) and grabbing (manipulating the molecule). Because translations and rotations of atoms are seen in real-time, it is also possible to draw the attention of others through physically "shaking" the region of interest. Referential support is also provided through spoken communication, which was transmitted via VoIP.

There are several interesting properties of using a dynamic bounding box in referencing tasks. Primarily, it has the capability of referring to more than one object (as do many image plane techniques); however, it accomplishes this not through projection, but in defining 3D space. Because of this, *spatial referencing* is possible - even when no objects are being referenced. For example, this is useful in clarifying statements such as "move that item *here*"; the word "*here*" refers to an arbitrary space in which no items

may exist. Consequently, the technique is also multi-modal in that it does not use geometric intersections to denote space; thus, it could be extended to refer to physical objects (and physical space). A drawback of using this approach as a referencing technique is that it relies too heavily on depth cues (from stereoscopy and occlusion), which was overcome (in our implementation) through highlighting; this makes it less desirable for referencing when geometric data of physical content is unknown.

Of interest was how to help scientists notify others of precisely what they are referring to, overcome occluded viewpoints, and clarify which context they are viewing the data - as the system supports the standard molecular visualizations, including "ball and stick" and space-filling algorithms. An inefficient approach is for users to stop their current task, move to the location of the other user and change their settings to match that of the other user; however, this loses the context in which they were previously working. In VR environments, it is a trivial to render the world from another viewpoint, which allows scientists to share views and reduce context-switch time. Though this concept is not new, we felt that including it would help in clarifying references.

System Architecture

At the core of the system is AMMP [115, 116], a free 2D (or command line) molecular mechanics simulator which provides the ability to model and analyze the dynamics of molecules, proteins and nucleic acids. Our architecture is based on a centralized version of server, which was modified to allow molecular state to be updated via network messaging. The server was written in C, is concurrent (using *pthreads*), communicates using Berkeley-style sockets (TCP/IP), and allows clients to freely connect or disconnect with the system without any adverse effects. An additional

benefit to this design is that molecular state remains from session to session until a new model is loaded, allowing scientists to work asynchronously. We take a distributed MVC (Model-View-Controller) approach to the design, decoupling the molecular data processing from the visualization (as well as leveraging from modern graphics hardware). This offloading of responsibility allows the client to render at higher framerates as well as customized visualizations. Any updates to molecular state (such as translation and rotation of atoms) are sent to this server, with new energies calculated and broadcast to all interested clients. Molecular mechanics therefore govern all client-side interaction (see Section 3.3 below). Given its current hardware configuration, AMMP can distribute real-time updates with just under 2000 atoms in the molecule.

Our system consists of any number of clients, the AMMP server, an orientation server, and communications server (see Figure 11). The orientation server is responsible for coordinating user state data by receiving then broadcasting client updates, such as hand location and orientation, finger states, bounding box information, scale, viewing position (which allows for quick switching of views), and user visualization preferences. This sub-system generates the most network traffic of all the subsystems – as the avatars are consistently updating their state; data sent from a client is assumed to be known by that client, and is therefore only propagated to others in the system. Given that several VoIP solutions already exist (many of which are free), in the interest of cost-efficiency, participants communicate using Skype (www.skype.com) which can support multiple participants.

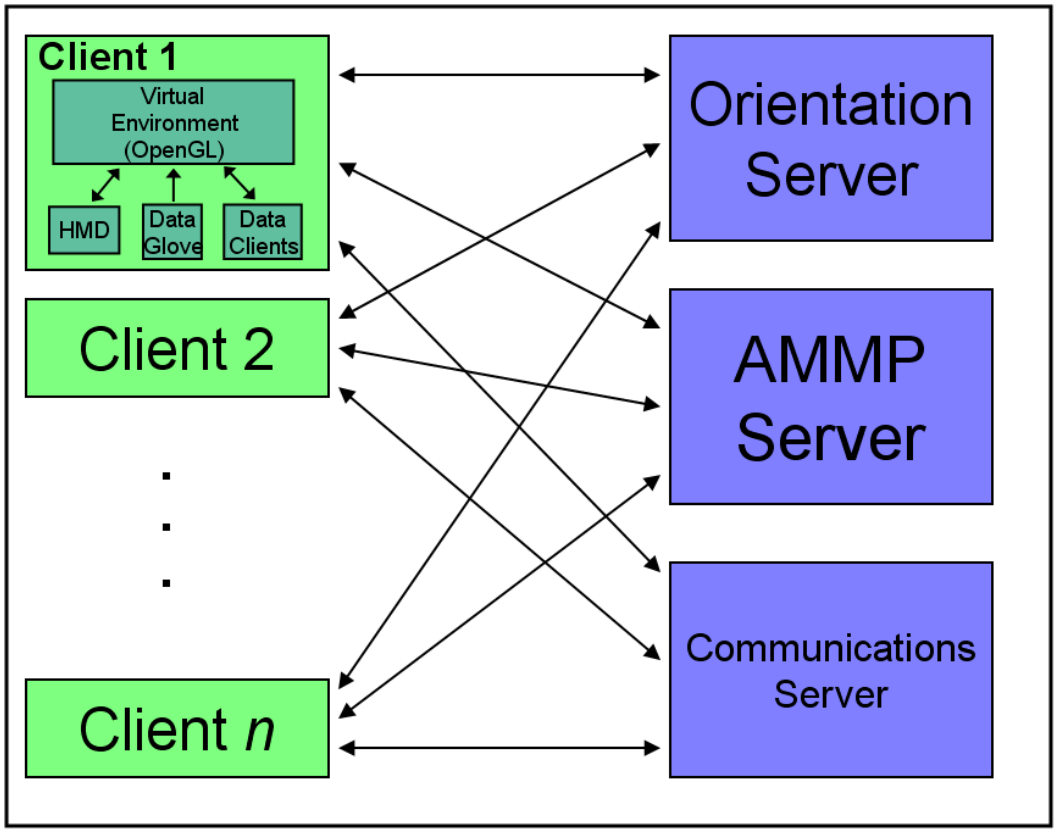

Figure 11 - System Architecture of AMMP-VIS

To achieve low cost, the client is designed to conform to heterogeneous hardware, including current generation consumer-level hardware; users can opt to use expensive equipment, such as high-end HMDs and data gloves, or use less-expensive displays, such as shutter glasses, or moderately-priced Icuiti Video Eyewear ($500) with a tracker, in conjunction with inexpensive commodity game gloves. Our test machine was a 2.6GHz machine with an ATI Radeon 9800 card, a VFX HMD and a P5 gaming glove (with infrared tracking). The total cost for our client test system was approximately $5000, much of which comes from the HMD. The client was written in OpenGL/C++, and provides a stereoscopic view through interlacing or page flipping.

3.2. Emphasizing the Area of Interest Using Shaders

When working with complex molecular models, where either the atom count is high or atoms are clustered together, the area of interest may be occluded by atoms or bonds that fall within the line of site, making it difficult, if not impossible to view the area. While HMD resolution is often seen as a limiting factor in most environments, there are considerably fewer pixels to render than a typical display, resulting in rendering rates that drastically exceed refresh rates; we can therefore leverage off of the recent availability of programmable graphics hardware (i.e. fragment shaders) to include an efficient blur shader in the environment – alleviating the occlusion problem while maintaining context and real-time performance [118]. While several physically-correct depth-of-focus algorithms have been created, we were interested in an efficient approximation specifically designed for this environment [119].

It was shown by Liu, et al. that the inclusion of shaders allows for a higher quality representation of atoms and improves users' visualization experience [120]. OpenGL, by default, shades objects by applying the Gouraud model – an algorithm that suffers from its inability to accurately render effects, such as specular highlights (see Figure 12). Fragment shaders work on a per-pixel basis, and allow the system to generate better quality images and more special effects, such as specular highlights, or effects like depth-of-field or "glowing". Figures 12, 13 and 14 compare the rendering of molecular models using the fixed-function pipeline and fragment shaders.

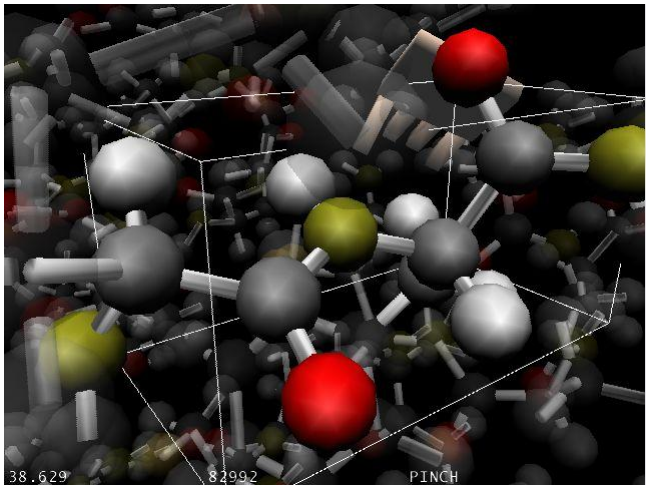

Figure 12 - Gouraud Shading - user defining area of interest

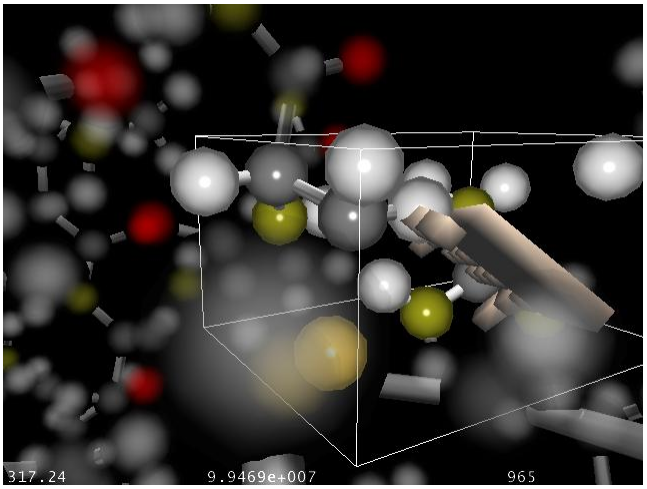

Figure 13 - Shader in ball-and-stick mode

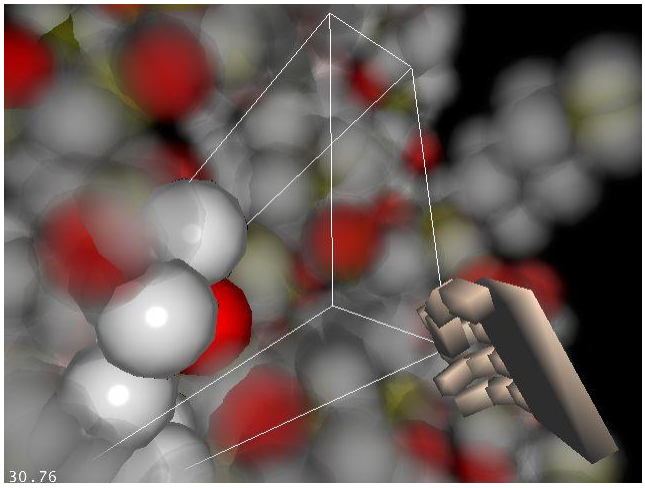

Figure 14 - Shader in space-filling mode

The shader implements the Phong shading model when atoms are selected; in the unselected state, they appear blurred. Blurring is achieved by using an algorithm similar to the Lambertian diffuse lighting equation, dismissing the specular highlight. The *alpha* component of each fragment is calculated by finding the dot product of the fragment normal *N* and the view vector ($N \cdot V$). This approach allows the center of the atom to remain opaque (up to a constant), with the alpha value falling to zero around the edges, creating a blur effect. Thus:

$$alpha = max\,(dot\,(N, V), 0.4);$$

Equation 3-1

This algorithm is appropriate to use when *dot* (*N*, *V*) decreases around the edges of a model, such as with spheres and cylinders, but is ineffective for use on concave objects. The value of 0.4 was chosen based on designer preference. Our system retains real-time performance (>30 fps) even with the shader enabled, with tests running nearly 3000 atoms visible at once; performance remains interactive (17 fps) with approximately 5000 atoms visible. Figure 15 compares the performance time between our shader versus the fixed-function pipeline (the default rendering method in OpenGL). Though the performance runs at nearly half that of the fixed-function pipeline, the figure clearly demonstrates the "wasted" rendering that occurs when the frame rate exceeds the display refresh rate.

It should be noted that if the frame rate drops below an undesirable value, the user can disable the shader (by toggling back to the fixed-function pipeline) when viewing the model from a distance to improve performance, and enable it again after zooming in

on a smaller section for a more detailed view. Given the relative simplicity and repetition of meshes, we are currently investigating ways of increasing the frame rate (such as through billboarding) to allow chemists to visualize extremely large models (those with more than 10,000 atoms).

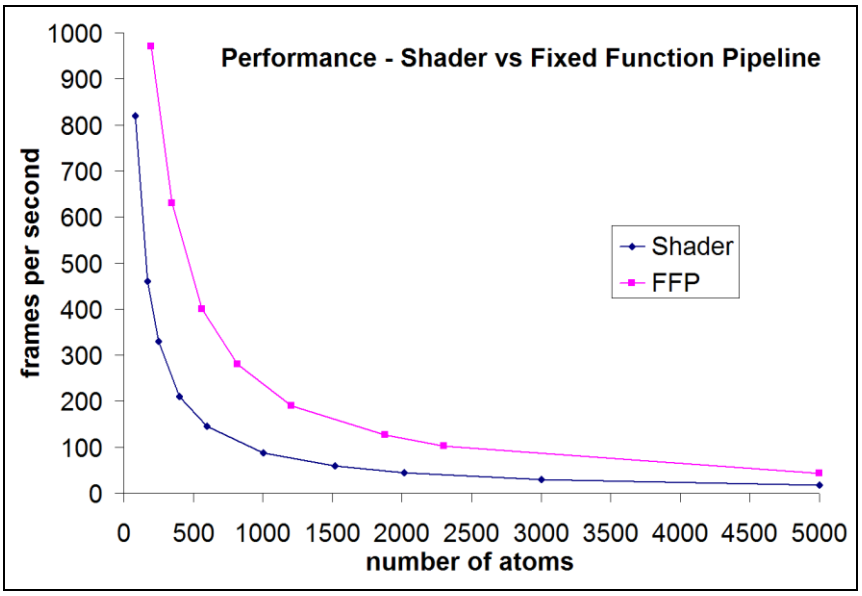

Figure 15 - Shader vs. fixed-function pipeline performance

3.3. The Transition to AR

Co-located AR increases efficiency between pairs of participants, allowing them to view the world in a more natural manner, and maintain work context [73, 77]. Thus, our interest shifted into transitioning the environment into the domain of AR, with the goal of allowing biologists and chemists to visualize molecular structures while seated at a conference table (see Figure 16). The first implementation was a direct modification of previous work, updating the client to display a video-textured background provided from a head-mounted camera. Because the molecular mechanics server was decoupled from the interface, no modification to the server was necessary. Further, the generic

client design permitted a straight-forward replacement of tracking systems; the head tracker was updated to receive VRPN[17] updates from an acoustic tracker (the InterSense-900)[18]. However, there was no way of interacting with the molecule: the system allowed only visualization. While the infrared tracker of the data glove was appropriate for interacting in virtual environments, tracking requirements are much stricter for AR. The hand must be tracked in world coordinates such that the physical and virtual worlds are accurately aligned within fractions of a degree. While the HMD was easily modified to mount an additional tracker, a second tracker that could be mounted on the glove was not available. An alternative tracking technology was needed, such as the vision-based ARToolkit [25]. It was at this time that the decision was made to transfer into a more flexible environment for designing AR applications: DART [121].

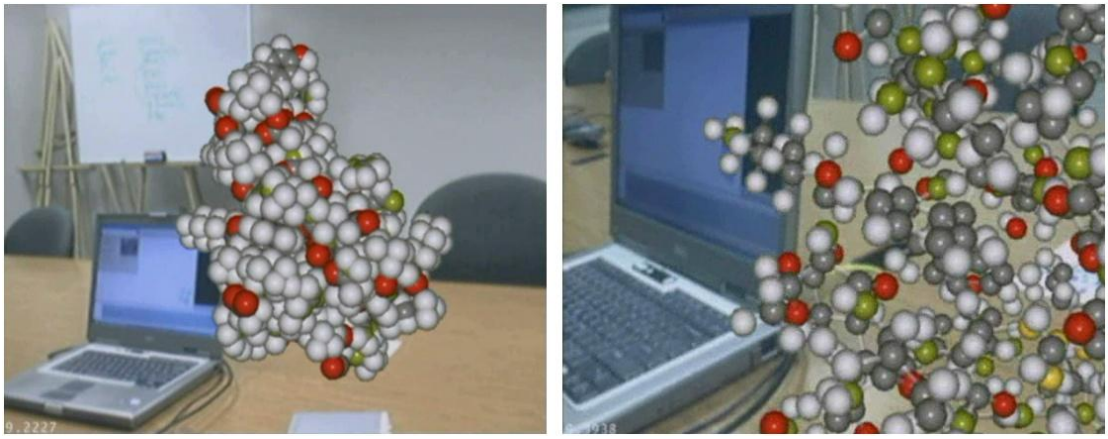

Figure 16 - Early AR environment for molecular modeling

In its current state, it was not possible to interact directly with the molecule, yet it *was* possible for participants in other realities; because of a centralized server, a

[17] Virtual Reality Peripheral Network. Copyright Russell M. Taylor II at the University of North Carolina at Chapel Hill

[18] Though this was not as straight-forward as expected; the coordinate system needed to be remapped.

participant could manipulate the molecule in the VR environment, and changes would be reflected in the AR environment (i.e. it was possible to see changes in structure and still receive energy updates). Such cross-reality collaboration is indeed possible (as well as useful), and raised a series of questions in how to best support inter-referential awareness in a wide variety of scenarios.

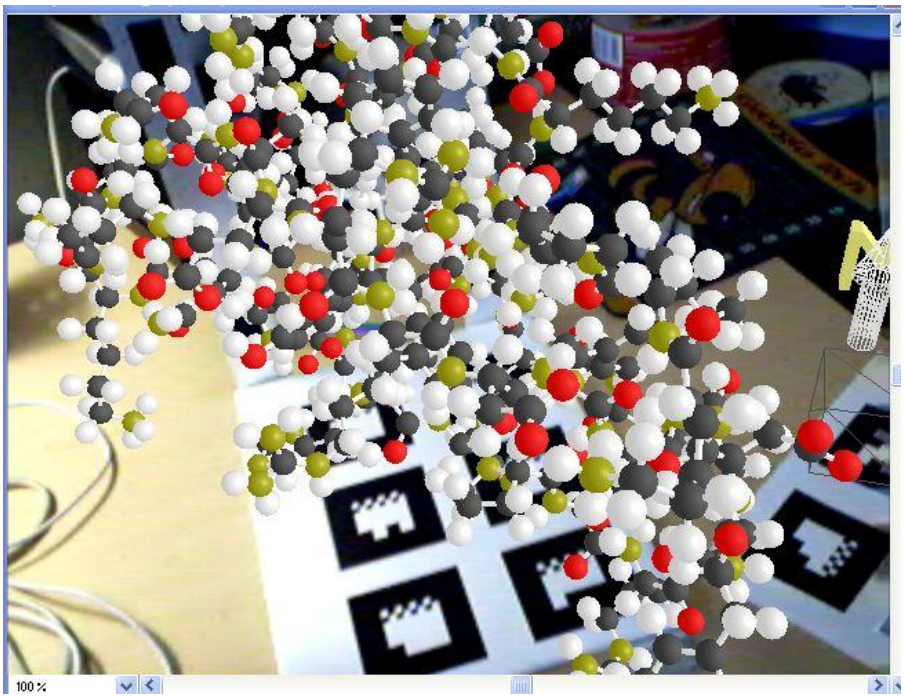

Figure 17 - The environment transitioned to the DART platform

Further, one of the members of the project expressed interest in supporting remote collaboration. Given the myriad of factors that were surfacing, there was a need to better understand how inter-referential awareness occurs in AR, the factors that affect it and methods that support it. These questions eventually led to the generic CSCW framework found in Chapter 4.

3.4. A Multi-modal Referencing Technique

In distributed collaborative augmented reality, participants are geographically separated and are capable of synchronously visualizing and manipulating physical and virtual content. An asymmetric subset of this configuration is that of a technician/expert pair. As described in Chapter 1, the technician wears a head-mounted display equipped with a camera - allowing an expert to view the workspace from the technician's point of view; depending on the implementation, the technician may be provided with a view of the expert's face. Experts need the ability to refer to content in the technician's space during the course of collaboration, yet very little (or no) geometric data of the remote environment is available; thus, virtual selection techniques are not applicable. While annotating the technician's view in 2D is certainly possible, references made using this technique are valid only if the technician remains still[19]. In the medical scenario, surgeons need the ability to make references at arbitrary depths, such as the surface of the skin (for cutting), below the skin, to embedded virtual objects, or within virtual objects. Further, they must be able to refer to content in a hands-free manner, as their hands are often occupied.

The above scenarios clearly indicate that new methodologies are required to achieve effective referencing in collaborative augmented reality. First, these techniques must be flexible enough to interact with both physical and virtual objects in a wide variety of scenarios. Second, given that there is often little geometric data available from remote environments, it can be difficult to refer to objects in space other than your

[19] And further, do not meet the requirements of AR according to Azuma [23].

own. Third, it can be difficult to refer to objects at arbitrary depths, such as those embedded within one another; even if geometric data is present, these objects cannot be physically "touched". Using the method described here, participants have the ability to consistently generate references to both physical and virtual content at arbitrary depths, without knowledge of the geometric structure of the object of reference. In addition, it can be extended to embed references in remote environments.

As with many other vision-based techniques, ours is inspired by stereoscopy. It requires the initiator to view the object of reference from two viewpoints, casting two rays which are ultimately culled; the minimum distance between these rays represents a point of reference. This approach does not rely on knowledge of the environment, and thus works for both physical and virtual objects. Additionally, it requires only relative positioning information, though we have implemented it using absolute positioning. It is mathematically straight-forward in concept and implementation (running in constant time), can be used as a generic referencing technique to make the system aware of space, and can be modified to allow a remote expert to make references. It was designed specifically to reduce the problem of referring to embedded virtual objects within a physical structure. This technique, however, is *not* intended to construct virtual models, but used to embed a 3D reference into an augmented space. While it is possible for the rays to remain present on the screen (in order to heighten awareness of the object and provide depth cues), Kiyokawa shows that users perform *worse* in AR environments when a viewing vector is rendered [78].

Our technique is a two-step approach [122]. The user first views the object of interest from one viewing point, casting a ray; from a secondary view, the user casts an

additional ray at the object. To make the selection more understandable, we provide an optional, semi-transparent crosshair.

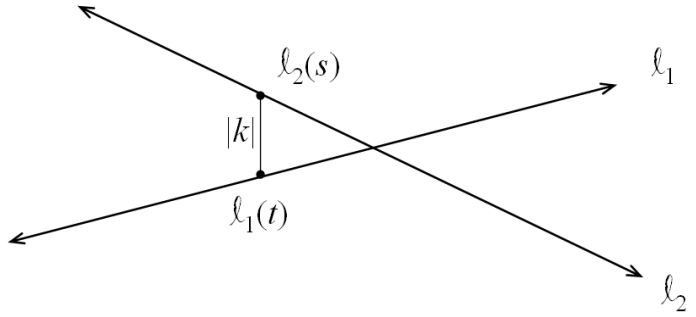

Figure 18 - A skew pair

Given two unique viewing vectors ℓ_1 and ℓ_2 of an object, it is possible to generate relative depth information in 3D space. These lines do not necessarily intersect, creating a *skew* (see Figure 18). The cross product of these vectors gives a third that is orthogonal to both. We must find a point on each line such that their distance represents the minimal distance *k* between the two lines. Note that when $k = 0$, the lines intersect. This can be represented with the following equality:

$$\ell_1(t) - \ell_2(s) = k\,(\vec{V_1} \times \vec{V_2})$$

Equation 3-2

Replacing with the vector equation of a line yields:

$$a + t\cdot\vec{V_1} - b - s\cdot\vec{V_2} = k\,(\vec{V_1} \times \vec{V_2})$$

Equation 3-3

where a and $\vec{V_1}$ represent the position and viewing direction from the first viewpoint, and b and $\vec{V_2}$ represent the second. Let $\vec{V_3} = (\vec{V_1} \times \vec{V_2})$. The equation can be viewed as a simultaneous equation of the three variables where t is the scalar associated with ℓ_1, s represents the scalar associated with ℓ_2, and k represents the minimum distance between the skew pair. Solving for these three variables, we put the equation into matrix form:

$$\begin{bmatrix} V_{31} & -V_{11} & V_{21} \\ V_{32} & -V_{12} & V_{22} \\ V_{33} & -V_{13} & V_{23} \end{bmatrix} \begin{bmatrix} k \\ t \\ s \end{bmatrix} = (\vec{a} - \vec{b}) \text{ and thus } \begin{bmatrix} k \\ t \\ s \end{bmatrix} = \begin{bmatrix} V_{31} & -V_{11} & V_{21} \\ V_{32} & -V_{12} & V_{22} \\ V_{33} & -V_{13} & V_{23} \end{bmatrix}^{-1} \begin{bmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{bmatrix}$$

Equation 3-4

Let the determinant

$$d = V_{31} \begin{vmatrix} -V_{12} & V_{22} \\ -V_{13} & V_{23} \end{vmatrix} - V_{11} \begin{vmatrix} V_{32} & V_{22} \\ V_{33} & V_{23} \end{vmatrix} + V_{21} \begin{vmatrix} V_{32} & -V_{12} \\ V_{33} & -V_{13} \end{vmatrix}$$

Equation 3-5

Solving for k, s, and t gives the following:

$$\begin{aligned} k = {} & (((V_{13} \cdot V_{22} - V_{12} \cdot V_{23}) / d)(a_1 - b_1)) + \\ & (((V_{11} \cdot V_{23} - V_{13} \cdot V_{21}) / d)(a_2 - b_2)) + \\ & (((V_{12} \cdot V_{21} - V_{11} \cdot V_{22}) / d)(a_3 - b_3)) \end{aligned}$$

$$\begin{aligned} s = {} & (((V_{33} \cdot V_{12} - V_{32} \cdot V_{13}) / d)(a_1 - b_1)) + \\ & (((V_{31} \cdot V_{13} - V_{33} \cdot V_{11}) / d)(a_2 - b_2)) + \\ & (((V_{32} \cdot V_{11} - V_{31} \cdot V_{12}) / d)(a_3 - b_3)) \end{aligned}$$

$$t = \begin{aligned} &(((V_{33} \cdot V_{22} - V_{32} \cdot V_{23}) / d)(a_1 - b_1)) + \\ &(((V_{31} \cdot V_{23} - V_{33} \cdot V_{21}) / d)(a_2 - b_2)) + \\ &(((V_{32} \cdot V_{21} - V_{31} \cdot V_{22}) / d)(a_3 - b_3)) \end{aligned}$$

Equation 3-6

Multiplying the scalars *s* and *t* by their respective viewing vectors yields two points, whose distance is *k*. The average of these two points is then used to create a bounding sphere of radius *k*/2. If the lines are parallel, the determinate will be 0, and no sphere is rendered. Additionally, if *s* and *t* are negative, the point of reference is behind the head, and consequently is not rendered.

Figure 19 - Referencing at arbitrary depths

Interpreting *k*

Given that *k* represents the minimum distance between the skew lines, in our original design, it was used as a way to visualize the error that occurs either through tracking or user error introduced from the selection; it was assumed that the user *intended* to create an intersecting line pair. Through experimentation, however, we have found that this value can be used as bounding geometry for the referenced object. In addition, we have chosen a sphere as bounding geometry. Depending on the kind of

reference needed, *k* can be visualized as a line running the length of an object, which may be useful when selecting objects of disproportionate length.

If this technique is used to create bounding geometry, then its use is restricted: the distance between the two viewing positions must be greater than the size of the reference. Otherwise, the minimum distance between the skew pair occurs *behind* the user, which will result in *s* and *t* being negative. Though using perspective projection techniques may help to overcome this, our preliminary implementation does not support this.

Use in Remote Referencing

The technique can be modified to allow for remote selection by broadening how the projected viewing vectors are defined. In the general case, the user is presented with a fixed crosshair. By allowing the crosshair to move based on mouse position, it is possible for the expert to cast rays. Using the same methodologies of raytracing, we can project rays through the image plane via mouse interaction. Thus, the expert can select objects of interest in the view of the technician.

Though it is yet to be implemented, we believe that freezing the frame - much in the same way that Bauer did - will allow this method to become easier to use [15]. Because orientation information is sent with each frame, the reference can be calculated "offline". We imagine this will be beneficial when referring to small objects, where precision is needed.

3.5. Discussion

Though we have not conducted formal user studies on the efficacy of the skew line technique, initial user feedback suggest the technique is intuitive and does not place

undue burden on the users. The participants were familiar with AR; we are interested to receive feedback from those who are unfamiliar with the area. Unlike many VR selection techniques, no burden beyond wearing the HMD is incurred.

As described previously, there are limitations to this referencing technique, such as its dependency on an accurate tracker, the user's ability to accurately position the crosshair, and in the case of the remote expert, the resolution of the camera and display. In addition, the user is required to perform a secondary step when embedding a reference (at the cost of efficiency), and is currently only able to create bounding geometry for relatively small objects. Thus, the relative movement of the user is directly dependent on the size of the desired reference. How well this technique performs in real AR applications has yet to be determined, though we imagine it will be a useful technique when objects are out of reach or are embedded within one another.

Finally, there is ongoing debate on whether the rays should be visualized during selection. Though rendering a viewing vector has been shown to be ineffective, it may be useful in the expert/technician scenario, heightening the awareness of general direction of the object as well as yielding depth information.

3.6. Summary

It was through the work in this chapter that the basic problem of inter-referential awareness was exposed. In the molecular modeling domain, there was a need to be able to generate spatial references. To address this, we provided a dynamic 3D bounding box, which could be seen by other participants during creation. This process allows users to select and refer to arbitrary 3D space. Unlike image plane techniques, this approach is capable of referring to empty space. Because the technique does not rely on geometric

intersection, it can be used to refer to physical content as well, though it may suffer from occlusion cues. We believe this referencing technique to be intuitive to those with even minor computing backgrounds, as it is analogous to the rubber-banding technique found in traditional 2D desktop systems. This technique was also supported by including real-time shaders to increase attention awareness. Further, the shader provides an un-occluded view of the referenced atoms – helping to reduce potential asymmetry between viewpoints. In transitioning the environment to AR, we realized that we needed ways to refer to both physical and virtual content in both local and remote environments. Our skew line technique allows users to generate references at arbitrary depths to both physical and embedded virtual artifacts in the environment. It can be extended to allow remote participants the ability to generate 3D references to content in the local environment. Further, such an approach does not require the use of hands, and thus is appropriate for scenarios such as collaborative AR surgery.

The research from this chapter demonstrates the complexity of referencing tasks in collaborative AR. Beyond the configurations of modality and space, if designers are to support inter-referential awareness among collaborators, they must understand the process of referencing, the factors that affect it, the methods that support it. Chapter 4 provides a generic, process-driven framework that can be applied to encompass domain-specific factors.

CHAPTER 4

A FRAMEWORK FOR INTER-REFERENTIAL AWARENESS

As computing evolves into new domains, we need methods of understanding how humans interact within them. The fundamental task of generating references seems nearly trivial; after all, we are well equipped to do this in the physical world and only rarely experience difficulty. With the arrival of CSCW, researchers began to realize some of the difficulties of referencing in computer-mediated environments; it is now understood that references comprise a critical part of communication, and must be supported if successful collaboration is to occur. Collaborative desktop applications exposed the fundamental difference between working in the physical and computer-mediated worlds; participants often have disparate views, and can no longer rely on many non-verbal forms communication, such as gesturing and eye gaze. As computing transitioned into 3D environments, it brought with it new possibilities in supporting collaborative tasks. In its initial concept, VR was supposed to provide us a synthetic world that closely parallels our own, but as researchers discovered, even basic tasks could be problematic; the non-tangible nature of virtual objects in combination with new spatial properties introduced interaction challenges. Just as in early CSCW, VR researchers realized that participants had difficulty making references, but the problem seemed to have compounded.

It seems that by re-introducing the physical world back into the workspace, we would be able to leverage off of the way that we naturally interact (and reference) within the environment. While this is partly true (in co-located scenarios), the difficulty of interacting with virtual objects is still present. Remote scenarios - such as the

expert/technician configuration – require the expert to become aware of the remote space as well as refer to the physical and virtual objects within it. AR also allows for interesting possibilities, such as virtual objects embedded within physical ones (e.g. the medical scenario). Because augmenting the view of the user occurs *last*, physical objects do not naturally occlude virtual ones, generating conflicting depth cues when physically pointing.

Support for inter-referential awareness is equally critical in AR as other environments, yet it is our observation that no systematic analysis has specifically addressed it; thus we began our work in designing a framework to better "mentally grasp" all of the factors that relate to this ostensibly simple task [123, 124]. Our hope was to provide methods for better understanding inter-referential awareness, help designers consider the factors that are involved when designing new referencing techniques, and analyze when referential ambiguity might occur.

We designed the framework to encompass factors and concepts in AR, but soon realized that it could be more broadly applied to the area of CSCW. Given the heterogeneity of media and myriad of interaction techniques that exist in general groupware however, the framework needed to be flexible. In this chapter, we present a unified and systematic way of encapsulating these factors through the creation of process-driven framework. It provides a formal method for describing inter-referential awareness, and serves as an approach that interface designers can use to better organize domain-specific factors. We present the framework generically - enumerating aspects found across the field of CSCW. However, we incorporate themes from AR where appropriate, show how this framework models inter-referential awareness in

collaborative AR environments, and enumerate some of the specific factors found in this domain. Next, we present taxonomy that classifies environmental factors that influence referencing. The framework and taxonomy presented in this chapter have been cyclically refined through user studies.

4.1. A Process-Driven Framework

The difficulty in generating meaningful references to objects within the environment varies with the communication medium, the application domain and context. Many distributed groupware systems support the use of audio, video, text, 2D and 3D space, and may be either synchronous or asynchronous. Devices for interacting within these interfaces can be awkwardly stretched to work across different dimensions, such as a mouse interacting with 3D content. Objects may be in any number of states, and may be referenced through various criteria. Though we most often focus on the challenges of distributed virtual space in CSCW, objects need not be digital - as real-world objects, such as those in mixed reality environments, are part of the natural context in shared spaces. Similar to virtual 3D content, these objects acquire additional spatial properties, such as proximity to participants and occlusion by other objects.

The Inter-referential Pipeline

We begin at an abstract level, viewing inter-referential awareness as a sequential process of *selection*, *representation* and *acknowledgement* (see Figure 20). In this context, the environment contains an implied set of participants and set of objects.

We describe *selection* as an atomic process in which, through the actions of an individual, a set of objects is chosen for reference. It is possible to decompose selection into a *cognitive cycle* (the mental process of determining the selection), and a *physical*

cycle (the act of making the supportive system aware of the objects). While the cognitive cycle must always occur, the physical cycle happens only if a computer-generated representation will support the reference[20]. Though specific to AR, when selecting physical artifacts for reference, the physical cycle often does not occur.

Figure 20 - The referential pipeline

We define *representation* as the means through which the attention of others is directed to a set of selected objects. Representation techniques are often visual, such as highlighting or alternate visualizations. Though discussed in more detail below, pointing is an example of a visual representation, just as the concomitant deictic speech that normally supports the reference is an auditory representation. This may at first seem counter-intuitive, but it can be argued that pointing is to make others aware of an object. One might suggest that this logic fails when discussing *gestural interfaces* – where the act of pointing informs the system of an object of reference; however, this is an alternative method in the physical cycle of the selection phase. Needless to say, the distinction between the selection and representation phase is subtle - especially when

[20] It will be interesting to see how brain interfaces might merge these two cycles into one. Thus, one could merely think of an object and have the system become aware of the selection.

addressing non-verbal referencing, such as changing ones pose, eye gaze, gesturing or using deictic speech. It is important, however, to clearly differentiate between the two.

The final phase of the pipeline is *acknowledgement*, which is the optional act of recognizing a reference and responding; this phase is heavily dependent on context. Of interest is how formally this occurs, as well as how it affects the behaviors of participants when this phase is absent. In some systems, an acknowledgement may be a gesture, utterance or physical action [6]. In other systems (e.g. distributed collaborative surgery), a *guaranteed* acknowledgement becomes increasingly important; ensuring that the reference was unambiguous can be of extreme benefit in mission-critical applications.

The Inter-referential Life Cycle

The pipeline described above describes how referencing occurs chronologically; however, it is necessary to incorporate other factors that influence it, including the available channels of communication, common ground between participants, relationships between artifacts and participants, as well as the properties of those objects; this creates the *inter-referential life cycle*. By intentionally avoiding domain-specific techniques, it can be more easily merged with existing ontology. For example, when applying this model to VR, the ontology developed by Bowman (on 3D selection techniques) can classify the selection techniques available for that domain. Figure 21 shows the integration of objects and participants as well as the relationships that exist among them. Further, it lists several of the spatial Object-Actor relationships and object states found within collaborative AR. Though more formally defined later, the figure is summarized here for clarity. The process begins with an initiator who has a set of

relationships with one or more shared objects. Using some selection technique, a set of objects of reference (0 or more) are chosen and represented to a set of reference receivers, each of whom have relationships with the objects. An acknowledgement may or may not be generated for the initiator by these receivers, though our studies show this can negatively affect referencing behavior. Note that the life cycle is independent of time, and therefore applies to asynchronous environments as well.

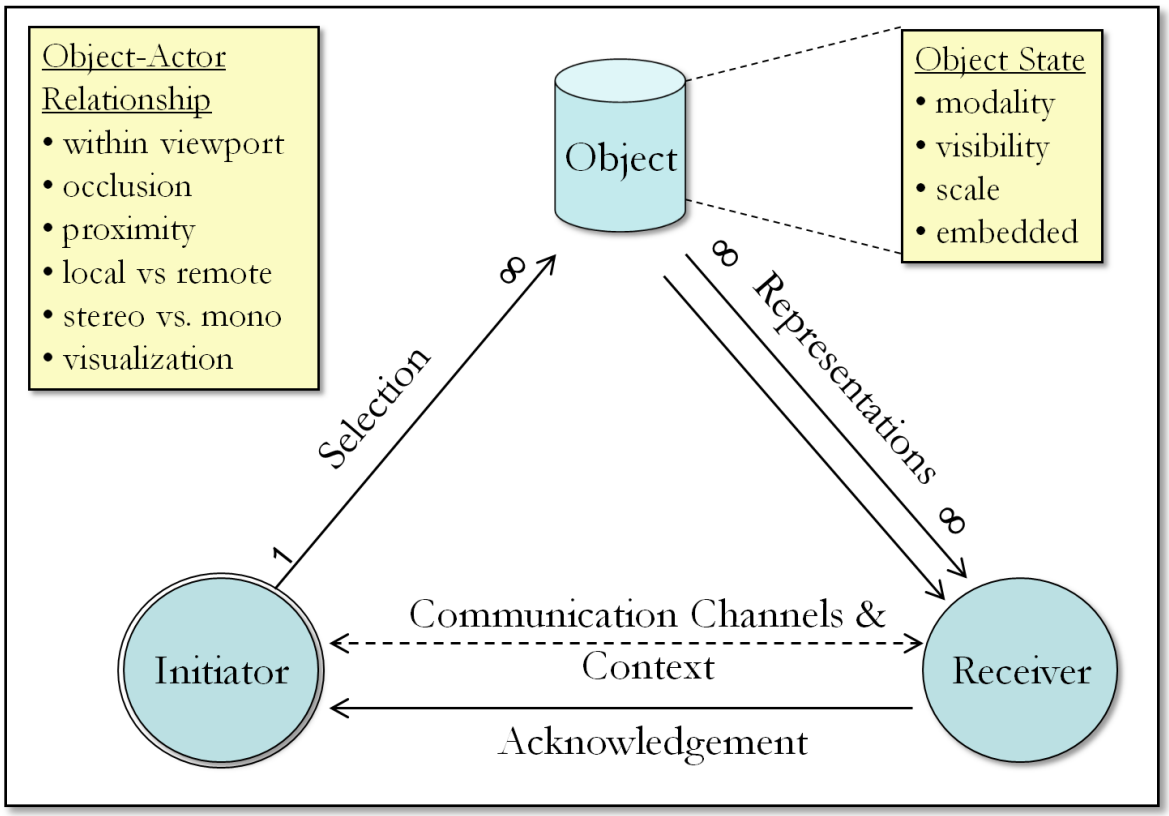

Figure 21 - The inter-referential life cycle (applied to AR)

The initiator and receivers share a *context*, which includes common ground, multiple channels of communication and the collaborative task. When applied to collaborative AR, these channels of communication may include spoken audio (or VoIP

for remote scenarios), shared video, object states (e.g. pose), or contextual information about other participants (e.g. current visualization). The figure above also lists domain-specific relationships that exist between participants and objects as well as the states those objects may be in; these states and relationships are listed in more detail in the next section.

4.2. Factors that Influence Referencing in Collaborative AR

Object-Actor Relationships

Referencing is affected by the environment that surrounds the users, the referencing techniques that are present, the skill of the participants in using them, and the relationships between the objects and the participants. Though the list is not exhaustive, we attempt to enumerate several spatial relationships that have an impact on referencing:

Occlusion: 2D artifacts can occlude other 2D objects in traditional applications, such as one application window in front of another. In 3D, however, as a user's position within the space changes, their relationships with objects change as well - potentially causing them to become occluded (including reference representations). Even though the object of reference is entirely unoccluded in the view of the initiator, it can be partially or completely occluded in the view of one or more receivers – creating asymmetry in the amount of information visible between participants. Occluded views are especially prevalent in scientific visualization, where objects can be tightly clustered (e.g. molecular modeling). The presence of occlusion should ultimately affect the representation technique used, and may be alleviated (or even eliminated) through a variety of techniques, such as sharing the viewpoint of the initiator, using transparency

shaders, or alternative techniques such as the "space distortion" method [13]. Further, though it is technologically straightforward for virtual objects to occlude physical ones, the opposite is not true; the pose and geometry of physical objects (including other participants) must be known for correct occlusion to occur. As shown in our studies, occlusive cues are beneficial in generating more-accurate references - especially when stereoscopy is not supported – as they are a strong indicator of relative depth; when they are not correctly supported, incorrect occlusive cues can be disconcerting to users.

Proximity: there has been a significant amount of research performed in selecting virtual objects at a distance, including image plane techniques, extended arms (such as the Go-Go technique) and WIM (World-In-Miniature) [61]. However, many of these techniques rely on intersection-based algorithms, and do not function well with physical objects unless their geometry and pose are known. In addition, physical objects that are at a distance cannot be "touched" (especially those in remote environments), and thus, it may be necessary to rely on computer-mediated techniques to reference them. Further, similar to VR selection techniques, we have found that the spatial relationship of the objects affects the accuracy of the interpretation of the reference, especially when simpler representations (e.g. a virtual arrow) are used.

Within View Frustum: many HMDs have a limited field of view[21], yet artifacts can surround the user. We must consider effective representation techniques that draw the attention of the receiver(s) when objects fall outside the view frustum, whether auditory (e.g. deictic references or ambient audio) or visual (e.g. Biocca and Tang's Attention Funnel).

---

[21] Typically between 24° and 55°

Local or Remote Environment: in purely virtual environments, local and remote objects are indistinguishable and most often rely on computer-mediated techniques for interaction. While many of the benefits of co-located collaboration are lost in remote AR environments, a virtual object can be easily shared in local or remote workspaces. However, remote physical objects present unique referencing challenges. The initiator must be aware the remote environment (including the collaborators and objects) as well as being equipped with referencing techniques that function across space.

Visualization: participants are assumed to have an independent view of the environment. Because views are customizable, participants can visualize virtual objects at different scales using a variety of visualizations (e.g. ball and stick and space-filling visualizations in molecular environments) which may cull out or present information in a context that makes a reference ambiguous.

In general, referencing becomes difficult when the relationship between the initiator and objects differs from the relationship between the receivers and the objects; these relational asymmetries are *compounding*, and present HCI and technological challenges. For example, if two participants are co-located and trying to refer to embedded virtual content, system designers must address how these artifacts are to be referenced; they may take advantage of the co-located environment and rely on non-verbal communication to support the referencing technique. However, when we extend this into a remote scenario, collaborators must be aware of each others' environment, their viewpoint, and have techniques that work across distance. Extending this further to include heavily occluded objects or those with little discernable visual information increases the complexity even more. Fortunately, much of this can be eliminated by

introducing a "What You See Is What I See" (WYSIWIS) interface - reducing the "awareness gap" between collaborators by allowing them to occupy the same physical space *virtually*; as our study suggests, this gap can be further reduced with the introduction of virtual reference points.

Object States

In the previous section, we discussed the relationships between the objects and participants, yet objects themselves can contain a variety of attributes that affect referencing – many of which must be distributed to other users to maintain awareness. These include:

Virtuality: Bowman et al. argue that the virtuality of an object will determine the modality of selection technique used [61]. While augmented reality simply augments an already complex environment with virtual artifacts, it may be necessary to track physical objects if they are included in the task; when the geometry of the object is known ahead of time, tracked physical objects can interact with virtual environments in natural ways (such as affording occlusion), allow for virtual representations of physical objects located in remote environments, and permit them to be referenced using virtual techniques. However, sharing the pose of physical objects with remote participants requires the information to be propagated across the network, giving rise to scalability issues.

Scale: because virtual objects have the ability to change scale, it can be cumbersome to refer to them. Scaling virtual objects is common for scientific visualization when virtual objects are loosely coupled (if at all) to physical objects - allowing humans to "fit" within the virtual workspace. For example, molecules must

obviously increase in scale while weather data (e.g. tornado wind vectors) must decrease in scale for meaningful interaction to occur. If people are collaborating at different scales, information may be visible to one party and not another. For example, the scale of the environment directly relates to the relative clustering of objects, and scaled-down data may lose visible features. Similarly, the scale of the physical workspace may restrict the number of participants, allowing few simultaneous references.

Embedded: one of the most impressive abilities of AR is in how it seamlessly integrates virtual objects within the physical environment; one variation of this is to embed virtual objects *within* physical ones, providing users with "X-ray vision" (e.g. AR surgery or visualizing the electrical system within a building). Embedded objects present a lower bound on the proximity between the objects and participants which will affect both selection and representation techniques. This scenario may require participants to refer to objects at arbitrary depths; for example, an architect walking through a building may refer to physical features within reach, well above their head, or in the case of plumbing and electrical systems, deep within floors or walls.

Visibility: though visibility pertains to virtual objects only, issues of privacy and ownership will inevitably evolve as collaborative environments contain a larger number of users; privacy techniques are being investigated by researchers [99].

Physical properties: references to physical attributes, such as weight, color and texture can help participants differentiate objects. When the physical properties of objects are too similar, there may be too little discernable information in the environment for deictic references to be effective. Thus, if a remote physical object is

represented by a local virtual one, physical attributes may need to be transmitted to help support deictic referencing.

4.3. Referencing as a Formal System

Without loss of generality, we make the following assumptions. First, all objects in the system have some method of being uniquely identified, either through a naming scheme, or a set of attributes (such as position). Without this property, it becomes difficult (if not impossible) for collaborators to differentiate between objects. Second, time is inherently part of any collaborative environment, and is used here to enforce the order of operations in the pipeline (represented as direction in the graph).

The graph is interconnected by both process and the relationships between the objects and participants. The environment contains a set of participants Z and an *initiator* $I \in Z$ who intends to generate a reference for one or more receivers $V \subset Z$. Note that V is not defined as $Z - I$, because a reference may not be intended for all recipients, and cannot be defined as $V \subseteq Z$ because it is implied that the initiator already understands the reference.

The set of all objects O is defined as a set of artifacts that exist in the shared space. In concordance with the definition of Rodden, objects might be people or other information [10]. These objects contain context- and domain-dependent properties that are separate from their relationship with the participants, such as *ownership*, *virtuality* or *locked* (held by another participant). The *objects of interest* O_i is a set of 0 or more objects that are intended to be selected by I through the cognitive cycle; thus, $O_i \subseteq O$ and $|O_i| \geq 0$. We further define O_a as the *actual* object set which contains the resultant set

of objects produced by a selection. For clarification, we use standard mathematical definitions to state that two sets are equal if they contain the same elements, regardless of ordering (and thus are equivalent as well).

A set of relationships R_1 exists between I and O. Each relationship contains domain-specific properties, such as the spatial relationships found in 3D environments. A secondary set of relationships R_2 exists between O and V, consisting of similar properties. Thus, $|R_1| = |O|$ and $|R_2| = |O| \times |V|$. These relationships will ultimately affect the selection technique that is chosen, and allows for independent representations for each relationship of R_2.

The set of relationships that exists between the initiator and the other participants is generically called a *context*, and includes the channels of communication and common ground [14]. This encompasses shared history, *information* and *social symmetry*, as well as the amount of *functional symmetry* in the channels of communication. Examples of these channels include auditory (e.g. speech), visual (e.g. shared video) and object state (e.g. cursor position or document data). The role of context is important in generating meaningful references; participants who view the environment from similar viewpoints and share similar experience will have less difficulty communicating than those where cognitive and communication asymmetries exist [6, 12, 14]. We agree with the claim by Billinghurst et al. that no collaborative application can be perfectly symmetrical, given the varied knowledge and experiences of the participants. To this extent, though it is indeed plausible to devise a metric for context (as a factor of participant and implementation symmetry), we believe this to be beyond the scope of this work.

Finally, we must also define the concept of a *cursor*, which is a location-based identifier within the shared space (e.g. a *telepointer* in Gutwin's work) which can exist in one, two and three dimensions; in co-located AR, the cursor may even be the participant's hand.

Formal Definition of Selection

We can now formally define *selection* as the specification of a resultant set of target objects using a set of rules for inclusion into (or exclusion from) set O_a; that is, selection is a set of functions reducing the universal set of objects $O \rightarrow O_a$. We say that a selection technique is *accurate* if $O \rightarrow O_a = O_i$. The selection technique chosen is highly dependent on context and object type and the relationship between the initiator and O_i (a.k.a R_1). For example, a set of rules may specify all objects that contain a specific property (such as *.txt files) or objects whose position fall into a given range (a.k.a - rubberbanding).

Hierarchically, the concept of selection can be further decomposed into *mode* (the medium through which the selection is made) to obtain more specific domains of selection. For example, audio interfaces are becoming more commonplace in mainstream applications. Examples of one-dimensional selection methods include queries, textual selection or "tabbing" through an interface. Second-dimension techniques include rubber-banding (or bounding box), knife tools, or pixel selection. Third-dimension examples include concepts of gesturing, raycasting, image planes and scaling (e.g. WIM). The scope of this hierarchy is purposely restricted to abstract concepts, allowing us to leverage off of existing ontology. For example, Bowman et al.

define a hierarchy for 3D selection techniques [15]. The selection hierarchy is extendable for future methods, including biometric techniques.

Formal Definition of Representation

As defined previously, *representation* is a means through which the attention of others is drawn to objects of interest (O_i). Representation techniques are designed to be easily perceived and are thus inextricably linked to human interpretation. Because of this, representation techniques are open to subjectivity, making it difficult to define formally, and are a primary source of referential ambiguity. However, it can be argued that it is the *intent* of a representation technique to draw attention to O_i. Any given representation *P infers* a set $O_p \subseteq O$ of 0 or more objects; that is, it is the attempt of the representation to draw the attention of the receiver to O_p. More specifically, the intent behind a representation is that $O_i = O_p$.

Most often, representation techniques are *visual* - such as changing the appearance of an object through highlighting or by gesturing with a cursor - or *auditory*, such as deictic speech. However, gesturing and deictic speech should not be confused with *selection*; through these actions, the intent is to draw the attention of others - the definition of representation. The object(s) of reference could also emit audio – the effectiveness of which can be demonstrated in our ability to locate a ringing telephone. Finally, though less common, a representation can be tactile. Examples include force-feedback mechanisms like haptic devices or hand-held game controllers that vibrate.

When designing interfaces to support inter-referential awareness, one must be aware that the perception of the representation varies with the relationships of R_2. Visual references may *not* be within view, and auditory references may be too far away to be

heard. Therefore, a representation should not only consist of some way of drawing the attention of the receiver(s), but should also *guide* them if the reference is beyond their perception.

It should also be made clear that the representation of an object should not be confused with the *feedback* received during the selection; feedback is intended as a confirmation of the selection *for the initiator*. In many situations, this feedback can be used as a representation for the receivers. For example, when highlighting text with a mouse, the background color of the text is often altered to confirm the selected text. In collaborative scenarios, this highlighting can also serves as a visual representation when referencing. However, it is the responsibility of the system to forward this representation in a meaningful way when receivers are in different computing contexts.

In distributed applications, it is evident that for the *system* to generate a representation, it must be aware of the selection. How then is gesturing, movement or deictic speech interpreted in this framework? They serve as a form of representation, with the argument that one does these actions in order to communicate with others. For example, in the work of Ott et al., participants were represented by a cone (thus, devoid of orientation information) and used proximity to refer to objects [13]; their change in position was a representation used to make others aware of their selection. Gestural interfaces admittedly blur the differentiation between selection and representation, and depend on the *role* of the system; if the system is to serve as a mediator between users, the gesture can be seen as a function to reduce $O \rightarrow O_a$, and is considered a selection technique. Finally, with a variety of communication channels available, it should be noted that multiple representations can be present as well - comprised of non-verbal

cues, deictic speech, movement and highlighting; deictic speech alone is meaningless unless it supports some other form of representation. This concept is denoted by the multiple arrows in the representation phase of Figure 21. It could be easily argued that redundancy in representations decreases ambiguity.

An overarching goal for an inter-referential system is to eliminate *referential ambiguity*. We argue that ambiguity occurs from poor selection techniques (during the physical cycle) or weak representations. When poor selection techniques are used, O_a does not equal O_i. Using weak representation techniques, O_p does not equal O_i or O_a. Thus, it can be stated that no referential ambiguity exists when $O_i = O_a = O_p$.

Acknowledgment

The final phase of the pipeline is *acknowledgement*, and is used as confirmation for the initiator that a reference is understood by *V*; depending on the formality of the reference, an acknowledgement is optional. However, our studies have shown that when the acknowledgement is not given, this negatively affects the efficiency of the participants in building tasks; the initiator becomes unsure of the state of collaboration, and often pauses. The receiver may become frustrated from waiting for further instruction. In the case of 3D references, we observed the initiator continuously referring to the same object – blocking the view of the builder – while the builder was waiting for the reference to clear the workspace to continue. Thus, acknowledgement can be seen as an important *social protocol*, and is dependent on the context of the environment. Acknowledgement can take on several forms, including speech, gesture (head-nodding) or action (such as acquiring the correct piece of a model). However, in formal systems where a *guaranteed* reference is required, the transpose of the graph can be taken such

that the initiator becomes the receiver (and vice versa) – requiring a complete re-selection of O_p. We observed this formality between dyads in our studies as a form of clarification by the receivers. Though this may seem unnecessary, it becomes increasingly useful in life-dependent scenarios such as collaborative tele-surgery.

Definition of an Inter-referential System

Given the definitions above, it is possible to define a *referential system* $\Re = (I, S, R_1, O, P, R_2, V, C, A)$ where:

- $I \in Z$ is the initiator of the reference
- S represents the selection technique, mapping $O \rightarrow O_a$
- R_1 is the set of relationships that exist between the initiator and O
- O_a is the set of actual objects selected
- P is the representation technique(s) that infers O_p
- R_2 is the cross of relationships that exist between O and V
- $V \subset Z$ is the set of receivers of the reference
- C is the context between I and each member of V, and
- A is the optional acknowledgement technique

4.4. Environmental Taxonomy

As shown in our user studies, inter-referential awareness is one of many forms of awareness that are affected by implementation issues. Here, we identify additional *environmental* parameters that affect this form of awareness by presenting taxonomy and include discussion of implementation requirements. It is our hope that this information will help developers understand how various configurations would affect referencing. Below, we describe the common factors that are found in collaborative applications and how they may be supported.

Time/Space – According to Bafoutsou et al., collaborative environments are often classified across *time* (synchronous vs. asynchronous) and *space* (co-located or remote) [40]. In AR, referencing is heavily influenced by spatial properties as well as spatial forms of communication, such as gaze, gesturing and orientation. We can extend the taxonomy to include the component of *dimension* which impacts how references will be made (see Figure 22). To better visualize this, the traditional 2x2 classification matrix can be rotated and extruded to include depth. We view this as an important distinction - as it cleanly addresses the object states and relationships described in Section 4.2, and accommodates spatial referencing. It also implies that participants have 3D viewpoints, and covers a wide range of scenarios such as references to remote physical objects.

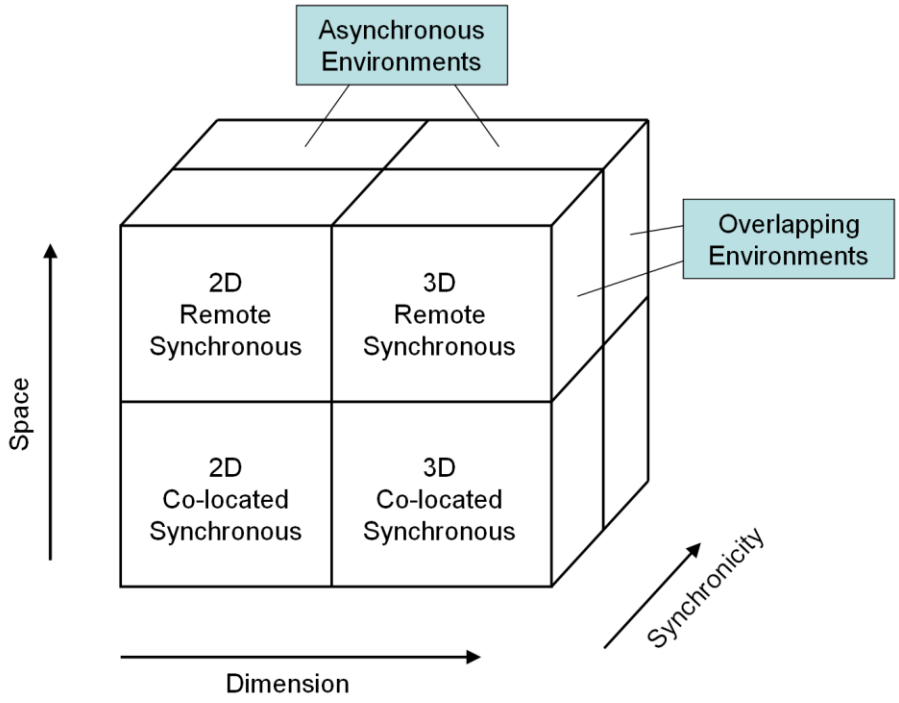

Figure 22 - Taxonomy of referencing spaces

However, including this extra dimension complicates system development and demands more from the underlying system. Remote collaboration in AR presents

challenges not found in other environments. For example, because space is comprised of both the physical and virtual objects that surround the user, the addition of each new remote participant brings with it the environment that surrounds *them*, creating a composite of *overlapping* environments; in other words, a coordinate system can be comprised of multiple parallel spaces. While distributed scene graphs exist, a more flexible abstraction is to view them as *aggregate* scene graphs. Such scenarios give rise to several technical questions, such as how parallel space can be visualized in a meaningful way, whether or not exact configurations of distributed space are necessary, and how participants can become aware of remote objects and generate meaningful references to them. Further, we need to develop efficient methods of distributing complex environments – especially with the initial burst of geometric data when joining these spaces.

Communication Channels: Though addressed in previous sections, communication channels are a critical component in supporting awareness – especially when referencing in synchronous environments. However, there is a direct correlation between the amount of available network bandwidth and the awareness that can be supported; slower networks afford less awareness. Several communication channels are needed for collaborative AR, drastically increasing network requirements. Object channels maintain the state of the shared artifacts within the environment, including pose and geometry. In remote scenarios, the object channel can also convey information about the participants, such as the location of their head and hands, and any 3D virtual cursors they use to make references. Video channels allow for participants to share viewpoints – helping them to become aware of the remote environment and establish or

clarify reference points. To support spatial tasks, this channel may be stereoscopic, requiring additional bandwidth. While co-located participants can freely speak with one another, distributed scenarios also require a separate audio channel to provide basic communication - allowing participants to make deictic references.

Task: referencing is also affected by the task, which can further dictate real-time constraints or alternate referencing techniques (e.g. hands-free referencing during collaborative surgery). The task will ultimately affect the kinds of referencing techniques that are supported, such as the spatial techniques in Chapter 3. In hard real-time scenarios, latency in the network can negatively affect communication channels – creating unnatural collaboration and causing deadlines not to be met. It may be necessary to "guarantee" that a reference is current through the use of a guaranteed protocol (e.g. TCP vs. UDP) at the cost of sending acknowledgement packets. Additionally, it is often valuable to see the reference being made in real-time – as it draws the attention of the receivers and puts the reference into context.

Further, it is necessary to classify references as *temporal*. If the network becomes unavailable in real-time scenarios and references are still visible, they become stale and may lead to ambiguity; thus, references should have a limited lifetime (i.e. a "time to live" after the discovery of a network or other failure). Stale references can occlude the view of the receiver (as discovered in our pilot study), and as our studies suggest, can be alleviated by allowing users to toggle the visibility of the reference. Further, because servers maintain state for asynchronous environments, collaboration can span days or weeks; references (such as annotations) for late-joining collaborators can become outdated if changes are made to the environment, and should be removable.

Workspace Placement: Kiyokawa et al. have shown that in AR, the placement of the workspace affects communication behaviors [78]. Placing the workspace between co-located collaborators allows them to see one another's gestures (as much as 95%), yet presents them with mirrored (asymmetric) views. Placing the workspace on the wall (e.g. a whiteboard) provides collaborators with similar views, but reduces these cues by placing a majority of them outside of the field of view.

Group Issues: Groups are comprised of users with varying backgrounds of experience and skill. As seen in our studies, some participants could accurately refer to objects, even when a stereoscopic view was not provided; others could be classified as having an extreme misunderstanding of the environment, a majority of which was ameliorated through training and the inclusion of shadows[22]. Because current AR technology is expensive, little research has examined collaboration for groups of more than a few users; as technology becomes more affordable, the possibility of large groups becomes feasible. However, these groups present several referencing challenges. More collaborators in an environment introduce the possibility of it becoming cramped, especially when the workspace is limited; in groups where there are many remote participants, virtual avatars (or pointers) must be identified with a user. Further, network requirements are similar to a combination of virtual environments (i.e. the propagation of the environmental scene graph) and traditional CSCW (by providing shared audio and video), and may suffer from similar scalability issues as the number of participants and objects increase. Finally, groups perform in a variety of ways, including their

[22] Much to our frustration, some participants *never* gave accurate references, which seemed to be directly attributed to how actively they play video games. This is of interest because "first-person shooters" do not provide a stereoscopic view, yet gamers appear to interpret depth correctly through other cues.

interaction pattern (such as one-to-one, one-to-many) and interactivity level; simultaneous video streams may raise awareness of multiple collaborators, but obviously requires more bandwidth.

Implementation Issues: throughout the studies presented in this dissertation, the systems have relied on visual tracking of fiducials to determine the pose of the user's head and hands as well as artifacts in the workspace. While this is a more cost-effective approach than other technologies (such as the IS-900 tracker), users can be reluctant to adopt referencing techniques when sporadic tracking occurs; instead they may rely on other forms of referencing (e.g. chaining or relative referencing). Depending on the capabilities of the system, objects may also be presented via a bioscopic or stereoscopic view. While our studies have shown that the inclusion of shadows positively affects referencing behavior, HMDs that support stereoscopy are likely to increase the accuracy of referencing.

Finally, for completeness, we can classify references in AR environments by their level of computer mediation and whether they are deictic or gestural (see Figure 23). While intuitively simplistic, this generalized model captures referencing techniques from multiple domains, such as highlighting and bounding. While highlighting is reserved for digital content (such as using a mouse to select a set of text), bounding can be both computer- and non-computer-mediated; for example, a pointing device can be used to outline a set of objects in computer-mediated manner, just as hands can be used to bound space or a set of real world objects; these reference techniques are a form of gesturing.

	Deictic	Gestural
Computer-mediated	Text/IM VoIP	Virtual Cursors Annotations
Non-computer-mediated	Speech	Pointing Gaze

Figure 23 - Classification of references in AR

4.5. Summary

In this chapter, we have constructed a generic, theoretical foundation for inter-referential awareness in CSCW, which includes a framework and taxonomy. The goal of this work was to help designers analyze their domain, as well as to consider factors that may influence referential techniques. We define this form of awareness as a process, consisting of the phases of selection, representation and acknowledgement, as well as a set of relationships between the shared artifacts and the participants. Further, we have provided a formal definition for an inter-referential awareness system, which allows us to describe referential ambiguity. We have applied this framework to collaborative AR, enumerating domain-specific components that affect references in these environments. Finally, we have provided taxonomy for classifying the environmental elements that impact referencing, as well as a generic classification of references. These concepts have been subsequently refined based on the user studies we have conducted.

This theoretical foundation organized the complexity of referencing in collaborative AR, as well as enumerated the factors that designers could potentially encounter; while this foundation incorporates a majority of referentially-related entities, realistically not all of these are simultaneously present in many scenarios; consequently,

the design space needs to be culled. Further, it is important to gain insight on how users behave within these environments and techniques that they prefer. Based on our framework, we conducted a pilot study with these goals in mind, focusing on scenarios that are most commonly found in the literature. The results of this study can be found in Chapter 5.

CHAPTER 5

UNDERSTANDING THE DESIGN SPACE OF REFERENCING IN COLLABORATIVE AR

Chapter 3 exposed some of the referencing difficulties we experienced in virtual environments. In transitioning the system into AR, we realized that the complexity of referencing is *compounding* across scenarios, and that we were in need of a framework for organizing concepts, relationships and domain-specific properties. Chapter 4 enumerated many of these factors for collaborative AR, and provided a framework and taxonomy to organize them. However, it is rare that *all* referential possibilities would be simultaneously present in one system; designers would be overwhelmed when trying to incorporate support for such scenarios. Further, a critical element was missing from our research: how subjects collaborate within AR environments. By using the framework as a foundation and focusing on the most common scenarios from the literature, the exploratory study presented in this chapter attempts to cull the design space of referencing in AR environments and gain a better understanding of the kinds of support participants prefer during collaboration.

To determine which referencing techniques are effective across different contexts, we extended the work in section 3.3 by developing an augmented reality prototype that supports specific interactive tasks required during co-located and remote collaboration; these tasks parallel those found in molecular modeling environments. We designed the prototype to support natural hand pointing, cues to designate selected portions of the model, and multi-modal interactions. We conducted this study to better understand how participants generate references to virtual and physical content and hypothesized that 1)

participants would prefer co-located interaction, 2) awareness of the viewpoint of others would be significant, and 3) shared video is effective when disambiguating references [125]. Participants were paired in groups and asked to collaboratively build physical and virtual models. Our scenarios are a cross between co-located and distributed collaboration, working with physical and virtual models, and using augmented techniques. We provided the users with a basic set of virtual referencing methods and encouraged natural interaction in order to observe which ones they use under certain conditions. This study suggests two design guidelines for collaborative AR systems:

1) Multi-modal referencing techniques should be provided
2) Shared viewpoints are a desirable medium of communication when generating references

While most CSCW research focuses on distributed scenarios, collaborative AR allows for participants to be co-located; we were subsequently interested in how participants would use physical gesturing as representations, such as pointing and gaze direction. Unlike WYSIWIS interfaces, each participant is required to have his or her own independent viewpoint - creating a disparity in the visual information presented to users and a potential for referential ambiguity; we believed that shared viewpoints would be effective in alleviating this asymmetry. We further believed that distributed scenarios present more referencing challenges co-located ones and wanted to explore methods for generating references to content in remote spaces; in these scenarios, the initiator must first become aware of the shared space, paralleling the expert/technician scenario described previously.

In this chapter, we describe how the study was designed to explore a variety of collaborative scenarios and techniques in both physical and virtual space. We also present our observations of user behavior, participant feedback and an analysis of the results.

5.1. Study Description

The goal of our study was to better understand the kinds of collaboration – specifically referential - that occur in AR environments while working with 3D molecular models. We used magnetic building toys composed of primitive geometric shapes to simulate the models (see Figure 24). These physical models incorporated magnets that mirrored the physical bonding of molecular structures; thus, the virtual and physical models behaved approximately to the same rule set.

Participants were grouped into pairs and asked to collaboratively build physical and virtual models in a variety of scenarios. In each scenario, participants were required to view the physical world through their HMDs. Throughout each exercise, one participant (known as the *guide*) could see a physical model in its target configuration (or *target model*) and was free to pick up or rotate this model to better understand its structure. Each model was comprised of a maximum of 7 spheres (6.3 on average) and 12 connectors (9.5 on average); correct coloring of the connectors was required. The second participant, or *builder*, had no prior knowledge of the model's structure, and could not view the target model. During the process of construction, the guide was allowed to make gestures and touch the shared (working) model, but could not manipulate it (i.e. the guide was not allowed to help build the working model). Participants were allowed to talk with one another at all times, including the remote

scenarios, where they were separated by a small barrier. Additionally, participants were free to move about the space, as well as rotate the workspace to help accommodate their actions. All scenarios within the study were timed, though limited to a maximum of 10 minutes to complete the model. Participants were interviewed after each scenario about the suitability of the interaction techniques and their ability to adapt to the technology.

Figure 24 - The Target Models

To serve as a base case, we "pre-piloted" this study using 4 students[23] in a remote, audio-only environment; the pairs were unable to see one another but were located in the same room, and could therefore use instruction only (or very weak deictic references). The builder in each pair constructed 3 models, similar to those shown above. However, it was such a horribly frustrating experience for the users that we decided to remove this configuration in the interest of time (no pair *successfully* completed a model after nearly 25 minutes). Further, because the intent was to observe behaviors in *AR environments*, it was of little consequence that this scenario was removed. The same 4 students were

[23] These were current students of the researchers, and were therefore ineligible for the study

used to refine the actual study, and played an important role in the development of virtual interaction techniques and discovery of limitations in the system.

A total of 8 groups participated, comprised of 12 males, 4 females, whose ages ranged from 19 to 38. Participants came from the general student body, though 13 of them were I.T. or computer science students; the others came from the fields of biology, psychology, and mathematics. Students of the researchers were not allowed to participate in the study. None of the participants indicated prior experience in virtual or augmented reality, though 13 had moderate experience with video games (where "*moderate*" is defined as more than 2 hours per week). Two of the pairs had prior experience working together as a team or similar communication experience. The overall experiment took 1.5 hours, and began with a 5 minute video to help explain how to construct virtual models; an additional 5 minutes of training time was given before the study began to train each participant individually in their role. Video feeds were duplicated to allow researchers to view the collaboration.

The design of the study was based on the framework presented in section 4.1 and configured in six ways. In all scenarios, users viewed the world through a HMD. The following scenarios were intended to represent a wide variety of applications:

1. Co-located/physical: participants were asked to build physical models using magnetic children's toys. This experiment (called "*the icebreaker*") allowed participants to become accustomed to a non-stereoscopic view of the world as well as become familiar with how to give or receive instructions from their collaborator. In this configuration, participants did not wear a data glove, and no virtual content was present.

2. Remote/physical: the builder moved into a location where they could not be physically seen by the guide, and the view of the guide was replaced with the video feed of the builder – giving the guide the exact view of the builder. To see the target model, the guide could look down (or use peripheral vision), or flip up their visor if necessary (though this did not occur). The builder was provided with his own view, and participants did not wear the data glove.

3. Co-located/virtual: participants shared the same physical and virtual workspace to build a virtual model. Participants were equipped with a tracked data glove to interact with the virtual models, and had independent viewpoints.

4. Remote/virtual: participants were provided a physically separate workspace, but a shared virtual one to build a virtual model. Each participant had an independent viewpoint and was equipped with a tracked data glove.

5. Co-located/augmented: participants shared the same physical and virtual workspace to build a physical model. Participants had independent viewpoints, could use both physical and virtual referencing techniques, and were equipped with a data glove.

6. Remote/augmented: participants were provided a physically separate workspace, but a shared virtual workspace to build a physical model. The view of the guide was replaced with the view of the builder (as in the remote/physical scenario). However, the hand of the guide was *locally tracked*, and mapped to a virtual pointer in the builder's remote physical space (see Figure 25); this allows the guide to make 3D virtual references to remote, physical objects (e.g. using their virtual arrow).

A summary of these configurations can be found in Table 1. A balanced Latin square design was used to order the virtual and augmented scenarios between groups to balance for learning effects in the technology and activity.

Table 1 - Summary of study configurations

	Physical models	**Virtual models**	**Augmented models**
Co-located	Scenario 1 *Independent views* Physical referencing, only	Scenario 3 *Independent views* Physical & Virtual referencing	Scenario 5 *Independent views* Physical & Virtual referencing
Remote	Scenario 2 *Shared view* Physical referencing, only	Scenario 4 *Independent views* Physical & Virtual referencing	Scenario 6 *Shared view* Physical & Virtual referencing

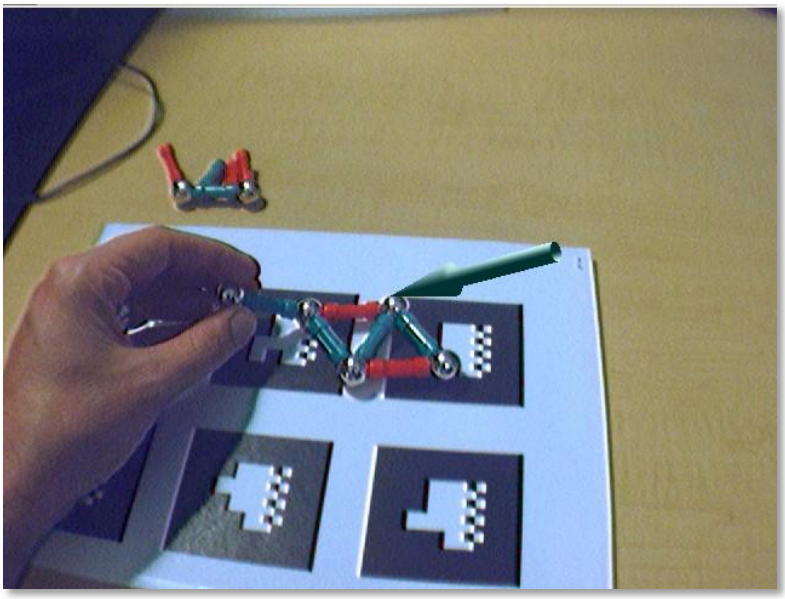

Figure 25- A 3D reference to a physical object from a remote participant

5.2. User Interface

In each scenario, participants were equipped with a HMD through which they viewed the physical world. While working in the virtual and augmented scenarios, the users wore a data glove which provided them a means to interact and reference using virtual techniques. In these scenarios, both participants had a virtual arrow attached to their glove which allowed either collaborator to select existing virtual spheres. Further, the head of each participant was tracked, which allowed a simple virtual head to be present (see Figure 26); if this became distracting, the visibility of the virtual head could be independently toggled for each user by the researchers. The virtual shared space was comprised of a color menu, as well as a workspace (surrounded by a wireframe box) where the virtual models were required to be built. In addition, each participant was equipped with a one-handed chorded keyboard - allowing the users more mobility around the workspace as well as the ability to touch type. Each participant used a maximum of three non-chorded keys on the keyboard.

To create new spheres, the builder would intersect the tip of his arrow into a special sphere (shown in Figure 26 as the light grey sphere to the left of the workspace) and then move his hand into the workspace. During this interaction, a new sphere was attached to the tip of their arrow, and the arrow turned red to denote that it was "sticky" (i.e. it was possible to translate the sphere). Toggling from sticky to non-sticky mode was done by pressing the '*a*' key (for "arrow"), which also allowed for *clutching*[24] of the virtual objects. To create a bond, the builder selected exactly two spheres and then

[24] A selected object can be repeatedly translated and released to "push" it to arbitrary distances.

pressed the '*b*' key (for "bond"). The builder could also select/de-select all spheres by pressing the '*d*' key. For example, to move the entire model, the builder could press '*d*' (selecting all spheres), press '*a*' (to translate) and then move his hand.

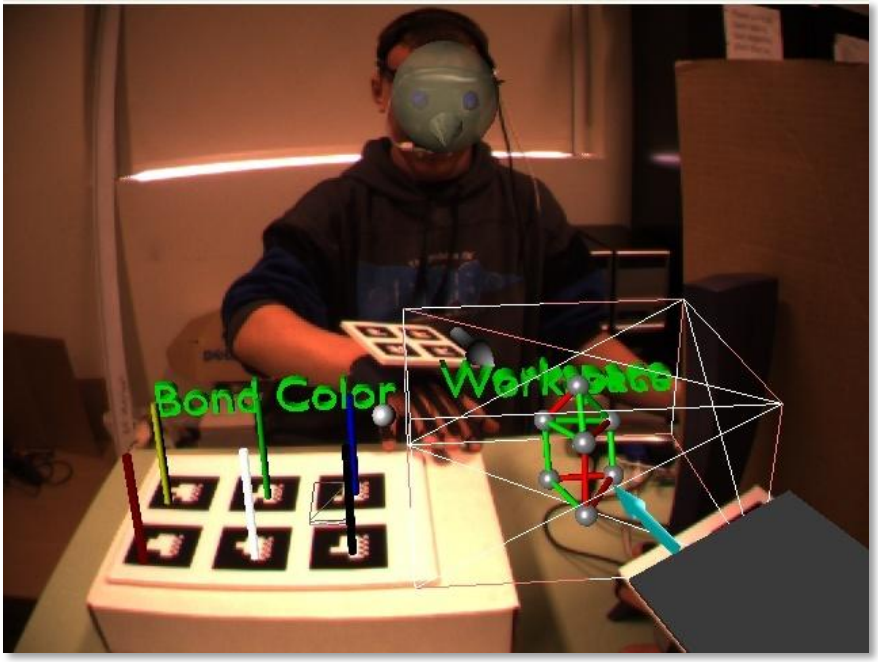

Figure 26 - Co-located collaboration from the builder's view

Besides pointing with their virtual arrow, the guide could also reference larger space by generating a dynamic 3-dimensional bounding box by pinching and dragging in 3D space (from Chapter 3). This action would cause all atoms that fell within the box to be selected, and those outside to be de-selected. The guide was also presented with the option of changing the representation of selected spheres, allowing them to *pulse* (through changing their transparency by pressing the '*e*' key), *color cycle* (by changing the colors over time by pressing the '*f*' key) or *shake* (from moving slighting back and forth by pressing the '*g*' key).

5.3. Implementation

The system is a modification of a collaborative molecular modeling environment governed by a molecular mechanics simulator. It was developed in DART, and uses the ARToolkit for tracking [25, 121]. Each participant was equipped with a modified eMagine HMD, which provided a bioscopic[25] 40° field of view with a resolution of 800x600 per eye. The builder's HMD contained a single PointGrey camera, which allowed for 30 frames per second with a resolution of 640x480. The HMD of the guide contained a DCAM camera, which provided a resolution of 640x480 at 15 frames per second. System state was maintained through VRPN shared memory.

In scenarios where virtual objects or virtual referencing techniques were present, participants were equipped with a P5 data glove. A customized VRPN server was written which allowed for tracking of the bend state of the fingers; pose estimations from the infrared tracker were ignored and replaced by poses from the ARToolkit. In our pilot configuration, each glove was attached with a small three-sided cube, with one marker per face; however, tracking was unstable - making interaction difficult. Instead, we opted for a single plane of four markers, approximately the size of the user's hand. While this configuration restricted the hand orientations that were possible, participants in the pilot study felt that the plane occluded less of their view, and because of the superior tracking, made the system more usable.

[25] a single camera feed replicated to both eyes

5.4. Observations

Though each group had unique methods for constructing the models, there were common threads in how they interacted in each scenario and the feedback they provided. In particular, we were interested in 1) the presence of common referencing techniques during interaction, 2) improvised techniques that emerged when the technology or scenario did not adequately provide support for referencing, 3) a "wish list" of features that would have helped them during collaboration, and 4) general problems they had while interacting within the environment. The referencing behaviors are summarized in Table 2.

Table 2 - Referencing behaviors (rows) by group (columns)

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**
Used shape	X	X	X	X	X	X		
Deictic chaining	X	X	X		X			X
Changed perspective		X		X	X		X	X
Color reference	X	X	X		X	X	X	X
Body references	X	X	X		X	X	X	X
References relative to model		X	X	X	X	X	X	X
Changed representations				X				
Arrow occluded		X					X	X
Liked shared view	X	X	X	X	X	X	X	X

In all groups and in all scenarios, the guide did a majority of the talking. Both guide and builder made heavy use of deictic speech (e.g. "*this*", "*that*" and "*here*"), indicating that spoken references are important. The builder's responses were most often

used short utterances to acknowledge a reference (e.g. "*Yeah, OK*" or "*mmm hmm*"), or to clarify a reference by asking a question (e.g. "*Did you mean this one*?") Further, we observed that the builder would often acknowledge a reference by selecting the sphere the guide was pointing towards, if the sphere had not already been selected.

Participants could be seen trying to establish common ground, describing the overall form of the model or asking questions such as "*Start with a pentagon. You* do *know what a pentagon is, right?*" When working with the physical models, participants began by collecting resources; one guide said "*you are going to need 7 yellow connectors and 6 atoms*". When building virtual models, a guide from one group used their virtual pointer with supportive speech to indicate both location and number of spheres, saying "*You're going to need balls here, here, here and here.*"

Over half of all groups made use of *referential chaining* – using the last place that was referenced to generate a new reference point. For example, one guide said "... *the connector where you most recently attached.*" When working with virtual models, guides initially used their virtual pointer to refer to the position of the first sphere. Once a sphere (or set of spheres) was established within the workspace, guides often made spoken references relative to the sphere most recently created. This chaining of references may be an artifact of the sequential nature of the task; however, it was most pronounced in the remote scenarios. As the model became more established, guides began working at a higher level of abstraction - using parts of the model to establish a relative reference (such as "*next to the red triangle*"). Further, seven of the eight groups used the color of the connectors as a reference point into the workspace. After only a brief time, several groups began to extend common ground by creating their own

vocabulary. For example, if a sphere contained only green connectors, it would be called the "*full green atom*" or the "*all green ball*". This behavior was noted across all scenarios.

In four of the eight groups, the guides asked at one point if the builders could see their virtual arrows. In two of these situations, the guides were observed making *projected references*, where their virtual arrow was located *between* their viewpoint and the object to which they were referring (see Figure 27). To the guides, the projected reference appeared to be correct, yet to the builder, the reference was ambiguous. During selection and reference generation, other participants were seen moving their pointer closer and further from themselves – leveraging from occlusion cues to more accurately determine depth.

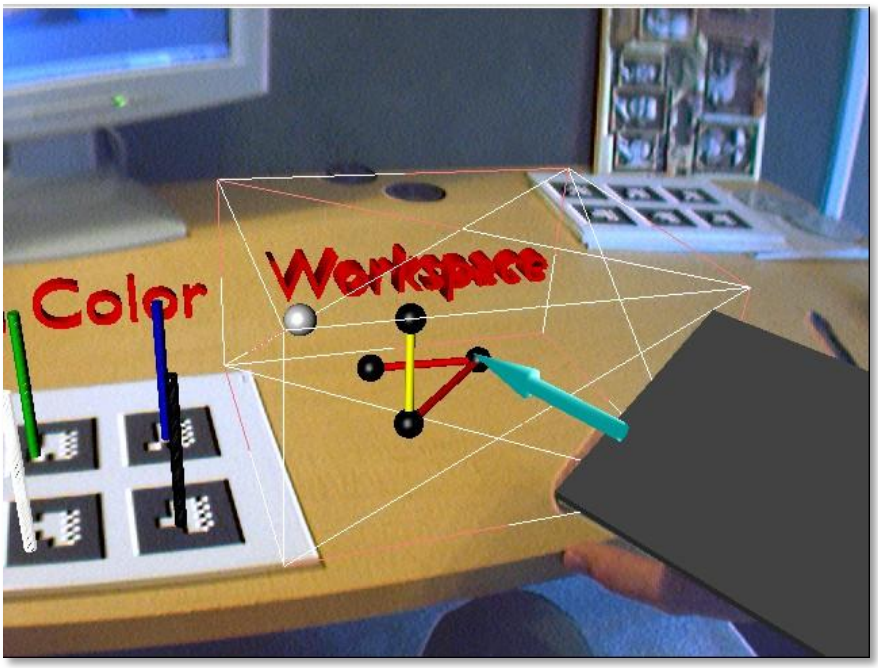

Figure 27 - an incorrect, projected reference (by the guide)

In the co-located/augmented scenario, a majority of the participants preferred to make references using their hands, though three of the guides *did* use their virtual pointers. In our last group, the guide felt comfortable describing the structure, with hands folded (see Table 3). Additionally, over half of the participants in co-located scenarios where virtuality was present needed to change their perspective, either by rotating the workspace or physically changing their location. Likewise, the builders explicitly moved the model to provide appropriate views to the guide, moved obstructing items, and verbally questioned ambiguous instructions.

While working in the remote/virtual and remote augmented scenarios, the virtual pointer was heavily used. After the reference was made, however, guides would occasionally leave their arrow in the same position for an extended period of time, occluding the view of the builders. Three of the builders made at least one comment that the virtual arrow was in the way. For example, one was heard saying "*Could you move your arrow?*" and another "*Your little arrow is in the way*."

As expected, the bounding box and alternate representations for selected spheres were only used occasionally. After the novelty of being able to change the representation, a majority of the guides left this option on "color cycling". Only one guide was noted as using the bounding box and representations to help make references – most often in a limited manner.

In scenarios where the guide's view was replaced with the view of the builder, participants made references relative to the builder's body (e.g. "*to your left*"). It was also in this scenario that the builders would often confirm the correctness of the model or acknowledge references by holding the model close to their face (or draw in near to

the model) and ask "*Like this?*" This behavior demonstrates that the builder was aware that the remote guide was "present" and could see their view.

Table 3 - Summary of techniques used by either participant

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**
1	H	H	H	H	H	H	H	-
2	H	-	H	-	-	-	-	-
3	-	HB	A	-	-	A	A	A
4	A	A	A	A	A	A	A	A
5	A/H	H	-	A/H	A/H	-	A	A
6	A	A	A	A	A	A	A	N/A

H = Hand, A = Arrow, B = Bounding Box

5.5. Participant Feedback

Because our HMDs were equipped with a single camera (i.e. non-stereoscopic), all groups noted the difficulty when both selecting and making references from lack of depth cues. To compensate, many said they used occlusion cues to determine depth, or incrementally changed their body position to obtain multiple views of the workspace. Some participants said that rotating the workspace was easier than moving their head (or physically changing their position), and thus preferred having an independent workspace.

Participants also expressed a desire for virtual reference points in the workspace when working with virtual models. They felt that this would allow them to more easily

disambiguate references, as well as overcome the lack of stereoscopy. Two groups explicitly suggested having a virtual grid that could be toggled between visible and invisible, while others described the need for similar virtual reference points.

Participants felt cramped in the co-located environment. First, the amount of available workspace was limited, and participants were required to keep their own fiducials (markers) within view while not occluding those of their collaborator. Users complained that they "*kept getting in each other's way*" (both virtually and physically) when trying to interact with the models, and felt that because of this tightness, needed the ability to cull out unnecessary virtual objects (such as the "workspace" text); while appropriate for novices, as users learned the system, the text was no longer needed and became an obstacle, particularly in the co-located scenarios.

All guides and a majority of the builders felt that sharing video was helpful when making references, especially in the remote/augmented scenario (using the virtual arrow of the guide). Many guides claimed that this scenario allowed them to generate better references because they could see from the builder's point of view. A majority of the participants preferred the remote/augmented scenario to the remote/physical, stating that the ability to point with a virtual arrow was helpful; one of the builders commented that "*it was good, because he [the guide] could point.*" Comparing the second and sixth rows of Table 3, when presented the option of using a virtual pointer, seven of the eight groups used it (the tracking failed in this scenario for the eighth group). In the remote/virtual scenario, almost all participants oriented the workspace to be able to view it from a symmetric perspective, which often resulted in the semi-transparent virtual head occluding their view. Participants seemed ambivalent about the virtual head, yet

when explicitly asked if they wanted it disabled, only one group desired to do so; we believe that this allowed the collaborators to better understand the viewpoint of the remote participant. In the co-located/virtual scenario (see Figure 26), builders only occasionally looked up, spending most of their time focusing on the workspace.

One guide suggested that the virtual arrow and finger should be aligned such that the finger tip and arrow tip intersect. They felt that while having a separate method for referencing physical and virtual content was appropriate, the fact that they were disjoint was unnatural - and by inferring multiple directions, could become ambiguous.

5.6. Discussion

Because referencing is a collaborative, rather than individual, task we were consequently interested in the typical cooperation and communication within each pair. There were, however, several surprises from the study. First, besides acknowledging references through utterances, we observed several builders formally re-select spheres, indicating that they were interested in clarifying the reference. We also observed that when references were *not* acknowledged, several guides left their reference in place; this occasionally blocked the view of the builder – hindering collaboration. We therefore view acknowledgement as an important social protocol. Second, the success of scenario 6 (remote/augmented) was unexpected, given that the guide was incapable of viewing their local workspace (including their own, physical hand). However, since their orientation towards the workspace was approximately symmetric to that of their collaborator, guides quickly understood the mapping between the virtual arrow and their physical hand (even though they existed in physically different locations). Third, participants choose to keep the semi-transparent, virtual head of the participant– even

when it partially occluded their view. The head was included to support awareness of the viewing direction of the collaborator, and was intended to strengthen presence in remote scenarios.

The intent of our study was to explore the collaboration styles that support manipulation in molecular modeling. The models provided many parallel characteristics to molecular structures. The activities of this study paralleled the scientific model use, but in the form of a toy that could be used by anyone, not just scientists with specialized knowledge; this approach allowed access to a broader participant pool - yielding insight into effective interaction guidelines when such extensive feedback would not have been available from the scientists.

While our participants were not scientists, they did bring specific skills and experiences. We noticed a direct correlation between the amount of video game experience and the ease with which virtual models were built (to our amazement, one student claimed to play 10 hours per day). This was especially apparent during the five minute training period before conducting the virtual scenarios; one builder constructed a model with approximately 15 spheres, each of which was highly connected.

By framing the interactions with respect to real-world scientific activities, we were able to narrow the design space of AR referencing interactions. These observations point to key points in the development of inter-referential awareness for cooperative collaboration:

1) Pointing benefits from the use of natural hand gestures.
2) Shared viewpoints are an effective medium of communication when generating references

Pointing - a fundamental requirement for collaborative augmented reality - was accomplished using both physical and virtual techniques. When working with co-located physical models, participants preferred to use their hands. When co-located with virtual techniques, they used both. However, when working with virtual objects or in remote situations (where they potentially had no other choice), the arrow was heavily used. Only the virtual pointer was used effectively and with any regularity; the bounding box and color change were not used to represent the items of interest.

The use of shared video served as an invaluable medium for several reasons. First, a single point of view is often inadequate to disambiguate specific objects, especially when stereoscopy is not available. Users physically manipulated their own body or their workspace to obtain a different view of the model. Our participants "manually" changed their view to see the workspace from their teammate's point of view, or to see around the occlusion of their teammate's head. Even when stereoscopy is not present, other factors (such as occlusion by other artifacts) can be overcome by multiple viewpoints. Second, shared video allowed guides to become aware of the remote environment (i.e. allowed them to identify which objects were present as well as their state), and reduces the context gap between collaborators by removing any viewpoint asymmetries that exist. This argument was strengthened in both in scenarios 2 and 6, where guides claimed this to be an effective form of communication - as it provided the "*exact view*" of the workspace from the view of the builder. The guides also indicated the ease with which references could be made, most often generating references relative to their "shared body". Given the success of these configurations, we recognize that allowing the

users to optionally exchange video can help to reduce referential ambiguity in collaborative environments.

While lack of stereoscopy may hinder interaction, it is not a requirement. Designers of 3D models often work with 2D representations during the process of construction through the use of multiple views, rotation and zooming. Further, many people successfully negotiate depth in non-stereoscopic first-person video games through the inclusion of perspective. Thus, augmented environments could leverage off of these features as well. To strengthen this argument, we observed participants changing their perspective - essentially rotating the model to deal with the lack of stereoscopy. The feedback of a few groups explicitly mentioned that having the ability to rotate the model within the workspace would help them collaborate. Adding multiple views to provide additional depth cues should be explored as an alternate to stereoscopic views.

Addressing the third key discussion point, we found that participants use (and requested additional) virtual tools to generate references more effectively. As discussed in Chapter 6, augmented environments can explicitly provide visual support for the verbal communication of references. We hypothesize that providing more virtual reference points will reduce the amount of workspace rotation that occurs when resolving ambiguous locations. Based on user feedback, the visibility of such reference points should be controlled by the participants to reduce virtual clutter. The study also suggests that when working with augmented and virtual models, a virtual reference point to where the *last* action occurred might help collaborators with sequential tasks - supporting referential chaining.

There were referential options in our system that were not used. The guide rarely changed the representation of selected spheres. Originally, selected spheres were white and unselected were black; however, white and black spheres tended to blend in with the environment, and cycling the transparency of selected atoms only worsened the situation. After learning how to change the representation, a majority of the guides left it in the color-cycling mode, though no explicit questions about representations were asked. The bounding box was used by just one group, and only a few times. We believe this is because of the given amount of space was relatively small, so volumetric references were not necessary for collaboration; users seemed to be more comfortable with the arrow, which seemed adequate for references to a specific point. The reluctance to adopt this referencing technique may also be an artifact of inaccurate tracking.

Another surprise was the difficulty groups had while working with virtual object in co-located space (scenarios 3 and 5). We expected co-located scenarios to be preferred by the participants. However, users preferred having their own workspace - one that they could manipulate. This reaction is attributed to several factors, such as the tightness of the shared workspace (approximately $30cm^3$). Increasing the available shared workspace (if available) or offering alternative, compact referencing techniques may help to alleviate this. Further, though some groups decided to sit next to one another, most groups placed the workspace between them (agreeing with Kiyokawa's findings [77]). This caused tracking problems – as the hand of the guide would occlude the markers that defined the world coordinate system. Again, offering a larger tracked area (when possible) might help alleviate both the tightness and the occlusion. These results suggest that when working with purely virtual content in a compact environment,

referential difficulties may be ameliorated if the workspace is duplicated and distributed - providing teams with more virtual space.

Finally, two of the guides experienced difficulties in making references, falsely assuming their virtual arrow projected correctly into the workspace (see Figure 27). We believe that this problem can be attributed to the lack of stereoscopy. The ambiguity behind this can be described by the framework in Chapter 4, and occurred for two reasons. First, while an orthographic projection of the arrow onto the sphere(s) of interest O_i *appeared* correct to the guide, nothing was technically selected, and thus O_i does not equal O_a. Further, because the arrow was distant from the workspace, the set of objects inferred (or O_p) was too broad, and thus O_i did not equal O_p. Note that if the projected arrow had actually selected correctly (where $O_i = O_a$) and a standard representation (e.g. 'highlighting') was used, ambiguity could still exist if the arrow infers a different set. Further, it should be noted that the projected reference became ambiguous because it was observed from a viewpoint other than that of the guide; if the builder shared the viewpoint of the guide, the reference would have more meaning. Using the framework also explicates why the virtual pointer and finger should occupy the same space: each is an individual representation, and thus can infer different object sets.

5.7. Summary of Exploratory Study

Mixed reality environments provide new opportunities for exploring and manipulating objects within three-dimensional space. Such environments enable medical and scientific researchers to collaboratively visualize and interact with models in ways not possible in the purely physical environment, and as such, we must better understand

how teams work together within mixed reality. In this chapter, we presented a prototype that instantiates a subset of selection and referencing interactions for mixed reality environments to determine which techniques are preferred by participants across a variety of collaborative scenarios. The study encouraged the cooperative building of physical and virtual models and leveraged natural physical interactions as a metaphor for the augmented techniques. This study contributes to the development of collaborative augmented reality environments by culling the design space of referencing and providing guidelines for future interaction implementation; it is our hope that our findings can be applied to general collaborative interactions, rather than expert and tutorial-style configurations.

The kinds of references generated are dependent on the media affordances of the mixed reality system and the context of use (e.g. co-located or remote environments). Overall, our study suggests that collaborators 1) need the ability to point (both physically or virtually), 2) exhibit many of the behaviors from general CSCW (which must be supported), 3) may have referencing challenges when stereoscopy is not provided and 4) use video sharing to effectively disambiguate other communication channels for selection and reference.

This study exposed some of the difficulties with referencing in AR, and provided us with possible solutions. There were several areas where this system could be improved by incorporating user feedback - exposing the need for a follow-up study. To overcome scenarios where space is limited, one approach may be in providing co-located participants the ability to replicate the shared workspace; this option is available for virtual content, but may be useful in accommodating referencing when a single

workspace has space constraints – creating a quasi-co-located scenario. Further, we were shown the strength of sharing viewpoints and our interest in understanding its impact on referencing increased; we felt the need for allowing dyads to arbitrarily share views during collaboration and studying the effects. Many of these ideas have been incorporated into a follow-up study, which can be found in Chapter 7.

This study also left us interested in sub-surface construction and related referencing techniques (see Figure 28). While this is a functioning part of this system, we felt that adding more scenarios was too much to include in one study. Our interest is in helping to support collaborative environments where, even when correct occlusion is present, physical referencing is awkward or impossible. For example, in collaborative augmented surgery, a virtual tumor may reside beneath the skull, and thus the proximity from which the reference can be made is restricted. In such scenarios, it is crucial that teams have the ability to make accurate references at arbitrary depths.

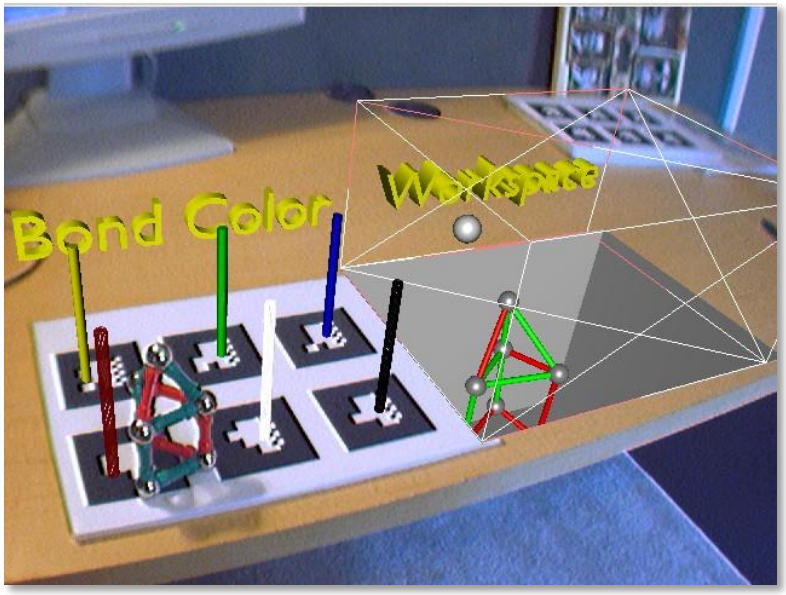

Figure 28 - A Sub-surface Model

Given the strong desire for stereoscopy by our participants, we are interested in its effects during referencing. Though Kiyokawa has shown that collaborative tasks are more efficient using stereo- vs. mono-scopic viewpoints, as with many other AR systems, stereoscopy is not currently supported by DART; however, the cost of transitioning to a different architecture is substantial. Therefore, we were interested in alternative approaches to lessening the effects of a bioscopic viewpoint. Several participants suggested overlaying a virtual grid as an approach to disambiguating references. Others suggested including a projected point directly below the tip of the virtual pointer – indicating the need for alternative depth cues, such as shadows. The efficacy of including shadows during referencing tasks can be found in Chapter 6. The effects of embedded virtual reference points can be found in Chapter 7.

CHAPTER 6

STUDYING THE EFFECTIVENESS OF VIRTUAL POINTERS

In studying embodied actions in collaborative workspaces, Robinson discussed the relationship that exists between the physical environment and the participants [39]. He noted that "*Pointing is the classic example of an action used to maintain indexicality*" and that "*The interpretation of what is being pointed at is dependent not just on the act of pointing but on other people being able to perceive what is being pointed at".* The exploratory study from Chapter 5 provides insight into how participants refer to objects across a variety of scenarios when equipped with virtual and physical referencing techniques; consequently, AR referencing techniques should function well across modalities as well as in co-located and distributed scenarios. One of the most primitive referencing techniques (and currently, one of the few) that satisfies these requirements is a virtual pointer. When considering Robinson's statement on the role of human perception in referencing, we were subsequently interested in the effectiveness of a virtual pointer as a referential representation.

In order to better understand its efficacy in collaborative augmented reality, this chapter presents a two-part study that independently examines how individuals both give and interpret references using this technique, as well as factors that influence accuracy [126]. Here, we present the results of these two sub-studies in the context of designing support for demonstrative referencing in collaborative augmented reality spaces. It was at this phase of research that we began to understand that when the probability of referential ambiguity is high, additional costs such as time, computational resources or alternative techniques can help reduce the ambiguity.

6.1. Motivation of Study

A common method of supporting references in collaborative environments is to include a virtual pointer for each participant; while primitive, this technique has several benefits in AR. Similar to physically pointing, they are multi-modal in that they can refer to physical and virtual content[26]. Virtual pointers are also flexible enough to work across remote and co-located scenarios, or environments that are a hybrid between the two. Additionally, they are analogous to the way humans naturally refer to objects - as they are an embodiment of direction. Finally, they can be spatially registered in 3 dimensions and are trivial to implement.

The ability to refer to artifacts is fundamental to collaboration, yet few studies have explored how to support this in AR. Here, we present the results of a study that explores the effectiveness of virtual pointers and, more importantly, in what context they may become ambiguous. To better understand the appropriateness of such a representation in the context of our framework, our study was decomposed into two, independent sub-tasks. The first half of the study, as described in section 6.2, examines how participants give references using a virtual pointers as well as factors that influence accuracy. This portion of the study is strongly tied with the behaviors we observed in the exploratory study of Chapter 5; many participants were observed giving projected references – or those which appear correct from the viewpoint of the reference initiator when "projected" into the environment. Thus, we were interested in ways of alleviating these through the inclusion of shadows (to provide depth cues) and the orientation of the

[26] However, occlusion between the virtual pointer and physical objects may or may not occur.

arrow (to eliminate a dimension when referencing). In concordance with Robinson, the second half of the study considers the human factor of referential *interpretation* (again, the representation phase), and examines how properties of the arrow (such as opacity, proximity as well as spatial configuration) affect this interpretation; this work is presented in section 6.3.

References can be comprised of several parts and often incorporate a visual representation to draw the attention of others, such as non-verbal cues (e.g. gesturing) or object highlighting; often, deictic speech (e.g. "*this*", "*that*") concomitantly supports the reference. Because trials were performed non-collaboratively, the effects of any deictic speech that would normally support the reference were removed, and thus, the efficacy of the arrow by itself could be studied. The research was piloted with 5 students who were not eligible for the study, and was conducted using 22 students from the general student population, whose ages range from 18-50+. 18 of the students had no previous experience in an AR or VR environments, with 4 participating in previous studies. All trial configurations were presented in a modified Latin Square arrangement (shuffling the order the sub-studies were conducted as well), including the order with which sub-study was presented first. Users were provided with a HMD equipped with a single camera, providing a bioscopic 40°-wide field of view at approximately 30 frames per second. In post interviews, students were given an opportunity to subjectively rate their experiences.

Our contributions include the presentation and analysis of results, our observations, user feedback, as well as a set of guidelines on the appropriate use of virtual pointers.

6.2. Giving References

Experiment Description and Hypotheses

In the first sub-study, we explored how accurately participants could refer to objects using a simple pointer across a variety of conditions. To generate references, users were given a paddle to which a virtual arrow was attached in one of two configurations. In the first configuration, the arrow was perpendicular to the plane of the paddle and thus oriented approximately in *parallel* with the view vector of the user. In the second configuration, the arrow was parallel with the paddle, and therefore generally *perpendicular* to the view vector of the user (i.e. pointing from the side); though the user could attempt to use this configuration to generate *parallel* references, the visual system prevented the paddle from being tracked at high angles, and would therefore fail. To better understand how *projected references* may be reduced, we included virtual shadows in some trial configurations, which appeared on the plane beneath the arrow and the target sphere (see Figure 29).

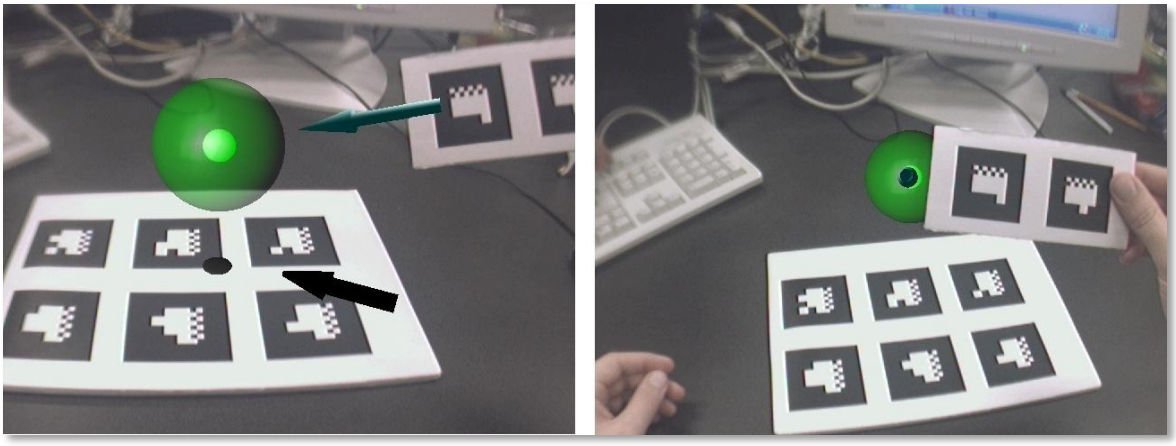

Figure 29 – Referencing in perpendicular (with shadows) and parallel

Users were asked to point to the exact center of a virtual sphere (known as the *target sphere*), which was surrounded by a larger, semi-transparent *barrier sphere* that restricted the proximity from which the reference could be made (i.e. the arrow tip could not come within the barrier). Once users believed that they were accurately referring to the center of the target sphere, they informed the researcher (usually through an utterance) and the next trial began. Accuracy was measured as the minimum distance between the center of the target sphere and the (non-visible) projected ray emanating from the arrow. To determine if the user was "in-line" with the arrow vector, we measured the distance between the arrow's projected ray and the user's view position; we felt this was a more appropriate metric than the angle between the camera and arrow vectors – as the camera vector and what the user is actually looking at can vary (in our implementation, up to 20 degrees). Each of the 22 participants completed 4 sets of 15 trials each (for a total of 1320 trials); however the first trial served as a "training session" to allow them an opportunity to become accustomed to the non-stereoscopic environment. During this time, we reiterated the instructions from the video and explained basic principles of occlusion, shadows, and how to point with the paddle; this trial was subsequently removed from the overall data set. We hypothesized that:

1) Referencing would be more accurate when the viewpoint of the participant is in-line (*parallel*) with the arrow than references generated at higher angles.
2) Increasing distance negatively impacts accuracy when references are made from the side (*perpendicular*).

3) Shadows would provide users with more meaningful depth cues, and thus make perpendicular referencing more accurate. Shadows would have little effect on the accuracy of parallel references.

Analysis

Overall, participants were most accurate when referencing in parallel with the presence of shadows than the other three configurations. Most importantly, we found clear evidence that accuracy significantly increases the more in-line the head position is with the arrow vector (see Figure 30 *a* and *b*). The cluster in Figure 30a shows that the vast majority of "in-line" references were less than 2cm off center, while Figure 30b shows a general shift (up) in accuracy. The right shift in data points from Figure 30a to 30b is caused by the enforcement of orientation in the perpendicular condition; it was impossible to reference in parallel with the arrow. However, the points in both of these plots are concentrated toward the left, indicating that few participants preferred to make references when their head position was far out of line with the arrow vector. Of particular interest is the appearance of a triangular "wedge" in both sets of data. Though there are outliers in this data (perhaps due to inaccuracies in the tracker), it can be clearly seen that there is a "cone" that emanates from the arrow which has serious ramifications on the context with which virtual arrows can be used. One must remember that when giving these references, all participants *believed the reference to be accurate*. Therefore, we can conclude – at least in non-stereoscopic environments – that pointing infers an area proportionate to how in-line the users view vector is with the arrow (i.e. the cone of inference decreases the more in-line the user is with the arrow vector).

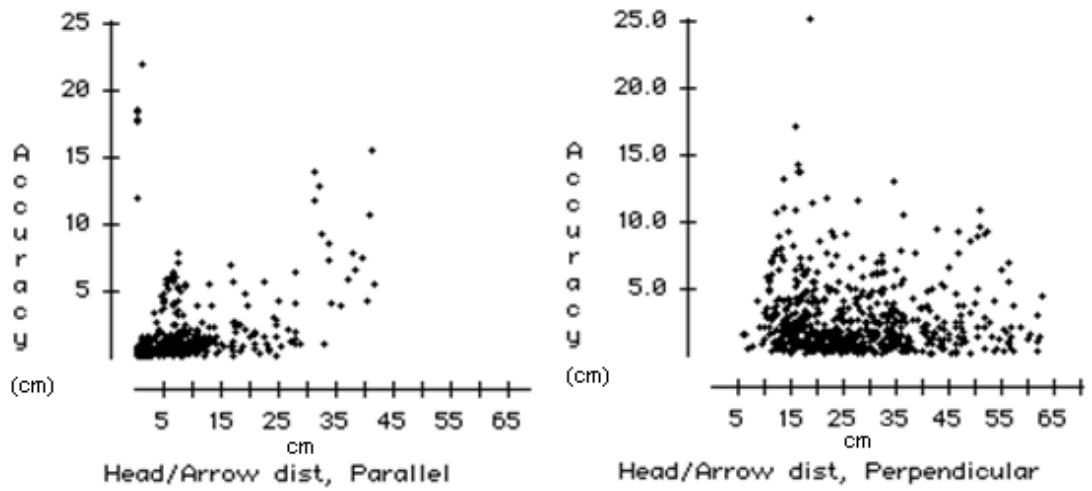

Figure 30 – Accuracy of a) parallel and b) perpendicular arrows

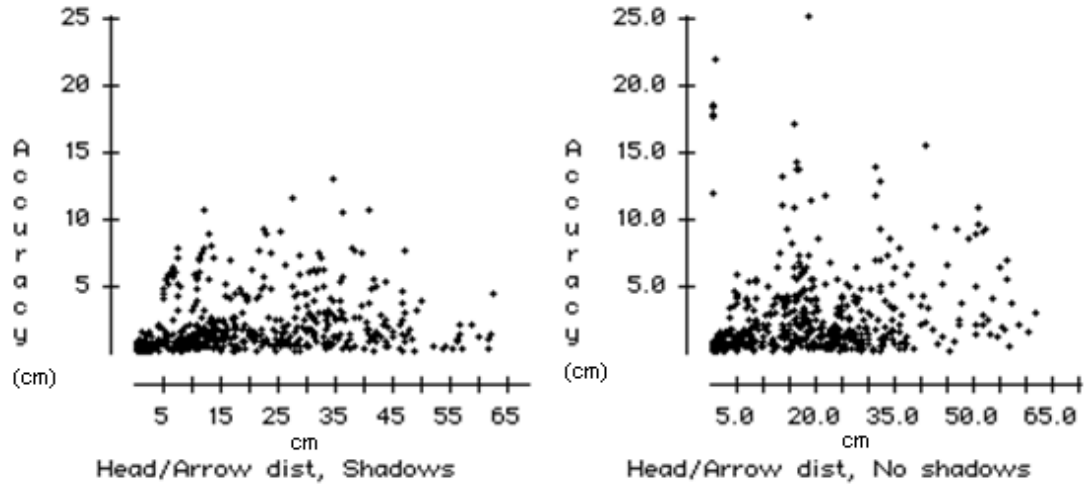

Figure 31 - Shadows vs. no shadow

The data plot illustrates a positive direction which indicates that as the user's view position became more distant from the arrow vector, larger error existed. The data yields an r^2 value of 0.298, and thus accounts for 29.8% of the variation in how accurate the participant was with regard to the distance of the head to the arrow vector. A majority of the clustered points in the lower-left of the plots represent those who gave "dead-on" references – or those where the head position, arrow tip and target sphere were generally in line.

The presence of shadows had a positive effect on the accuracy of the references in the perpendicular configuration (Figure 31). Surprisingly, though less pronounced, it could be seen that shadows increased the accuracy in the parallel configuration as well. The combined attributes of shadows and parallel arrows provided the most accurate results (see Figure 32). Based on a two sample t-test, we found a statistically significant ($\alpha = .05$) difference in accuracy between the mean values of the data when shadows were present and when shadows were absent. The shadow sample mean for distance accuracy was 1.1406 cm with standard deviation of 1.6113 cm, as compared to the no shadow sample mean for distance accuracy was 1.5604 cm with standard deviation of 3.1966 cm.

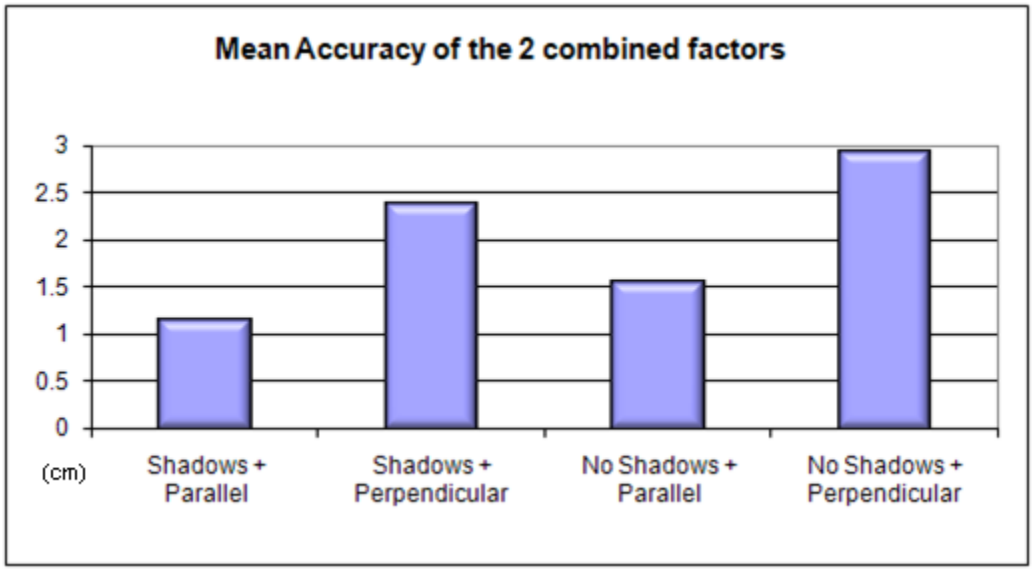

Figure 32 - Accuracy comparison of all 4 scenarios

In the perpendicular configuration, distance was a negligible factor in accuracy (taking into consideration how accuracy decreases proportionately with distance). It had little to no effect on referencing in parallel. Further, there was no significant correlation

between accuracy of referencing and the proximity between the head position and arrow tip.

Of interest is the relationship between referencing time and accuracy (see Figure 33). We placed no time constraints on the time to make the reference - emphasizing only accuracy. Those times that were exceedingly long represent users who repeatedly "poked" the barrier sphere or multiple viewpoints to gain a better understanding of the relative depth of the object. We found no difference in the time taken to make the reference in the parallel vs. perpendicular configurations.

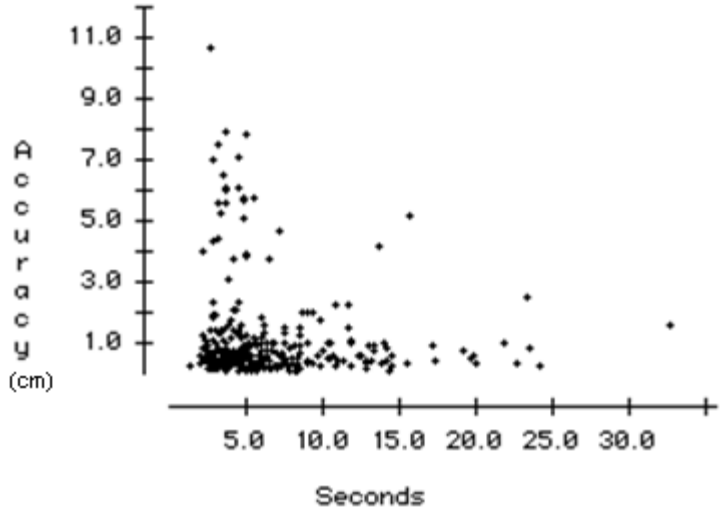

Figure 33 – Referencing time vs. accuracy

Observations and User Feedback

Even when shadows were present, some participants still had difficulty understanding relative depth. In scenarios where no shadows were available, participants were observed gaining depth cues through occlusion between the arrow and spheres and/or multiple viewpoints. To our surprise, in the perpendicular configuration, several users were seen refining their references by repeatedly poking the barrier sphere (from the side) until they were confident the tip of the arrow was the same relative depth as the

center of the sphere. This repeated poking took a significantly longer period of time (as shown in Figure 33 as the data points to the right), yet in post interviews, provided users with accurate enough depth cues to reduce referencing to a two-dimensional problem; once confident of relative depth, they would adjust the angle of the paddle to point to the target sphere, most likely leveraging from proprioception[27] cues. Other participants were seen trying to use the perpendicular configuration in a parallel way; once they realized this was not possible, they opted for as "near-to-parallel" as they could achieve.

Table 4 – Average accuracy (in cm) of users who preferred the perpendicular configuration

UID	**Parallel**	**Perpendicular**
1419	0.68159	1.78266
1421	1.60163	3.83754
1426	6.47181	4.73206
1432	0.42831	2.07348
1435	0.77428	2.27034
1437	2.25233	1.48723
1438	0.38435	1.22959

Many participants based their arrow configuration preference by how *natural* they believed it to be. Of those who responded, 63% explicitly mentioned that parallel referencing was more natural than the perpendicular orientation. In this configuration, one participant commented "*I feel like I have a better sense of accuracy*", while in the

[27] The awareness of the position of a person's body.

perpendicular configuration another claimed "*I had to think* [more] *about the relationship between the ball and the pointer*." 7 participants (32%) preferred the perpendicular configuration and the remaining 5% believe them to be equivalent. However, their results do not confirm this. As shown in Table 4, we extracted those who preferred the perpendicular configuration to examine their accuracy, and found that a majority of these participants were significantly more accurate in the parallel configuration. This contradictory preference is interesting, and requires more investigation. For those who actually *were* more accurate in the perpendicular configuration (UIDs 1426 and 1437), they were considerable less accurate overall – regardless of configuration.

Even though participants received no feedback on the accuracy of their referencing during the trial, of those who responded, 72% felt that the arrow was more effective with shadows present – often expressing that they provide an extra dimension of information and instill confidence that their reference was accurate. One participant, after running through a set of trials in which the shadows *were* included, claimed that the shadows were irrelevant; however, when this cue was taken away (by chance, in the next trial set), they commented "*Wow, that's harder! I didn't realize I used* [the shadows] *that much*". Five of the participants preferred no-shadow configurations, claiming the shadow to be distracting or of no use; however, their performance was significantly more accurate in scenarios where the shadows were present.

6.3. Study of Interpreting References

In the second sub-study, we were interested in exploring how participants interpret references from a virtual pointer as well as how contextual (and other) factors

might influence the interpretation. In each trial, participants were presented with a configuration of 8 cubes, and asked to identify the cube to which the arrow was referring; each cube was uniquely labeled with a value between 1 and 8. The virtual arrow was comprised of an open cylinder (i.e. one in which the ends of the cylinder were absent) and cone, and appeared in one of two modes. In the first configuration, the back face of the cone was not rendered, allowing the users to view through the cylinder to obtain a non-occluded view of the direction of the arrow (see Figure 34). The second configuration was opaque in that both the front and back faces were rendered. The arrow distance varied in each trial, and would point to one of the eight cubes. In some of the configurations, participants could freely move the workspace, while in others, the workspace was mounted on a picture frame (see Figure 35); though they were not able to move the workspace in the later configuration, they were allowed to change their viewpoint by moving their head or body.

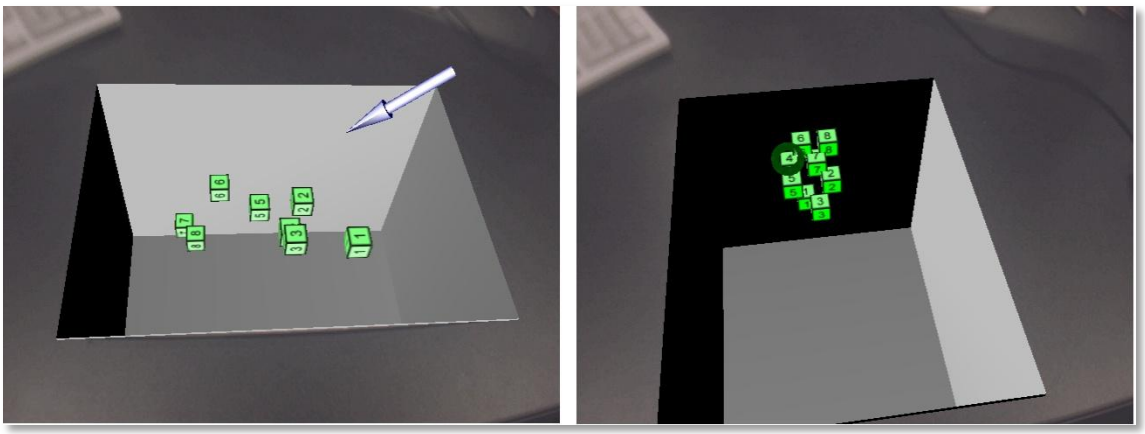

Figure 34 - Opaque and see-through arrows

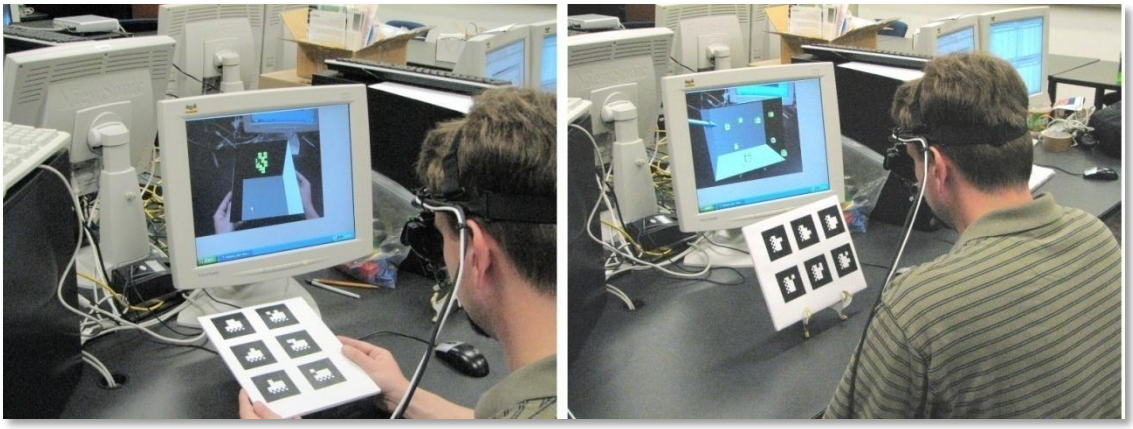

Figure 35 - Moveable vs. stationary scenarios

Finally, we were interested in studying how the spatial relationship of the cubes affected the accuracy of interpreting the reference – simulating clustered scenarios that are found in molecular modeling. Within each cube configuration, one or more cubes were target candidates, based on the properties of the cube configuration; the system chose one of these cubes at random, and changed the orientation of the virtual pointer to refer to it (i.e. the arrow was mathematically guaranteed by the system to point to the middle of exactly one). Though the cubes did not overlap one another, in some trials, the target cube may have been partially (but never fully) occluded by others when viewed from the direction of the arrow. Specifically, the cube configurations were:

0) *Tube* – mostly sparse, but in a long cylindrical pattern. Any of the 8 were target candidates
1) *Tight cluster* – all 8 appeared within close proximity to one another. Any of the 8 were target candidates

2) *Small cluster* – a mostly sparse configuration, yet 3 of the cubes were tightly clustered; any of those 3 were target candidates
3) *Sparse* – all cubes were candidates and were scattered throughout the workspace
4) *Staircase* – the cubes appeared in a descending staircase form, any of which were candidates.

Users were seated in a chair facing the default workspace, and given a few minutes before each trial set to become familiar with the environment, such as viewing the environment from different perspectives (by holding the fiducials) and identifying cubes by number. The experiment used the same 22 participants from the first sub-study, with each evaluating 60 references (30 with the moveable workspace, and 30 with the stationary – again, for a total of 1320 trials). Participants were asked to identify which of the eight cubes the arrow was referring to – calling out its number once they felt comfortable they understood the reference.

We hypothesized that:

1) Users will try to line up with the arrow, and those that do will be more accurate and take less time in deciding than those who do not
2) Spatial configurations will have a significant impact on accuracy of interpreting the reference
3) See-through arrows would improve confidence and time to respond, but not improve accuracy
4) Increasing the distance between the arrow and the cube to which it referred would only minimally affect the accuracy and time to interpret the reference

Analysis

By recording the minimum distance between the head and the arrow's projected ray at the beginning and end of each trial, we found that participants moved "in-line" with the arrow – such that their viewpoint was near-parallel to the direction of the arrow. When the angle relative to the participant's view vector was increased to an "uncomfortable" direction, the participants' response times were longer and far less accurate by at least 15% of the time. The majority of these larger, uncomfortable angles were enforced by the non-moveable workspace configuration; in some cases, it was almost impossible for the participant to become "in-line" with the arrow because of physical limitations.

Table 5- Cube configuration and accuracy

Cube Display	**Percent Correct**	**Percent Incorrect**
Tight Cluster	75.8%	24.2%
Tube	76.5%	23.5%
Small Cluster	86.0%	14.0%
Sparse	90.9%	9.1%
Staircase	92.4%	7.6%

The spatial configuration of the cubes had an effect on the accuracy, with the clustered configuration acquiring the most inaccurate responses (see Table 5). Using the data μ (seconds for correct responses) of 6.39 with a corresponding standard error of 0.615 seconds and the Gossett t-model, one can be 95% confident that the mean

selection time for a *correct* response by the 22 participants was anywhere from 5.111 seconds to 7.669 seconds. Comparatively, the μ (seconds for *incorrect* responses) was 12.16 with a corresponding standard error of 2.673 seconds. This supports the conclusion that the time taken to respond was directly related to the accuracy of the response; participants who took longer to respond were usually less accurate. In collaborative pairs, such delays may indicate confusion (in the acknowledgement phase), and therefore may be alleviated through social protocols, such as a re-iteration or alternative representation.

The accuracy of the see-through vs. opaque arrow was inconclusive. Regardless of distance, the performance was consistent for both configurations. Surprisingly, the distance between the arrow and the target cube had the greatest impact on the accuracy of the interpretation, with a minimum success rate of 94% or more for when the arrow was close to the target cube.

Observations and User Feedback

Participants *overwhelmingly* felt the need to line up with the virtual arrow; consequently, they preferred the scenario when the workspace was moveable. When asked why, users commented that it was easier to move the workspace than themselves, with one user stating "*it was a whole lot easier for me to move my hands than my head*" and another that "*there were more angles to view from."* In configurations where the workspace was stationary, participants rocked back and forth to view the references from multiple perspectives – especially when they couldn't line up with the arrow. This configuration was especially troubling to several individuals when it was nearly physically impossible to line up with the arrow. Much to the surprise (and paranoia) of

the researchers, several participants contorted their bodies (with heavy leaning and some getting out of their chairs) to come into line with the arrow or to gain a better viewpoint.

Though it is evident that users will attempt to view the arrow along its length, observation suggests that the data presented here is a loose upper-bound of the extent to how much this occurs. Even though the distance between the head and the arrow vector was measured at the beginning and end of each trial (i.e. the time at which the user spoke the number), a more accurate measurement would have been to record the *minimum* distance that occurred. Participants often tried to view the reference from multiple viewpoints – examining the cubes, then the arrow, then cubes again; as a result, they sometimes responded in the middle of this behavior, causing the data to be skewed away from being "in-line".

Of those who responded, 94% of the participants explicitly preferred the hollow arrow, often claiming that "*it acted like a* [gun] *scope*." It is believed that this technique *guided* the user to view from a specific viewpoint, which then provided them a more-accurate "orthographic" view – essentially eliminating a dimension as described by the participants in post-interview responses. When asked if any new features would help with the task, many participants suggested variations of a ray (e.g. a variable-length ray). This need became evident through observation as well; one user was seen physically extending the virtual arrow with his finger in order to more accurately gauge the direction of the arrow.

The overall success of the staircase configuration was a surprise; we believed that in cases where the direction of the arrow was nearly in line with the staircase, interpreting the reference would become difficult. However, even minor differences in a

second dimension seem to provide enough information to determine the point of intersection between the line of the staircase and arrow vector. Other configurations suggest that there exists a *cone of inference*: interpretations become less accurate with distance. Further, because the environment was non-stereoscopic, participants repeatedly refined this cone through multiple viewpoints.

6.4. Discussion

We have shown evidence that, when giving references using a virtual arrow, users are more accurate when in line with their arrow, and often prefer parallel to perpendicular configurations. Similarly, those that interpret the references using this technique prefer to be in line as well. These seemingly contradictory requirements strengthen the argument for the availability of shared viewpoints between the reference initiator and receivers: without it, one of these parties is forced into a less-accurate referencing scenario.

It was hypothesized that distance would only partially affect the accuracy of the response, given our pilot observations on how participants physically position themselves to be along the length of the arrow. However, the diameter of the arrow cylinder was fixed, and thus, when viewed at a distance, inferred a larger area. Given that the vast majority of participants preferred this technique, consideration may be given to decreasing the diameter of the cylinder to reduce the inference - especially when the object of reference is distant; however, dynamic diameters require the system to be aware of the object of interest, which is similar in nature to selection.

It could be argued that the virtual pointer is too ambiguous of a reference representation and a simple highlighting scheme should be used instead. However, this

approach is insufficient for physical objects unless its pose and geometry are known ahead of time. Further, our observations from Chapters 5 and 7 indicate that it is natural for users to point to objects using both physical and virtual techniques; the human finger is similar in properties to the arrow. Finally, if objects *are* to be highlighted, they must first be selected, and are subject to many of the complex attributes found in Chapter 4.

Our overall understanding of referencing in AR is that there is a cost associated with disambiguating references, which can manifest itself in the form of time or computing resources. If shared viewpoints disambiguate references as the data suggests, additional network bandwidth processing must be allocated. More time taken to generate accurate references reduces the efficiency of the group. If alternative techniques are used, they must be multi-modal; otherwise, users will be forced to switch between multiple referencing technologies. Further, while potentially more powerful, alternative referencing techniques might require specialized hardware and the training of participants.

6.5. Summary

Though pointing is one of the most natural ways to refer to objects, when the probability of referential ambiguity increases because of environment factors (such as clustering, occluded viewpoints, etc), additional referencing support should be provided; this support could be in the form of alternative referencing techniques, additional time taken to ensure accuracy, or techniques that support referencing - such as shared video or embedded reference points. However, this comes at the cost of group efficiency (i.e. time), effort from the users in the form of training and usability, or additional computational resources that support them (e.g. network bandwidth, advanced rendering

techniques, or specialized hardware). Conversely, when the probability of ambiguity decreases, these forms of support can be relaxed.

Researchers often over-emphasize the virtual aspects of AR systems; admittedly, this study has as well. Here, we have studied the virtual pointer – as it is capable of referencing multi-modal content. However, if other virtual selection techniques are to function in AR (such as raycasting), geometric representations of the physical objects in the environment are required a priori; in the case of dynamic environments, they must be tracked as well. When this knowledge is not available, reference techniques should be multi-modal (e.g. a physical+virtual laser pointer or those that do not rely on intersection) - placing less of a burden on the user by not requiring separate referencing techniques.

This study yields insight into the effectiveness of giving and receiving references using a virtual arrow. We found that when giving references, accuracy increases when the view of the user closely parallels the direction of the arrow, and are less susceptible inaccuracies caused by distance. Further, the inclusion of shadows helps to resolve depth, and subsequently allows participants to generate more-accurate references. Similarly, the second half of the study showed that, to better interpret the reference, participants preferred to line up with the arrow, and took more time and were less accurate when they could not. Further, the accuracy of the responses was sensitive to the configuration of the environment; when multiple objects fall within the direction of the arrow or when the proximity of the arrow to what it is referring to increases, references can become ambiguous. This study also indicates that the arrow infers a conical space which increases when viewed at higher angles. When combined, both

parts of this study demonstrate the need for shared viewpoints; without them, one party is forced into a less desirable referencing scenario.

CHAPTER 7

FOLLOW-UP STUDY AND ARCHITECTURE

The exploratory study presented in Chapter 5 provided insight into the referencing behaviors that pairs exhibit during collaborative building tasks. We found that multi-modal referencing techniques should be provided in co-located and remote collaboration, and that a shared viewpoint is effective in generating and clarifying references to artifacts in remote workspaces. To better understand how the *environment* could passively support inter-referential awareness, we conducted a follow-up study where participants were given a similar building task (to enforce collaboration); this research examines several of the concepts of environmental taxonomy found in Chapter 4. We restricted our focus on supporting references in remote scenarios and, based on user feedback from the previous study, incorporated virtual reference points into the workspace in the form of a 2D grid. Additionally, by modifying the underlying system, the environment permitted arbitrary sharing of viewpoints between participants. In our study, this capability was naturally limited to the guide to simulate the expert/technician scenario as well as prevent the builder from having visual access to the target model.

When extending the underlying system to support shared viewpoints, we discovered a variety of ways in which augmentation and tracking could occur; though implementation is dependent on the context and constraints of the application, it leads to an interesting discussion on flexibility and design. Participants now receive *multiple* video streams (potentially simultaneously), and therefore tracking and augmentation can occur independently on either one; further, augmentation can occur before or after the video stream is sent to remote participants – or even multiple times. We take a

subscription-based approach to sharing video where each camera is evolved into a video server, and allow systems to negotiate the potentially heterogeneous settings (such as FPS, dimension, bit depth, format, etc.) found in many collaborative AR environments.

The results of this chapter further strengthen the argument of availability of a shared viewpoint in collaboration, clarify the role of virtual referencing techniques during collaboration, and validate many of the referencing and behaviors from the exploratory study. Further, we discuss our architectural design of shared video and the implementation details of remote referencing.

7.1. Follow-up Study

An important consideration when designing any system is in understanding how participants will use it and the kinds of support that are required for them to successfully collaborate. In this study, we were interested in further observing participant behavior as well as in receiving subjective user feedback about their experiences. This follow-up study explores how participants refer to objects in remote scenarios, and is similar in task to the research in Chapter 5. We focus on remote scenarios as they present unique challenges not found in co-located collaboration; many of the non-verbal forms of communication are removed, and we examine how to alleviate this by supplementing the environment with computer-mediated techniques.

Knowledge was externalized (for the guide only) in the form of a virtual model representing the target configuration (see Figure 36). The builder was required to construct its physical equivalent in a remote environment using wooden blocks. A total of 16 users participated - some of whom participated in the exploratory study, and were subsequently grouped into 8 pairs; roles were negotiated within each pair before the

trials began. Subjects were interviewed after each trial, where they were asked to describe the level of support the environment provided, and asked to disregard the complexity of the model as a factor. Subjects were also allowed to rank the environments relative to one another after all trials had been completed.

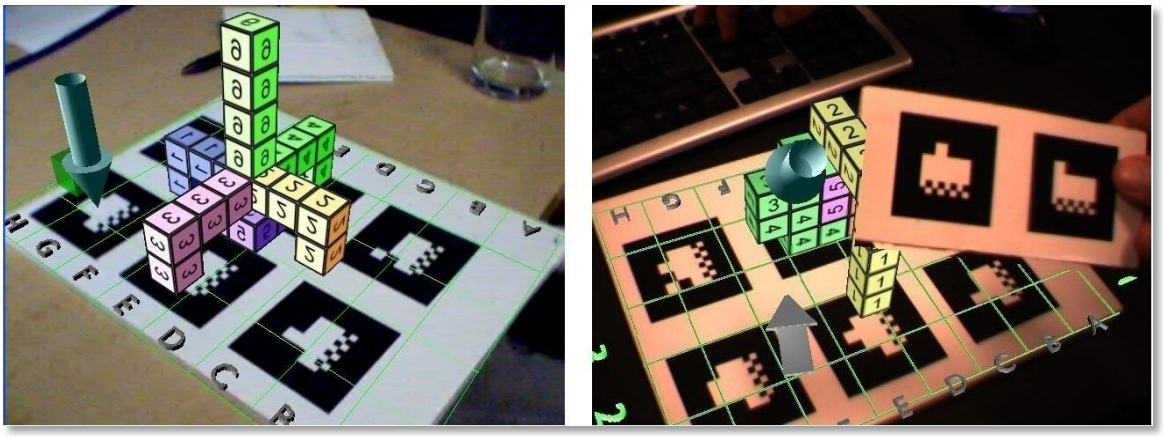

Figure 36 - Configurations from the guide's view

Study Design and Setup

Using participant feedback from our prior study, we were interested the efficacy of including a virtual grid to serve as an embedded reference point in the local workspace of each user. We were further interested in the role of the virtual pointer during collaboration. In all configurations, the guide was able to view their local environment (where the virtual model was present) as well as toggle their video feed to view the remote workspace from the view of the builder. Unlike the exploratory study where each subject was equipped with a Twiddler (a one-handed keyboard to prevent them from searching for keys), video toggling was the only option for the guide - and

was accomplished by pressing the space bar[28]. No keyboard interaction was necessary for the builder. The subjects were separated by a small barrier and were able to talk with one another, but were physically unable to see one another. Both the builder and guide were given approximately 5 minutes to become familiar with the system. As most of the burden fell on the guide, this time was used to train them in how to toggle the video as well as how to become familiar with their local, visual tracker. Similar to the study in Chapter 6, a small panel of fiducials was given to the guide, which allowed them to make references using a shadowed, virtual arrow.

The study was a modified Latin Square arrangement of the following configurations:

1) *video-only*: the guide could toggle the view between the local and remote environment. No virtual referencing techniques were allowed.
2) *video+arrow*: the same as scenario 1, but the hand of the guide was tracked within their local environment. When viewing locally, the arrow of the guide appeared in world coordinates for both participants; thus, the workspace could be viewed independently by each user, and the orientation of the guide's arrow was viewpoint-dependent. However, when viewing the remote environment, the guide's hand became relative to their shared viewpoint, providing them with a 3D cursor into the remote workspace (i.e. the guide's hand was tracked using camera coordinates and mapped into the builder's camera coordinates such that their views were identical).

[28] We chose the space bar because it is unique in size as well as placement on the keyboard, and thus is easily identified through the HMD. However, identification occurred only at the beginning of each trial.

3) *video+grid*: same as scenario 1, but a virtual grid appeared over the workspace. No arrow was present.
4) *video+arrow+grid*: a combination of scenarios 2 and 3.

Each model was comprised of 6 pieces, though the configurations differed significantly in piece orientation and the degree to which pieces occluded one another. The pieces themselves were shaped similarly to those found in the classic video game, Tetris.

Though performance time was measured, we were most interested in observing how references varied across conditions, and how participants perceived the ease of giving and receiving instructions to complete the task. Teams were interviewed after each trial and rated their experiences on how well the environment supported referencing (excellent=4, good=3, fair=2 or poor=1). Further, they were given a chance to discuss and rank the scenarios relative to one another after all trials had been completed (from 1 to 4, with 1 being the most preferred). Specifically, we hypothesized that:

1) the *video+arrow+grid* configuration would be favored over other configurations as it provides the most support, with the video-only configuration being least-favored. We believed that redundant referential options would clarify ambiguous references.
2) guides would spend a majority of their time viewing the remote video feed

3) the arrow would be useful until the task was completed, but the grid would only be useful initially. Thus, the *video+arrow* configuration should be preferred to the *video+grid* configuration.

Observations and User Feedback

Training effects had a significant impact on the time to complete the task. Figure 37 shows that teams were far more efficient in the last trial than the average of the first three - regardless of configuration. As teams became comfortable with the task, common ground was established; over time, guides appeared to give better instructions (and references) and builders carried them out more efficiently. For example, it was necessary to establish a common *mental coordinate system* within the group; "*up*" for some meant farther away from their body – as if the workspace were a whiteboard – while for others, the same word meant increasing in elevation.

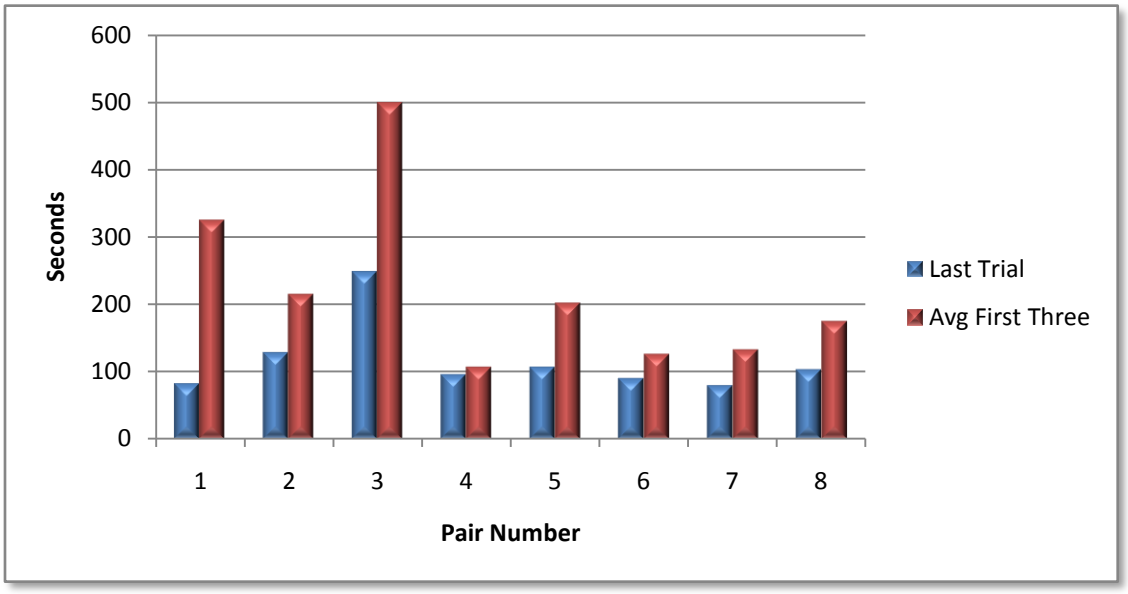

Figure 37- Training effects between pairs (1-8) in seconds

Through observation and post-interview questions, we found that participants found it crucial to leverage off of the virtual referencing techniques to establish an *initial* reference point in the workspace. Additionally, our observances reinforce the idea that participants make heavy use of *referential chaining* – or using the last referenced point as a relative basis to create a new one. We further noted that, even though virtual shadows were included, the bioscopic view caused a few guides to (*still*) give *projected references* (i.e. the arrow tip was placed approximately halfway between their viewpoint and the object to which they were referring, creating ambiguous references). Fortunately, guides most often worked in the remote view (64% of the time on average) in which projected references have meaning for both participants.

Two of the guides experienced difficulties in hand tracking (caused by lighting conditions and limited field of view of the camera), causing them to become frustrated and use the arrow less; subsequently, the pairs rated this configuration lower. However, when referential chaining or relative referencing *failed*, these same guides reluctantly returned to using the arrow (at which point they quickly clarified the reference). One (right-handed) guide was seen pointing with his left hand, even though his remote colleague could not see the gesture.

When asked to rank the environments relative to one another on how well they supported collaboration (with 1 being the *best* and 4 being the *worst*), the *video+arrow+grid* configuration was favored (see Figure 38).

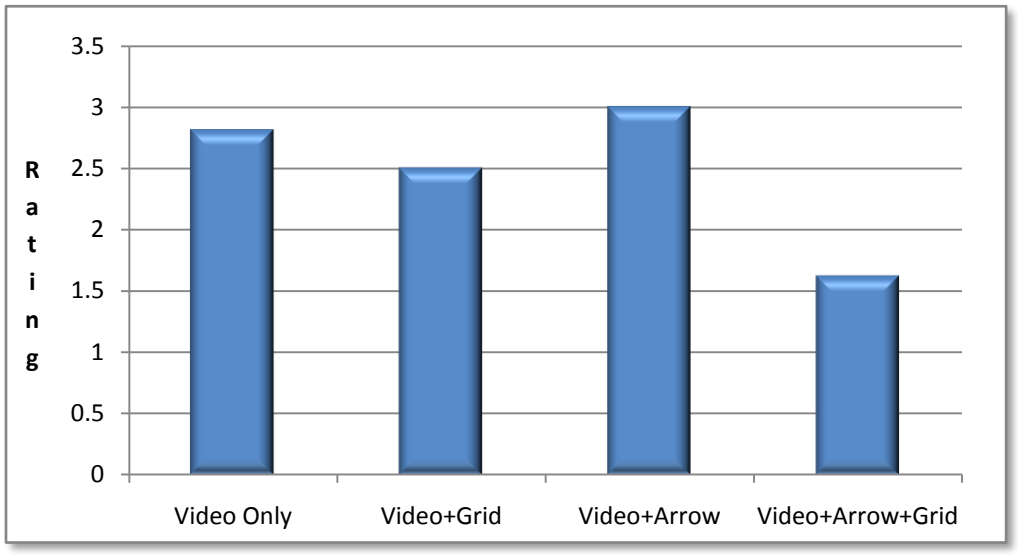

Figure 38 - Relative rank of environments

When asked how well each environment supported collaboration *without* regard to one another (with excellent=4, poor=1), we found that, as expected, the *video+arrow+grid* scenario was the most preferred (Figure 39). Surprisingly, the *video+grid* configuration was preferred over *video+arrow*, and the *video-only* was rated higher than *video+grid*.

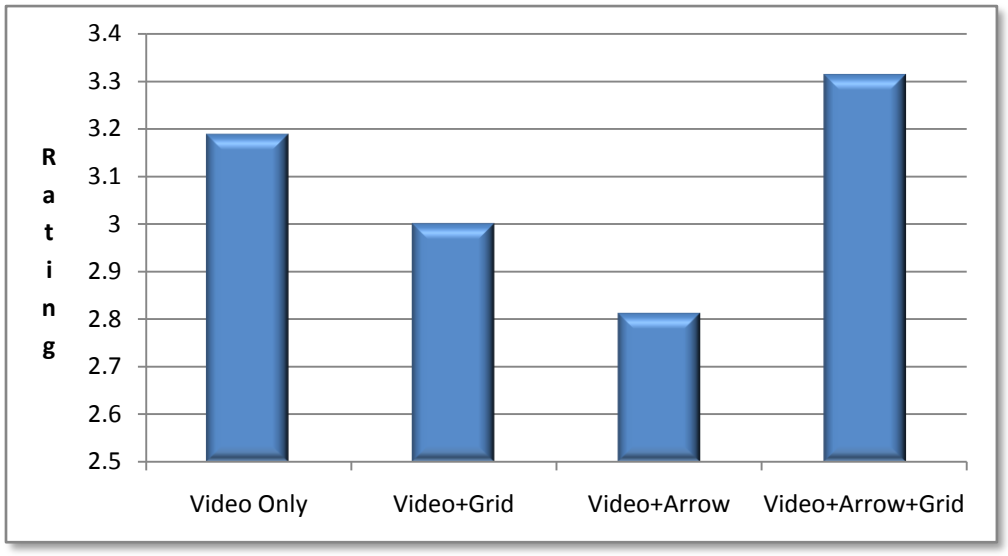

Figure 39 - Independent rating of referential support

The configuration had *no* significant impact on the number of times the guide switched between local and remote video, with the average number of switches to be 20.125 (see Figure 40).

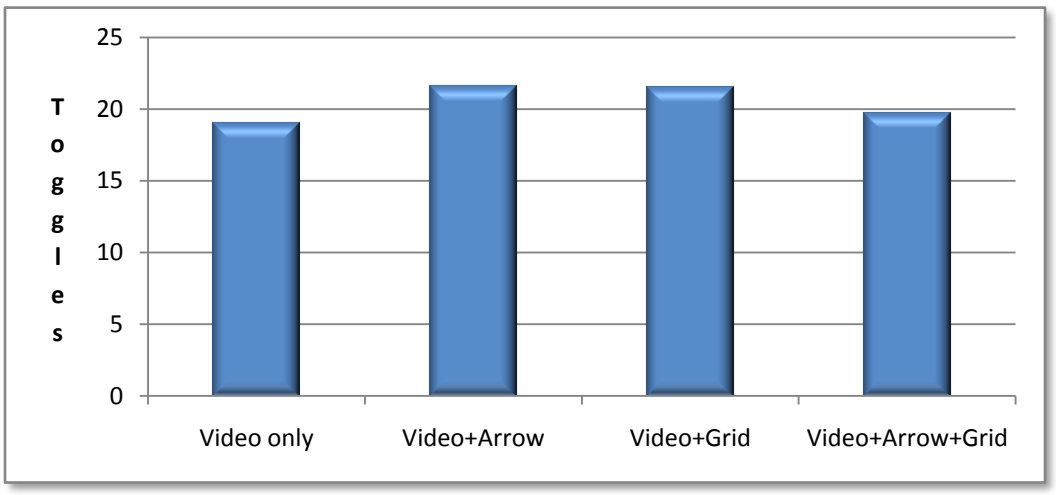

Figure 40 - Number of video toggles during task

7.1.Discussion and Related Work

It came as a surprise that the *video+arrow* configuration rated significantly lower than the *video+grid* configuration. We believed that the arrow was a much more interactive way of referring to content, and given that the grid was only 2-dimensional, would be superior in the space that could be referenced. However, these results may be an artifact of the implementation; when asked which techniques might better support the act of referencing, almost all guides expressed the need for better tracking. Further, many groups clearly stated that the arrow was an *invaluable* tool, which was also observed in their behavior when references could not be clarified using speech alone. The ranking of the *video+grid+arrow* scenario leads us to believe that the combination

of techniques was most beneficial to participants, and increased the perception of referential support the environment provided.

During the course of each trial, there was a decline in the use of both the virtual arrow and grid. Once a point was established, virtual techniques became less important; the "connectedness" of the model seemingly aided the collaborators more than which techniques were present. We imagine that, if given an extremely disjoint model (e.g. one in which relative references would become more ambiguous and referential chaining could not be used), or if the task is more exploratory than constructive, virtual referencing would play a more important role. Further, several groups mentioned that the grid and arrow should be visible only when needed, and thus their visibility (or transparency) be toggle-able; when not being used, the arrow became both misleading and distracting- as it ambiguously referred to where it was last tracked. Similarly, many builders argued that, because physical objects (e.g. their hands and the blocks) did not occlude the virtual grid, at times the grid obstructed more than aided the construction – often stating that the grid "*floats above*" the workspace (see Figure 41).

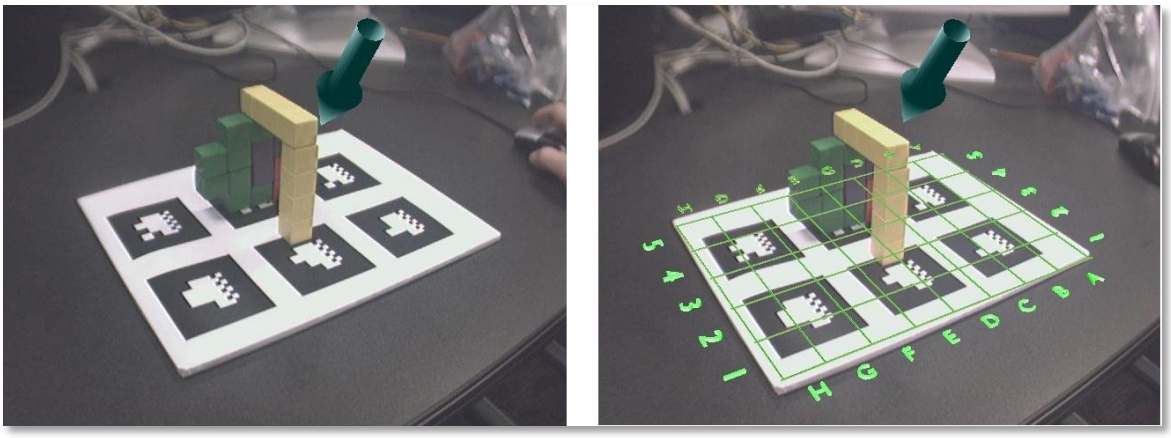

Figure 41 - The "floating grid"

As shown in Figure 37, the average time to complete the task varied widely among groups, yet the final trial time was consistently less than the average of the first three, demonstrating the effects of training. Much of this variation can be attributed to how well the guide can *give instructions*; for example, the guide in group 3 apparently had difficulty in understanding the spatiality of the model, and as a result gave several ambiguous references.

Finally, while this study focused on remote scenarios, many of these results have application in co-located environments as well; referencing techniques must always be present. For example, though participants may exist in the same physical space, their viewpoints will differ; thus, factors such as occlusion create asymmetric viewpoints of the referenced object, and the opportunity for ambiguity increases.

7.2. System Support for Referencing

The studies were implemented using the rapid prototyping system, DART [121]. However, to provide access to the view of other participants, the architecture needed to be extended to allow remote users to independently subscribe to each local camera. Given that our implementation relies on video-based tracking (the ARToolkit), this presents an interesting new "option" in which the tracking and augmentation processes occur; both of these can occur independently from one another using different video feeds. For example, in the guide/builder scenario described above, we required that (in some configurations) the guide's hand be tracked in their local environment (using the local video feed) while augmenting the view from the builder's environment. While it is indeed possible to augment and track using only the remote video, it is the hybridization

that makes it possible to see the virtual arrow in the context of the remote workspace from a shared viewpoint.

Shared Video

In the collaborative study, the interface presented to the guide allowed them to toggle between viewing their local video feed or that of the remote participant. For this to occur, the local camera needs to supply video to the local system as well as those subscribed to it - regardless of the number of users - and ensure that the local frame rate remains relatively unaffected. Further, because of the heterogeneity found in many AR systems, it should work across different resolutions and frame rates, depending on the network resources available.

Our solution is to allow each camera to become a concurrent video server. Each server negotiates video options with the first client to connect, including frame rate, color depth, resolution and video compression type. For efficiency, if a second client connects, the server refuses to negotiate, and defaults to the parameters from the first negotiation; otherwise, an undue computational burden could be placed on the server by providing arbitrary video feeds. To perform negotiations, when the client connects, it sends its preferences to the server. The request can be symmetrical with that of the server, in which case no video conversion needs to occur. This is unlikely for a variety of reasons however, so the server (in its current implementation) chooses to "downsize" to the lesser-of-all options, and returns these to the client. This approach allows the local video system to run using its preferred settings (if computationally feasible), and remain independent from lower requests. For example, if a client negotiates a lower frame rate, the local environment can still receive frames at the higher frame rate, propagating the

latest frame buffer to the clients when necessary. For compression, the protocol supports raw, uncompressed frames, or frame-based compression (JPEG), though the code can be easily extended to include other compression techniques. We leveraged the fact that our "remote collaboration" had a gigabit Ethernet connection, and thus we had better performance using uncompressed video.

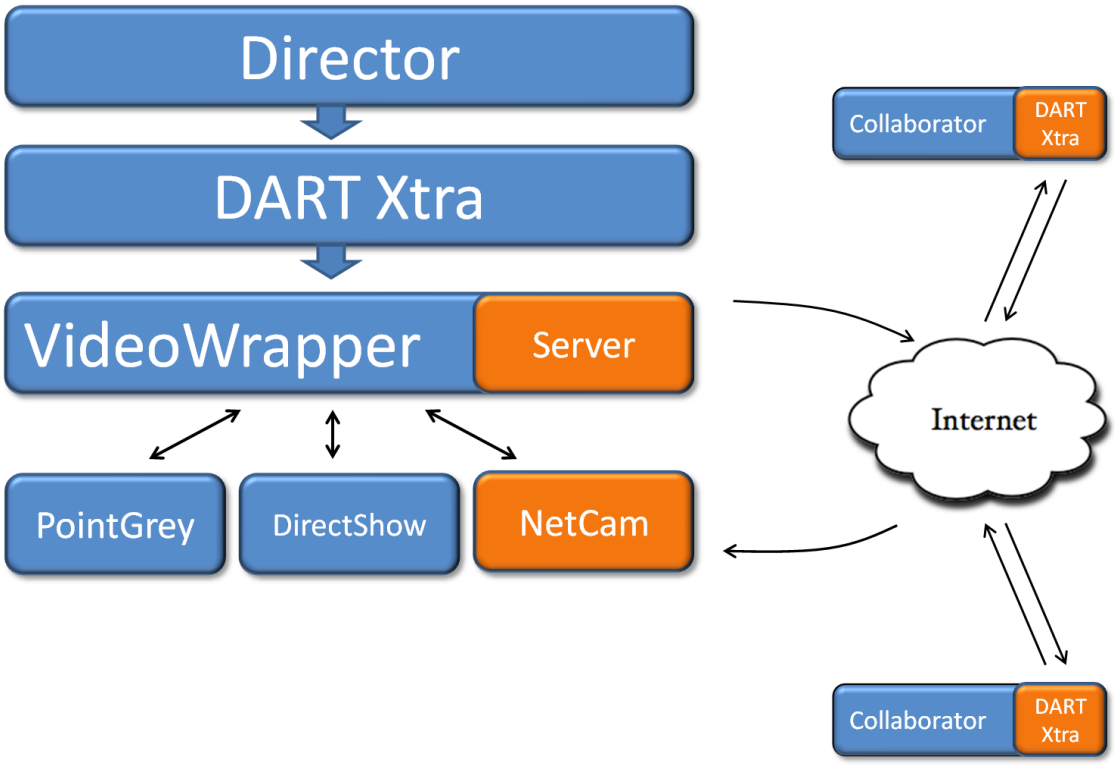

Figure 42 - Layered architecture of DART

DART is comprised of a series of layers, and allows developers to work at higher levels of abstraction – either graphically (using Director's interface) or using a scripting language (see Figure 42). Director then communicates with the DART Xtra (a plug-in) – which is responsible for exposing the functionality of a series of libraries. One of these

libraries is a *VideoWrapper*, which is responsible for providing a generic interface between the Xtra and more-specific camera types; each camera must have a wrapper as well to conform to the VideoWrapper's interface. Thus, calls to the VideoWrapper (such as *getFrame*) propagate to a specific camera wrapper, which then call camera-specific functions. In essence, the VideoWrapper allows the Xtra to generically interface with a wide variety of cameras and query or modify any exposed parameters.

To obtain the desired behavior, modifications to each of these existing levels were required. At the highest level (through scripting), it was necessary to provide an interface for the user to subscribe to remote cameras (entering IPs, port numbers, and the options previously described) and allow local camera(s) to become video servers. These high-level functions, in turn, propagate into the Xtra, which are further propagated into the VideoWrapper. If a camera is designated as a video server, it may spawn one or more threads as each remote client connects; each thread runs independently to handle the various streams to each client.

Along with other camera types (such as DirectShow, DCAM or PointGrey), a new *NetCam* camera type was created. This form of camera initially connects with an existing, remote camera, and receives a video feed using the negotiated protocol. In the case of frame-based video, once a frame is received, it replies with an acknowledgement. This approach accommodates occasional network delays by sending frames only when the client has received a previous one, thus adapting to the network by skipping video frames; it further serves to reduce network traffic when viewing local feeds (i.e. it is not necessary to pull video from a remote camera if it is not being viewed).

Local vs. Remote Tracking

Using a video-based tracking system while receiving multiple video streams presents some interesting possibilities. It is possible for users to toggle between local and remote video feeds in a very normal way, with augmentation and tracking using one stream as a data source; of interest is *when* tracking and augmentation occur, and *which* video stream is used. In our implementation, augmentation occurs only from the local system (i.e. a video server will never send video that is already augmented with virtual objects). Thus, if viewing a remote environment, each frame is augmented as if it were the local video feed. However, tracking can occur on arbitrary video streams (or both), which was a requirement in our implementation.

Recall that when the guide is viewing their local video feed, their hand is tracked with regard to world coordinates. Thus, if the builder changes their viewpoint, such as rotating the workspace, the arrow rotates accordingly; the pair is working "independently" in this sense. However, when the guide switches to the remote video feed of the builder, the arrow must appear in the exact pose to both participants; this, provides the builder a "third arm" with which the guide can reference physical objects 3-dimensionally in the remote environment.

To accomplish this, the hand of the guide was tracked in the camera coordinate system using their local video feed, resulting in a standard 4x4 matrix for the object. This matrix was forwarded over a VRPN connection to the builder, where it was used to place the arrow within the builder's field of view. Though the viewpoint of the guide could be forwarded over the network just as easily (changing the coordinate systems to

correctly place the arrow), we felt camera coordinates were cleaner. However, such a straight-forward approach has tradeoffs. While ostensibly simpler, different cameras often require different calibration - which will ultimately affect their distortion matrix. Thus, using this technique, cameras that are incorrectly calibrated can cause inconsistencies in what is supposed to be a consistent view.

An overall advantage of sending an object's camera coordinates and video frames independently is that the tracking and frame rates can be independent. This is especially pronounced in scenarios where network lag is a dominating factor. For example, if the guide is receiving 1 fps from a remote video stream while the builder is locally receiving 30 fps, the guide can still generate references at interactive rates for both himself and the builder. While the references have the potential of becoming stale in scenarios such as these (where the remote viewpoint changes frequently), many scenarios require the builder to remain relatively stationary. In other words, by streaming the pose of the arrow as a separate communication channel, the environment of the guide can be more responsive.

7.3. Summary

In this chapter, we have presented the results of our follow-up study, which suggest several design principles. First, this reaffirms many of the user behaviors we observed in the exploratory study in Chapter 5. Subjects made heavy use of deictic speech in supporting their gestures, used referential chaining, and made references relative to their shared viewpoint as well as to objects that were present in the workspace. Participants also referred to the models using both their physical hand (for

the builder) and the virtual pointer (for the guide), reaffirming the effectiveness of pointing (both virtually and physically) in multi-modal environments.

Perhaps the most significant impact on task efficiency can be seen in the effects of training. Pairs become more efficient with time, which may be an artifact of understanding the task and the establishment of common ground. Paralleling this theme – though significantly harder to quantify - we observed that guides gave better references over time which were more cleanly interpreted by the builder than in earlier trials; this could be seen in the development of a shared mental coordinate system. Part of this may be attributed to becoming familiar with interacting within the environment (including the tracker), but we believe that a vital part of common ground is in understanding how collaborators give and receive references. Overall, our observations suggest that there is a social component to referencing and that skill in making references improves as users become familiar with the environment.

The study also clarifies the role of virtual referencing techniques: they are important in establishing references when other forms of referencing (e.g. referential chaining) are not sufficient. Their importance was noted in generating initial reference points; once this point is established, relative references have meaning and are a more efficient (or at least preferred) method of referencing. Subjects claimed that the *video+arrow+grid* configuration gave the best support for referencing; however, the virtual artifacts became disconcerting to users when occlusion between the physical virtual was not supported, as shown by the "floating grid" effect. We also believe that the low-quality tracking negatively impacted the rating of the *video+arrow* configuration. Based on our observations, we can conclude (as suggested in Chapter 4)

that implementation issues will affect the willingness of participants to adopt the referential techniques that are provided; when tracking fails or is sporadic, participants begin to rely less on the virtual techniques, and more on other forms of referencing (such as verbal communication).

This is not to dismiss these techniques, however. Subjects found them indispensible in clarifying references when other techniques failed; we find this especially true in remote environments. From this we can conclude that virtual referencing techniques are context-sensitive and should be used sparingly and in intelligent ways. For example, placing the grid *around* the workspace (i.e. removing the lines) may produce less occlusion errors – creating a more natural environment. Further, the visibility of virtual techniques must be toggleable (by either the user or the system) to ensure that they do not occlude the view of the workspace.

As with previous studies, we argue the importance of a shared viewpoint in maintaining inter-referential awareness. In this chapter, we have discussed our architecture and ways in which referencing can be supported when multiple video streams are present. Augmenting the video before propagating it to others provides the referencing technique to be synchronized with the video frame, and can potentially save bandwidth by removing the propagation of locally tracked objects. This is the approach we used in the exploratory study, where the augmented feed was sent to the guide. However, the reference was in world coordinates, and weakens the mental mapping between the hand and arrow (for the guide) if the workspace is moved. In other words, if the builder altered their viewpoint to view from the side, axes would essentially be transposed.

In the study presented in this chapter, the local stream was tracked while augmenting the remote feed - allowing for camera-coordinate referencing into the remote environments; this can occur locally at much higher frame rates - independent of the remote video feed. In our study, a network delay would imply fast referencing over slow video for the guide, yet (oddly) interactive rates for the builder; the same applies if the *builder* were provided with a virtual pointer. While this approach is beneficial in environments where little head movement occurs, it suffers from the potential misalignment between frames and the virtual pointer.

Finally, it should be noted that, while referencing in camera coordinates, projected references have *meaning*. Even though they may technically be incorrect for those viewing the environment from an alternate viewpoint, a shared video feed provides the correct context to disambiguate this kind of reference.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

We have presented our work on inter-referential awareness in collaborative augmented reality environments; however, there is still an outstanding question: does the "ultimate referencing technique" exist? If so, we know that, at a minimum, it must be multi-modal in its ability to refer to physical and virtual artifacts in the environment, and its representation must be able to clearly infer a set of objects. Because many environments are distributed, it must function in co-located and remote space. It must allow references to be generated at various granularities and depths, and be sensitive to the context of the environment. The success of a referential technique is also influenced by environmental factors, skills of the participant, and the *task*; fortunately, concurrent support of all of these requirements is often not necessary. We ultimately believe that it is more appropriate to ask if the ultimate *set* of techniques exist; as developers, it is our responsibility to provide those that best match the factors in the environment.

It is also natural to ask that if physical techniques work in physical environments, and virtual techniques work for virtual ones, why not simply use them independently? Other than obviously burdening the user with multiple methods of referencing, we must remember that for AR to be useful, it must "*go beyond*" the capabilities found in current applications and therefore it allows for unusual situations not possible in other environments (see Chapter 1). For example, we have shown how AR can be used to refer to physical (and virtual) artifacts in remote environments. If virtual techniques are used, they must be able to reference physical content (crossing modality); purely physical approaches must be able to refer to virtual content, and may require remote

hardware (such as the shoulder-mounted laser pointer) to represent the intentions of the expert. In the medical scenario, virtual objects are embedded within physical ones. Purely virtual techniques cannot refer to physical artifacts unless their geometry is known ahead of time and require the use of the hands - which may be preoccupied during surgery. Even if physical geometry *were* known a priori, the problem is essentially *volumetric* – where any arbitrary depth is within the range of referential possibility. Because many virtual techniques rely on image plane algorithms or other forms of intersection, they would (at best) infer a volume of space in this scenario. Further, the proximity from which physical references are made is restricted, creating potential for ambiguity; even hybrid techniques (e.g. the 3D bounding box) that define space cannot be applied here because of physical limitations. Finally, if "ideal" single-mode techniques *were* incorporated into the environment, we still must cleanly address other factors that influence references, such as occluded viewpoints. Thus, we are seemingly always presented with a scenario in which the success of the technique is sensitive to context, and where the *environment* plays an important supporting role.

Throughout this dissertation, we have demonstrated some of our approaches to address a subset of these issues, as well as a theoretical framework that describes it. We explored the properties and limitations of a common, multi-modal referencing technique: the virtual pointer. As it is critical to understand how users behave in these environments and which referential techniques and scenarios they prefer, we have also presented the results from three user studies. Based on our findings, we believe that supporting inter-referential awareness comes at a cost, which is described in Section 8.1. Later, we proffer a set of design principles, and discuss future work.

8.1. The Cost of Unambiguous Referencing

There is an overarching theme to this research: when the probability of referential ambiguity is high, there are additional costs in supporting references; this probability is related to several factors. Of fundamental importance are the channels of communication that are made available within the environment. Multiple channels provide support for *multiple representations* (such as deictic speech, gesturing and shared viewpoints) that ultimately strengthen the reference; when one or more of these channels are absent, referencing can suffer. Several environmental factors influence this probability as well, including spatial configuration, the amount of discernable difference in objects, the presence of occlusion, embedded reference points, depth cues (such as stereoscopic views or shadows) and implementation issues (such as tracking). The properties of the referential technique must be considered, including its appropriateness for the task (e.g. set selection or arbitrary depths), how well it infers a set of objects, and the skill of the participants in both giving and interpreting the references (i.e. ease of use). In addition, we have presented several contextual factors, which are listed in the object-actor relationship box of the framework.

The costs of lessening the effects of these manifest in the form of additional computational and hardware resources, time or less efficiency. To overcome occluded viewpoints, we leveraged from modern graphics hardware to include shaders in the environment. In raising the awareness of the expert (e.g. familiarize them with the remote environment), we supported shared viewpoints and VoIP, requiring significantly more network bandwidth and resources than when these options were not present; while minimal, additional bandwidth is also required to disseminate the pose of a 3D cursor (in

the case of the virtual arrow, ~1920Bps - with many of these being "tinygrams"[29]). In Chapter 6, additional time was taken by several subjects to generate more accurate references (by repeated poking) - at the cost of efficiency. The initial cost of "training" between pairs in Chapter 7 yielded a significant increase in efficiency - as participants better understood the task and established common ground (such as a shared mental coordinate system for referencing); as a result, we observed that guides became more comfortable with the environment (specifically the tracking technology), and that pairs gave and interpreted references more efficiently. While the multi-modal skew technique is able to reference at arbitrary depths, it requires user training and may be less efficient than other techniques in other contexts. Further, alternative referencing methods can require specialized hardware, which can be costly or not widely available.

These costs are most pronounced in remote scenarios, where mediated communication channels are provided to synthesize co-located collaboration. Co-located participants benefit from sharing the same environment (and thus share more common ground), which is often simulated for remote participants through shared audio and video. Our pre-pilot studies suggest, as does the literature, that if audio-only conditions are provided in remote construction tasks, considerably more time is needed to complete them. It would seem obvious then to include shared viewpoints for such tasks. However, if designers opt for this support, they must choose between mono- and stereoscopic streams. While stereoscopy can potentially double the required bandwidth, *not* including it comes at the cost of depth cues - which according to our findings, can result in potentially ambiguous referencing. In raising awareness in larger groups, multiple video

[29] Tinygrams are packets of information where the header information drastically outweighs the payload.

streams (from multiple participants) require additional processing time if they are displayed simultaneously.

Fortunately, when the probability of referential ambiguity decreases, these forms of support may be relaxed; for example, when enough discernable information is available within the environment, participants can rely on simpler referencing techniques, such as the virtual arrow. Further, alternative techniques can be a viable option when more-costly ones cannot be supported; instead of providing a shared stereoscopic view of a remote environment, skilled participants can often leverage from other depth cues (e.g. virtual shadows and multiple viewpoints) at significantly less cost.

8.2. Summary of Design Principles

The techniques prescribed here are based on user observations, feedback, performance, as well as our development experience. Many of these refer to the way the environment can better support referencing. They are summarized below in hopes of providing guidelines for developers of collaborative AR systems.

Support Appropriate Referencing Techniques

Because AR environments mix real and virtual content, the referencing techniques that are provided should be multi-modal. During our user studies, we observed that it is very common for co-located participants to refer to physical and virtual artifacts through physical *and* virtual pointing – both of which are similar in properties. As suggested in post-interviews however, when multiple methods of referencing are present, they should refer the same content (e.g. the virtual arrow and physical finger should line up). Otherwise, the receiver must determine through context which technique is currently being used, which can result in ambiguity.

Further, the referencing technique(s) that are provided must match the task. In scientific visualizations, participants must have methods of simultaneously referencing multiple objects – or even empty space – often when no discernable features exist in the data. In our research, we have implemented a dynamic bounding box which dually serves as a reference and selection tool. When combined with shaders, this technique is useful in un-occluding the area of interest; it is suggested that occluding objects be de-emphasized while emphasizing referenced objects. Further, we have examined a hands-free technique for referring to multi-modal content at arbitrary depths. While this may be useful in medical and expert/technician scenarios, it is obviously not an appropriate choice for molecular modeling environments.

Provide Multiple Channels of Communication

The behavior found in collaborative AR parallels that of CSCW, and requires basic support for communication. Participants in our studies made heavy use of deictic speech and consequently, an audio channel must be provided to support them in remote scenarios. Because references are usually comprised of multiple representations, such as gesturing combined with speech, multiple channels should be included to support this. Examples include a channel for a 3D cursor, or shared visual channels. The benefits of shared video channels are described later.

Provide a Reliable Implementation

Our studies show that when the implementation does not provide reliable support, subjects are reluctant to adopt referential techniques and "fall back on" more stable mediums such as verbal instruction. In our implementation, the visual tracker was susceptible to the lighting conditions of the environment. Further, given the limited field

of view of the camera, references could be made only within a restricted area. When using the virtual arrow in remote scenarios, guides often needed the ability to place the reference *just* outside of the view frustum, causing the fiducial to be clipped, and the tracker to fail. In co-located scenarios, the act of referencing often occluded the view of the markers for one participant, again causing the tracking system to fail. Thus, it is recommended that a wide, accurately tracked workspace be provided.

Virtual Referencing Techniques must be Context-Sensitive

Our studies have clarified the role of virtual referencing techniques in collaborative spaces. Based on observation, virtual techniques are especially important in establishing an initial point of reference. When the environment is devoid of discernable features (such as during scientific visualization), virtual reference points can be provided to supplement the environment. We believe them to be of more use when relative references become difficult to generate, such as when objects are disjoint. However, their role is dependent on the task, must be sensitive to context and often relies on social protocols to function correctly. While at the beginning of the task the arrow and grid were deemed useful by subjects, once physical blocks were placed, relative references were sufficient and virtual support was no longer needed; the virtual objects ultimately cluttered the task space. Further, the "floating grid" problem created confusion in the workspace- demonstrating the importance of occlusive cues between virtual and physical artifacts. It is recommended that the visibility of these objects be "toggleable" – perhaps "timing out" after periods of inactivity. Further, if physical/virtual occlusion is not present, virtual reference points should lie outside the task space. Socially, our pilot study revealed that until the guide receives an

acknowledgement from the builder that the reference is understood, it is held in place - occasionally occluding the view of the builder.

Support Referencing in Camera and World Coordinate Systems

When in participants are working independently within their local environment, it is natural to refer to objects in *world coordinates*; by changing viewpoints or moving the workspace, the reference remains relative to the object to which it was referring. However, when sharing viewpoints, references should be made in *camera coordinates*, allowing collaborators to share an exact view of the workspace. Camera coordinate referencing is also one approach to overcoming projected references. Because the references are spatial, a hybrid variation of this allows camera coordinates to be used during *reference creation*, but once the reference is made, returns to world coordinates – allowing the receiver to view the reference from multiple angles.

Provide Depth Cues

Several subjects from our studies had difficulties referring to content when depth cues were not present - generating *projected references*; we observed these behaviors from the guide in the pilot study and the pointing task in Chapter 6. While at first this may appear applicable only to virtual pointers, this may have an impact on co-located collaboration, where the probability of physically pointing (with a finger) is high. We discovered that the inclusion of virtual shadows provides important depth cues, which significantly increased accuracy in pointing tasks. Thus, when stereoscopy is not supported, it is recommended that additional depth cues be included, such as shadows, haze or everyday objects whose size is known.

If Virtual Arrows are Used…

An entire chapter of this dissertation was devoted to understanding the referential properties of a 3D virtual pointer. This flexible technique is common in remote scenarios, is multi-modal, and simple to implement. If a virtual arrow is supported, we found that participants make more accurate references when the arrow is generally parallel to their view vector, and becomes increasingly less accurate when this angle increases. Thus, the orientation of the arrow relative to the control to which it is bound should be flexible. Further, participants are more efficient (and generally prefer) to *interpret* references in parallel, creating a physically impossible requirement with regard to the reference initiator; this can be overcome through shared viewpoints. We also discovered that environmental factors, such as spatial configuration of objects and arrow-object distance, influence the interpretation of this technique. We have evidence that arrows have a "*cone of inference*" which can be ambiguous when multiple objects fall within it. Therefore, in scenarios where it is difficult or impossible to place the arrow close to the object of reference (e.g. embedded virtual objects), a distance-tolerant referencing method - such as shared video with parallel references or alternative technique – should be considered.

Support Shared Video in All Scenarios

The positive impact of shared video on referencing cannot be over-emphasized. As described by the CSCW literature, it helps in establishing common ground more efficiently. Based on our observations, its flexibility helps in clarifying references in a variety of ways. First, shared viewpoints alleviate projected references – as these kinds of references cannot be made; by providing context, projected references have meaning.

If arrows are supported and approximately parallel to the view vector, a distance-tolerant referencing technique is available. Further, when the view of the reference initiator is shared, it can be used to overcome asymmetry from occluded viewpoints. We believe the advantages of shared viewpoints to be a cost worth incurring in every collaborative AR application. More broadly stated, we argue that *there should be no difference in the referential support provided for co-located or remote scenarios*.

8.3. Future Work

A limiting factor in the design of new techniques is that a majority of them are incapable of functioning across modality. Realistically, this is an artifact of our current technology, and will likely be overcome. The emergence of cutting-edge technology, such as the "depth camera", allows a system to acquire the projected depth of physical objects in the environment - similar in nature to the *z*-buffer in graphics hardware. This technology does not solve all of the problems we have addressed, but allows us to re-examine the potential of using virtual techniques. Further, knowledge of the physical environment can allow for natural occlusion between physical and virtual objects, which as our studies suggest, is important when embedding virtual reference points as well as in giving references. This knowledge can be beneficial to remote collaboration as well – providing the remote participants with depth cues when remote stereoscopic views are not provided. Further, this information may be transmitted on a frame-by-frame basis, allowing the remote environment to be locally re-constructed for the expert in order to generate references in world coordinates.

As described by our framework, a reference is comprised of selection and representation. First, just as references are often supported by deictic speech, we believe

that new, multi-modal representations should be investigated - perhaps supplemented using other sensory representations such as auditory cues. These may manifest as a device that functions in dual modes, such as a physical laser pointer that projects a virtual ray; while combined into one device, toggling would be required to eliminate referencing two points. Second, we believe that representation is a much underserved area, as ultimately this is what draws the attention of the user. Given that AR can visualize objects that are occluded by physical barriers (such as walls) and can augment physical objects with virtual information, we are provided with an ample toolset for designing powerful representations. We are also interested in refining our theoretical framework. Upon re-examination, we believe more emphasis should be placed on *task* - promoting it as a factor.

Our work has focused on dyads, but little work has been performed in large group interaction. This is understandable, as it is cost-prohibitive to obtain the equipment necessary to support such research, and challenging to identify a willing population of participants. It is easily imagined that if such spaces did exist, the environment could quickly become cluttered, and identifying remote participants could be difficult. Further, each participant brings with him a set of complex relationships and environmental factors (found in the framework of Chapter 4) – compounding the problem. As hardware becomes more available and AR systems become commonplace, methods that support referencing in these environments should be investigated.

We have seen the importance of environmental support in referencing, and must continue to investigate its role. In our implementation, we provided a 2D grid for a 3D task, and thus the support was likely less than ideal; we believe that better approaches

exist. We are also interested in studying the tradeoff between cost and referential support, and alternative ways of supplementing the environment when "ideal" options are technologically unfeasible.

We believe that to establish credibility of the skew pair technique, user studies must be performed. Specifically, we need to study its effectiveness in referencing volumetric space, and compare its efficiency (in both accuracy and time) to a baseline technique (e.g. raycasting). We must also implement the remote scenario to better understand its limitations. These studies will require the use of a more powerful tracker, such as the IS-900 used in section 3.4. Further, there are several "theme-and-variations" to some of our approaches. The skew-line technique can be modified to *automatically* cast a series of rays - once a pre-defined difference in position has been reached. If the variables *s* and *t* are negative, the 3D point can remain in space, or can be attached to the user's viewpoint to provide a fixed-length ray. A variation of our remote pointer allows the expert to control two virtual hands in the remote environment - essentially providing the technician with a second pair of hands.

This dissertation is presented in hopes of providing the community with a better understanding of inter-referential awareness in collaborative augmented reality, and has opened several directions for future research. Ultimately, however, our understanding of this field will require more creative thought, more user studies, and a willing population of scientists to perform the research.

BIBLIOGRAPHY

[1] M. Billinghurst, S. Weghorst, and T. Furness III, "Wearable Computers for Three Dimensional CSCW," Proc. of *International Symposium on Wearable Computers (ISWC)*, Cambridge, MA, 1997.

[2] M. Bajura, H. Fuchs, and R. Ohbuchi, "Merging Virtual Objects with the Real World," *Computer Graphics*, vol. 26, pp. 203-210, 1992.

[3] J. P. Mellor, "Enhanced Reality Visualization in a Surgical Environment, A.I. Technical Report No. 1544," 1995.

[4] W. E. L. Grimson, G. J. Ettinger, S. J. White, and T. Lozano-Perez, "An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 129-140, 1996.

[5] C. R. Johnson, R. Moorehead, T. Munzer, H. Pfister, P. Rheingans, and S. Yoo, *NIH-NSF Visualization Research Challenges Report*: IEEE Press, 2006.

[6] A. van Dam, A. S. Forsberg, D. H. Laidlaw, J. J. LaVoila, and R. M. Simpson, "Immersive VR for scientific visualization: a progress report," *IEEE Computer Graphics and Applications*, vol. 20, pp. 26-52, 2000.

[7] N. Akkiraju, H. Edelsbrunner, P. Fu, and J. Qian, "Viewing geometric protein structures from inside a CAVE," *IEEE Computer Graphics and Applications*, vol. 16, p. 58, 1996.

[8] Z. Ai and T. Frohlich, "Molecular Dynamics Simulation in Virtual Environments," *Computer Graphics Forum*, vol. 17, pp. 267-273, 1998.

[9] M. Nelson, W. Humphrey, W. Gursoy, A. Dalke, L. Kalé, R. D. Skeel, and K. Schulten, "NAMD -- A parallel, object-oriented molecular dynamics program," *International Journal of Supercomputer Applications and High Performance Computing*, vol. 10, 1996.

[10] C. Cruz-Niera, R. Langley, and P. A. Bash, "VIBE: A virtual biomolecular environment for interactive molecular modeling," *Computers and Chemistry*, vol. 20, p. 469, 1996.

[11] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations," *Journal of Computational Chemistry*, vol. 4, pp. 187-217, 1983.

[12] S. Su and R. B. Loftin, "A Shared Virtual Environment for Exploring and Designing Molecules," *Communications of the ACM*, vol. 44, pp. 57-58, 2001.

[13] N. Elmqvist, "BalloonProbe: reducing occlusion in 3D using interactive space distortion," Proc. of *Virtual Reality Software and Technology (VRST)*, Monterey, CA, 2005.

[14] R. E. Kraut, Miller, M., Siegel, J., "Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication," in *Computer-Supported Cooperative Work (CSCW)* Cambridge, MA, 1996.

[15] M. Bauer, G. Kortuem, and Z. Segall, ""Where are you pointing at?" A Study of Remote Collaboration in a Wearable Videoconferencing System," Proc. of *International Symposium on Wearable Computing (ISWC)*, 1999.

[16] H. Slay, M. Phillips, R. Vernik, and B. Thomas, "Interaction Modes for Augmented Reality Visualization," Proc. of *Australian Symposium on Information Visualization*, Sydney, Australia, 2001.

[17] M. Billinghurst, Bee, S., Bowskill, J., Kato, H., "Asymmetries in Collaborative Wearable Interfaces," Proc. of *The Third International Symposium on Wearable Computers. Digest of Papers*, 1999, pp. 133-140.

[18] M. Agrawala, Beers, A., Fröhlich, B., Hanrahan, P., McDowall, I., Bolas, M. , "The Two-User Responsive Workbench: Support for Collaboration Through Individual Views of a Shared Space," Proc. of *ACM Special Interest Group on Graphics and Interaction (SIGGRAPH)*, 1997.

[19] P. Milgram, Kishino, F., "A Taxonomy of Mixed Reality Visual Displays," *IEICE Trans. Information Systems*, vol. E77-D, No 12, December 1994.

[20] I. E. Sutherland, "The Ultimate Display," IFIP Congress 1965.

[21] W. Broll, E. Meier, and T. Schardt, "The Virtual Round Table: A Collaborative Augmented Multi-User Environment," Proc. of *Collaborative Virtual Environments (CVE)*, 2000.

[22] J. R. Vallino, "Introduction to Augmented Reality," www.se.rit.edu/~jrv/research/ar/introduction.html, 1998.

[23] R. T. Azuma, "A Survey of Augmented Reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, pp. 355-385, 1997.

[24] J. Hollen and S. Stornetta, "Beyond being there," Proc. of *Conference On Human Factor in Computing Systems (SIGCHI)*, 1992, pp. 119-125.

[25] H. Kato and M. Billinghurst, "ARToolKit, Technical Report," Hiroshima City University 1999.

[26] K. N. Kutulakos and J. R. Vallino, "Calibration-Free Augmented Reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 4, pp. 1-20, 1998.

[27] T. Höllerer, Feiner, S., Terauchi, T., Rashid, G., Hallaway, D., "Exploring MARS: Developing Indoor and Outdoor User Interfaces to a Mobile Augmented Reality System," *Computers and Graphics*, vol. 23(6), pp. 779--785, 1999.

[28] K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, and E. W. Looney, "Twiddler Typing: One-Handed Chording Text Entry for Mobile Phones," Proc. of *Conference on Human Factors in Computing Systems (SIGCHI)*, Vienna, Austria, 2004.

[29] S. Weghorst, "Augmented Tangible Molecular Models," Proc. of *International Conference on Artificial Reality and Telexistence*, Tokyo, Japan, 2003.

[30] I. Poupyrev, D. Tan, M. Billinghurst, H. Kato, H. Regenbrecht, and N. Tetsutani, "Tiles: A Mixed Reality Authoring Interface," Proc. of *InterACT*, 2001.

[31] J. Rekimoto, "Transvision:A Hand-Held Augmented Reality System For Collaborative Design," Proc. of *Virtual Systems and Multimedia (VSMM)*, 1996.

[32] H. Ishii and B. Ullmer, "Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms," Proc. of *Conference On Human Factors in Computing Systems (SIGCHI)*, Atlanta, GA, 1997.

[33] B. Ullmer, H. Ishii, and D. Glas, "mediaBlocks: Physical Containers, Transports, and Controls for Online Media," Proc. of *ACM Special Interest Group on Graphics and Interaction Techniques (SIGGRAPH)*, 1998.

[34] R. Raskar, G. Welch, K. L. Low, and Bandyopadhyay, "Shader Lamps: Animating Real Objects With Image-Based Illumination," Proc. of *Rendering Techniques*, 2001.

[35] H. Hua, L. D. Brown, and C. Gao, "A new collaborative infrastructure: SCAPE," Proc. of *IEEE Virtual Reality (VR)*, 2003.

[36] E. T. Hall, *The Silent Language*: A Fawcett Premier Book, 1959.

[37] Mehrabian, *Silent Messages*: Wadsworth Publishing Company, 1971.

[38] A. J. Sellen, "Remote Conversations: The Effects of Mediating Talk With Technology," *Human-Computer Interaction*, vol. 10, pp. 401-444, 1995.

[39] T. Robertson, "Building Bridges: Negotiating the Gap Between Work Practice and Technology Design," *International Journal of Human-Computer Studies*, pp. 121-146, 2000.

[40] G. Bafoutsou, Mentaz, G., "Review and functional classification of collaborative systems," *International Journal of Information Management*, vol. 22, pp. 281-305, 2002.

[41] J. Hindmarsh, Heath, C., "Embodied reference: A study of deixis in workplace interaction," *Journal of Pragmatics*, vol. 32, pp. 1855-1878, 2000.

[42] C. Gutwin, Greenberg, S., "A Descriptive Framework of Workspace Awareness for Real-Time Groupware," in *Computer-Supported Cooperative Work*. vol. 11 (3-4): Kluwer Academic Publishers, 2002.

[43] H. H. Clark, Brennan, S. E., "Grounding in Communication," in *Perspectives on Socially Shared Cognition*, L. B. Resnick, Levine, J., Teasley, S. D., Ed. Washington, DC: American Psychological Association, 1991, pp. 17-149.

[44] P. Dourish, "Awareness and Coordination in Shared Workspaces," Proc. of *Computer Supported Collaborative Work*, 1992.

[45] V. Bellotti and S. Bly, "Walking away from the desktop computer: Distributed collaboration and mobility in a product design team.," Proc. of *Conference on Computer Supported Cooperative Work (CSCW)*, Cambridge, MA., 1996, pp. 209 - 218.

[46] K. Schmidt and L. Bannon, "Taking CSCW Seriously," Proc. of *Computer Supported Collaborative Work (CSCW)*, 1992.

[47] C. A. Ellis, S. J. Gibbs, and G. L. Rein, "Groupware: Some Issues and Experiences," *Communications of the ACM*, vol. 34, pp. 39-58, 1991.

[48] M. Stefik, D. G. Bobrow, G. Foster, S. Lanning, and D. Tartar, "WYSIWIS revised: Early experiences with multiuser interfaces," *ACM Transactions of Office Information Systems*, vol. 5, pp. 147-186, 1987.

[49] S. Whittaker and O. O'Connaill, "The Role of Vision in Face-to-Face and Mediated Communication," Proc. of *Video-Mediated Communication*, 1997.

[50] S. R. Fussell, Setlock, L.D., Kraut, R.E, "Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks," in *Conference on Human Factors in Computing Systems (SIGCHI)* Ft. Lauderdale, FL, 2003.

[51] W. Gaver, "The affordances of media spaces for collaboration," Proc. of *Computer Supported Cooperative Work (CSCW)*, 1992.

[52] A. Ranjan, Birnholtz, J.P., Balakrishnan, R., "An Exploratory Analysis of Partner Action and Camera Control in a Video-Mediated Collaborative Task," in *Computer Supported Collaborative Work (CSCW)* Banff, Alberta, Canada, 2006.

[53] J. Ou, Fussell, S. R., Chen, X., Setlock, L.D., Yang, J., "Gestural Communication over Video Stream: Supporting Multimodal Interaction for Remote Collaborative Physical Tasks," in *International Conference on Multimodal Interfaces* Vancouver, British Columbia, Canada, 2003.

[54] L. Cheng and J. Robinson, "Dealing with Speed and Robustness Issues for Video-Based Registration on a Wearable Computing Platform," Proc. of *International Symposium on Wearable Computing (ISWC)*, 1998.

[55] N. Sakata, T. Kurata, T. Kato, M. Kourogi, and H. Kuzuoka, "WACL: Supporting Telecommunications Using Wearable Active Camera with Laser Pointer," Proc. of *International Symposium on Wearable Computing (ISWC)*, 2003.

[56] S. Benford, J. Bowers, L. Fahlén, and C. Greenlaugh, "Managing Mutual Awareness in Collaborative Virtual Environments," Proc. of *Virtual Reality Software Technology (VRST)*, Singapore, 1994.

[57] K. M. Curry, "Supporting Collaborative Awareness in Tele-immersion." vol. Master of Computer Science Blacksburg, VA: Virginia Polytechnic Institute and State University, 1999.

[58] J. Hindmarsh, Fraser, M., Heath, C., Benford, S., Greenlaugh, C., "Object-Focused Interaction in Collaborative Virtual Environments," *ACM Transactions on Computer-Human Interaction*, vol. 7, pp. 477-509, December, 2000 2001.

[59] D. A. Bowman, D. B. Johnson, and L. F. Hodges, "Testbed evaluation of virtual environment interaction techniques," Proc. of *Virtual Reality Software Technology (VRST)*, London, UK, 2001.

[60] D. A. Bowman, J. L. Gabbard, and D. Hix, "A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison Methods," *Presence: Teleoperators and Virtual Environments*, 2002.

[61] D. Bowman, E. Kruijff, J. LaVoila, and I. Poupyrev, *3D User Interfaces: Theory and Practice*: Addison-Wesley Professional, 2004.

[62] I. Poupyrev, T. Ichikawa, S. Weghorst, and M. Billinghurst, "Egocentric Object Manipulation in Virtual Environments: Empirical Evaluation of Interaction Techniques," *Computer Graphics Forum*, vol. 17, 1998.

[63] I. Poupyrev, M. Billinghurst, S. Weghorst, and T. Ichikawa, "The Go-Go Technique: Non-Linear Mapping for Direct Manipulation in VR," Proc. of *User Interface Software and Technology (UIST)*, 1996.

[64] M. R. Mine, "Working in a Virtual World: Interaction Techniques Used in the Chapel Hill Immersive Modeling Program," University of North Carolina Technical Report TR96-029 1996.

[65] A. Olwal, Feiner, S., "The Flexible Pointer: An Interaction Technique for Augmented and Virtual Reality," Proc. of *ACM Symposium on User Interface Software and Technology (UIST)*, Vancouver, B.C., 2003, pp. 83-84.

[66] J. Liang and M. Green, "Geometric modeling using six degrees of freedom input devices," Proc. of *Conference on CAD and Computer Graphics*, Beijing, China, 1993, pp. 217-222.

[67] A. S. Forsberg, K. Herndon, and R. Zeleznik, "Aperture Based Selection for Immersive Virtual Environments," Proc. of *User Interface Software Technology (UIST)*, 1996.

[68] J. S. Pierce, A. S. Forsberg, M. Conway, S. Hong, R. Zeleznik, and M. R. Mine, "Image plane interaction techniques in 3D immersive environments," Proc. of *Symposium on Interactive 3D Graphics*, 1997.

[69] R. Stoakley, M. Conway, and R. Pausch, "Virtual Reality on a WIM: Interactive Worlds in Miniature," Proc. of *ACM Conference on Human Factors in Computing Systems (SIGCHI)*, 1995.

[70] S. Zhai, W. Buxton, and P. Milgram, "The Silk Cursor: investigating transparency for 3D target acquisition," Proc. of *Conference on Human Factors in Computing Systems (SIGCHI)*, 1994.

[71] M. Billinghurst, Weghorst, S., Furness, T., "Shared space: An augmented reality interface for computer supported collaborative work," Proc. of *Collaborative Virtual Environments Workshop*, Nottingham, Great Britain, 1996.

[72] D. Schmalsteig, A. Fuhrmann, Z. Szalavari, and M. Gervautz, "Studierstube - An Environment for Collaboration in Augmented Reality," Proc. of *Collaborative Virtual Environments (CVE) Workshop*, Nottingham, Great Britain, 1996.

[73] M. Billinghurst, Kato, H., "Collaborative Augmented Reality," *Communications of the ACM*, vol. 45, pp. 64-70, 2002.

[74] P. Renevier, Nigay, L., "Mobile collaborative augmented reality: the augmented stroll," Proc. of *Engineering for Human-Computer Interaction (EHCI)*, 2001.

[75] H. Kuzuoka, "Spatial Workspace Collaboration: A SharedView Video Support System for Remote Collaboration Capability," Proc. of *Conference on Human Factors in Computing Systems (SIGCHI)*, 1992.

[76] H. Ishii, M. Kobayashi, and K. Arita, "Iterative Design of Seamless Collaboration Media," *Communications of the ACM*, vol. 37, pp. 83-97, 1994.

[77] K. B. Kiyokawa, M. , S. E. Hayes, A. Gupta, Y. Sannohe, and H. Kato, "Communication Behaviors of Co-located Users in Collaborative AR Interfaces," Proc. of *International Symposium of Mixed and Augmented Reality (ISMAR)*, 2002.

[78] K. Kiyokawa, Iwasa, H. Takemura, H., Yokoya, N., "Collaborative immersive workspace through a shared augmented environment.," Proc. of *The Society for Photo-Optical Instrumentation Engineers (SPIE)*, 1998, pp. 2--13.

[79] M. Billinghurst, J. Bowskill, M. Jessop, and J. Morphett, "A Wearable Spatial Conferencing Space," Proc. of *International Symposium on Wearable Computing (ISWC)*, 1998.

[80] C. Cruz-Niera, D. J. Sandin, and T. A. DeFanti, "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE.," Proc. of *ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH)*, 1993.

[81] I. Poupyrev, R. Berry, J. Kurumisawa, and K. Nakao, "Augmented Groove: Collaborative Jamming in Augmented Reality," Proc. of *ACM Special Interest Group on Graphics and Interaction Techniques (SIGGRAPH) Conference Abstracts and Applications*, 2000.

[82] M. Billinghurst, Kato, H., "Real World Teleconferencing," Proc. of *Conference on Human Factors in Computing Systems. SESSION: Late-breaking results: novel collaborative paradigms.*, 1999.

[83] S. Prince, A. D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. Kato, "3-D Live: Real Time Interaction for Mixed Reality," Proc. of *Computer Supported Collaborative Work (CSCW)*, New Orleans, LA, 2002.

[84] X. W. Zhong, P. Boulanger, and N. D. Georganas, "Collaborative Augmented Reality: A Prototype for Industrial Training," Proc. of, 2001.

[85] T. B. Moeslund, Storring, M., Broll, W., Aish, F., Liu, Y., Granum, E., "The ARTHUR System: An Augmented Round Table," *Journal of Virtual Reality and Broadcasting*, vol. 34, 2003.

[86] A. Butz, Hollerer, T., Feiner, S., MacIntyre, B., Beshers, C., "Enveloping Computers and Users in a Collaborative 3D Augmented Reality," Proc. of *2nd IEEE and ACM International Workshop on Augmented Reality (IWAR)*, 1999.

[87] H. Regenbrecht, Wagner, M., "Interaction in a Collaborative Augmented Reality Environment," Proc. of *Conference on Human Factors in Computing Systems (SIGCHI)*, 2002.

[88] J. Rekimoto and Y. Ayatsuka, "CyberCode: designing augmented reality environments with visual tags," Proc. of *Designing Augmented Reality Environments (DARE)*, Elsinore, Denmark, 2000.

[89] S. Feiner, B. MacIntyre, T. Höllerer, and A. Webster, "A Touring Machine: Prototyping 3D Mobile Augmented Reality System for Exploring the Urban

Environment," Proc. of *International Symposium on Wearable Computers (ISWC)*, 1997, p. 74.

[90] G. Reitmayr, Schmalstieg, D., "Collaborative Augmented Reality for Outdoor Navigation and Information Browsing," Proc. of *Symposium Location Based Services and TeleCartography*, 2004.

[91] A. Stafford, Piekarski, W., Thomas, B. H., "Implementation of God-like Interaction Techniques for Supporting Collaboration Between Outdoor AR and Indoor Tabletop Users," Proc. of *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2006.

[92] R. Wichert, "Collaborative Gaming in a Mobile Augmented Reality Environment," Proc. of *EUROGRAPHICS - Ibero-American Symposium in Computer Graphics*, Guimares, Portugal, 2002.

[93] W. Friedrich, "ARVIKA - Augmented Reality for Development, Production and Service," Proc. of *International Status Conference HCI*, 2001.

[94] B. Brown, I. MacColl, M. Chalmers, A. Galani, C. Randall, and A. Steed, "Lessons From The Lighthouse: Collaboration In a Shared Mixed Reality System," Proc. of *Conference on Human Factors in Computing Systems (SIGCHI): People at Leisure: Social Mixed Reality*, 2003.

[95] B. Bell and S. Feiner, "Augmented Reality for Collaborative Exploration of Unfamiliar Environments," Proc. of *NSF Lake Tahoe Workshop on Collaborative Virtual Reality and Visualization (CVRV)*, Lake Tahoe, NV, 2003.

[96] B. H. Thomas and W. Piekarski, "Glove Based User Interaction Techniques for Augmented Reality in an Outdoor Environment," *Virtual Reality: Research, Development and Applications*, vol. 6, 2002.

[97] P. Renevier, L. Nigay, J. Bouchet, and L. Pasqualetti, "Generic Interaction Techniques for Mobile Collaborative Mixed Systems," Proc. of *Computer-Aided Design of User Interfaces (CADUI)*, 2004.

[98] A. Fuhrmann, Schmalstieg, D., "Multi-Context Augmented Reality," Institute of Computer Graphics, Vienna University of Technology (Technical Report) 1999.

[99] A. Butz, C. Beshers, and S. Feiner, "Of vampire mirrors and privacy lamps: privacy management in multi-user augmented environments," Proc. of *User Interface Software and Technology (UIST)*, 1998.

[100] A. Fuhrmann, Hesina, G., Faure, F., Gervautz, M., "Occlusion in Collaborative Augmented Environments," *IEEE Computers and Graphics*, vol. 23, pp. 809-819, 1999.

[101] C. Furmanski, R. T. Azuma, and M. Daily, "Augmented-reality visualizations guided by cognition: Perceptual heuristics for combining visible and obscured information," Proc. of *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2002.

[102] D. Schmalsteig, Fuhrmann, A., Hessina, G., Szalavari, Z., Encarnacao, L. M., Gervautz, M., Purgathofer, W., "The Studierstube Augmented Reality Project," *Presence*, vol. II, pp. 33-54, February, 2002 2002.

[103] D. Schmalstieg, Fuhrmann, A., Hesina G., "Bridging Multiple User Interface Dimensions with Augmented Reality," Proc. of *3rd International Symposium on Augmented Reality (ISAR 2000)*, Munich, Germany., 2000.

[104] H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto, and K. Tachibana, "Virtual Object Manipulation on a Table-Top AR Environment," Proc. of *International Symposium on Augmented Reality (ISMAR)*, Munich, Germany, 2000.

[105] E. D. Mynatt, M. Back, R. Want, and R. Frederick, "Audio-Aura: Light-Weight Audio Augmented Reality," Proc. of *User Interface Software and Technology (UIST)*, 1997.

[106] M. Anabuki, H. Kakuta, H. Yamamoto, and H. Tamura, "Welbo: An Embodied Conversational Agent Living in Mixed Reality Space," Proc. of *Conference on Human Factors in Computing Systems (CHI) Demonstrations*, 2000.

[107] E. Woods, Billinghurst, M., Looser, J., Aldridge G., Brown, D., Garrie, B., Nelles C., "Augmenting the Science Centre and Museum Experience," Proc. of *2nd International conference on Computer graphics and interactive techniques* Australasia and South East Asia, 2004.

[108] P. Wellner, "Interacting with paper on the DigitalDesk," *Communications of the ACM*, vol. 36, 1993.

[109] B. Ullmer and H. Ishii, "The metaDesk: Models and Prototypes for Tangible User Interfaces," Proc. of *User Interface Software and Technology (UIST)*, 1997.

[110] F. Biocca, A. Tang, C. Owen, and F. Xiao, "Attention funnel: omnidirectional 3D cursor for mobile augmented reality platforms.," Proc. of

Conference on Human Factors in Computing Systems (SIGCHI), Montreal, Canada, 2006.

[111] M. Tönnis, C. Sandor, G. Klinker, C. Lange, and H. Bubb, "Experimental Evaluation of an Augmented Reality Visualization for Directing a Car Driver's Attention," Proc. of *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2005.

[112] J. W. Chastine, J. C. Brooks, Y. Zhu, G. S. Owen, R. W. Harrison, and I. T. Weber, "AMMP-Vis: A Collaborative Virtual Environment for Molecular Modeling," Proc. of *ACM Symposium on Virtual Reality Software and Technology (VRST)*, Monterey, CA, 2005.

[113] J. W. Chastine, Zhu, Y., Brooks, J., Owen, G. S., Harrison, R. W., "A Collaborative Multi-View Virtual Environment for Molecular Visualization and Modeling," Proc. of *Coordinated Multiple Views (CMV)*, London, England, 2005.

[114] X. Chen, I. T. Weber, and R. W. Harrison, "Molecular Dynamics Simulations of 14 HIV Protease Mutants in Complexes with Indinavir," *Journal of Molecular Modeling*, vol. 10, pp. 373-381, 2004.

[115] R. W. Harrison, "Stiffness and Energy Conservation in Molecular Dynamics: an Improved Integrator," *Journal of Computational Chemistry*, vol. 14, pp. 1112-1122, 1993.

[116] R. W. Harrison, "Integrating Quantum and Molecular Mechanics," *Journal of Computational Chemistry*, vol. 20, pp. 1618-1633, 1999.

[117] K. M. Curry, "Supporting Collaborative Awareness in Tele-immersion," in *Master Thesis, Department of Computer Science* Blackburg, VA: Virginia Tech, 1999.

[118] J. W. Chastine, J. C. Brooks, Y. Zhu, C. Owen, R. W. Harrison, and I. T. Weber, "Emphasizing the Area of Interest Using Real-Time Shaders," Proc. of *ACM Special Interest Group on Graphics and Interactive Techniques (SIGGRAPH Poster Session)*, Los Angeles, CA, 2005.

[119] P. Haeberli, Akeley, K., "The accumulation buffer: hardware support for high-quality rendering," Proc. of *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, Dallas, TX, 1990.

[120] F. Liu, G. S. Owen, Y. Zhu, R. W. Harrison, and I. T. Weber, "Web Based Molecular Visualization using Procedural Shaders in X3D," in *Proceedings of ACM SIGGRAPH Conference Web Program* Los Angeles, California, USA: ACM Press, 2005.

[121] B. MacIntyre, M. Gandy, S. Dow, and J. D. Bolter, "DART: A Toolkit for Rapid Design Exploration of Augmented Reality Experiences," Proc. of *User Interface Software and Technology (UIST)*, Santa Fe, NM, 2004.

[122] J. W. Chastine, "A Skew-Line Referencing Technique for Collaborative Augmented Reality," Georgia State University, Technical Report 2006.

[123] J. W. Chastine, Zhu, Y., Preston, J.A, "A Framework for Inter-referential Awareness in Collaborative Environments," Proc. of *2nd IEEE International Conference on Collaborative Computing*, Atlanta, GA, 2006.

[124] J. W. Chastine, Zhu, Y., "The Cost of Supporting References in Collaborative Augmented Reality," Proc. of *First International Conference on Immersive Communications (IMMERSCOM - under review)*, Verona, Italy, 2007.

[125] J. W. Chastine, Nagel, K., Zhu, Y., Yearsovich, L., "Understanding the Design Space of Referencing in Collaborative Augmented Reality Environments," Proc. of *Graphics Interface (GI)*, Montreal, Canada, 2007.

[126] J. W. Chastine, Nagel, K., Hudachek-Buswell, M. H., Zhu, Y., "Studies on the Effectiveness of Virtual Pointers in Collaborative Augmented Reality," Proc. of *International Symposium of Mixed and Augmented Realities (ISMAR - under review)*, Japan, 2007.