RC23652 (W0507-023) July 1, 2005
Computer Science

IBM Research Report

A Study on the Manipulation of 2D Objects in a Projector/Camera-Based Augmented Reality Environment

Stephen Voida
GVU Center
College of Computing
Georgia Institute of Technology
85 5th Street NW
Atlanta, GA 30332

Mark Podlaseck, Rick Kjeldsen, Claudio Pinhanez
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

A Study on the Manipulation of 2D Objects in a Projector/Camera–Based Augmented Reality Environment

Stephen Voida
GVU Center, College of Computing
Georgia Institute of Technology
85 5th Street NW
Atlanta, Georgia 30332 USA
svoida@cc.gatech.edu

Mark Podlaseck, Rick Kjeldsen, Claudio Pinhanez
IBM Research, T.J. Watson
19 Skyline Drive
Hawthorne, New York 10532 USA
{podlasec, fcmk, pinhanez}@us.ibm.com

Abstract
Are the object manipulation techniques traditionally used in head–mounted displays (HMDs) applicable to augmented reality based projection systems? This paper examines the differences between HMD– and projector/camera–based AR interfaces in the light of a manipulation task involving documents and applications projected on common office surfaces such as tables, walls, cabinets, and floor. We report a Wizard of Oz study where subjects were first asked to create gesture/voice commands to move 2D objects on those surfaces and then exposed to gestures created by the authors. Among the options, subjects could select the object to be manipulated using voice command; touching, pointing, and grabbing gesture; or a virtual mouse. The results show a strong preference for a manipulation interface based on pointing gestures using small hand movements and involving minimal body movement. Direct touching of the object was also common when the object being manipulated was within the subjects' arm reach. Based on these results, we expect that the preferred interface resembles, in many ways, the egocentric model traditionally used in AR.

Categories & Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *input devices and strategies, interaction styles*; I.3.6 [**Computer Graphics**]: Methodology and Techniques – *interaction techniques*; I.4.0 [**Image Processing and Computer Vision**]: Miscellaneous – *gesture recognition*

General Terms: Experimentation, Human Factors.

Keywords: Augmented reality, augmented workspaces, Wizard of Oz study, user–centered design.

INTRODUCTION

Techniques for the manipulation and control of virtual objects have been extensively studied in the Virtual Reality (VR) community [9, 27, 28], particularly in the context of egocentric manipulation [4, 15, 21, 31]. Most of these techniques and results can be applied directly to Augmented Reality (AR) environments based on head–mounted displays. However, it has not been studied whether these manipulation techniques are the most appropriate for the emergent projector/camera–based AR environments [19, 22, 24, 25] where imagery can only be overlaid on the surfaces of the furniture and objects in the environment.

For instance, consider the "Go–Go" technique [21], which uses a virtual hand floating in the visualization field as a surrogate for "touching" the virtual objects. To reach distant objects, the "Go–Go" technique employs a mapping of the arm movement that non–linearly amplifies the movement at the end of the arm's reach, significantly expanding the user's reach.

However, in projector/camera–based AR environments, it is impossible to render the virtual hand floating in midair. Instead, the hand would have to be visualized as moving across the projectable surfaces of the environment. Since in most practical cases of AR these surfaces are not smoothly joined or even continuously connected, the image of the virtual hand is likely to jump in an unnatural way as the arm movement progresses.

This paper reports a study that takes a user–centered design (UCD) approach to the problem of determining manipulation techniques for projector/camera–based AR environments. In this study, subjects are invited both to create their own set of manipulation techniques and to use a set defined by the authors based on classic desktop and gesture–based manipulation techniques. Through this UCD approach, we try to avoid the tendency of VR, AR, and ubiquitous computing practitioners to create new interaction paradigms that require a lot of learning from users (for example, [5]).

The experiment described in this study was performed in the context of an application that aims to use AR techniques in an office environment. In particular, our experiment asks subjects to manipulate typical computer documents and applications being projected on office surfaces such as tables, desks, walls, cabinets, and on the floor.

We start this paper by examining some key differences between projector/camera–based and head–mounted display (HMD)–based AR environments, and a brief descrip-

tion of the use of the augmented office scenario space. We then introduce the main issues we identified for the design of manipulation techniques in such situations. The main part of the paper describes the study, presents its results, and discusses the possible implications of the results on the design of projector/camera–based AR user interfaces.

PROJECTOR/CAMERA VS. HEAD–MOUNTED AR

Augmented reality has traditionally focused on systems and applications using head–mounted displays (HMD). However, a number of researchers have advocated the use of projected displays instead of HMDs, as mentioned by Azuma et al. [1] in their survey of AR techniques. For example, Rekimoto and Saitoh [25] use a projector to show the contents of videotapes and documents placed on a table; Raskar et al. [24] use a projector to change the color of the surface of objects; Lai et al. [13] use a steerable projector system to move computer applications to the wall and tables in an augmented office environment; and Pingali et al. [18] augment a store shelf with information about the products on display.

A major advantage of projector/camera–based AR is that there is no need to compensate for the movements of the user's head. Although it is still necessary to register the projector system to the environment, the problem is dramatically simplified due to the fact that the projector is fixed in space. In particular, problems with registration lag and delays are eliminated (as long as the physical projection surface is not moving).

However, projection–based systems suffer from a major drawback: graphics can only be rendered on environment and object surfaces and, due to occlusion, some of these surfaces may be unavailable for projection. Additionally, display quality is affected by the environment lighting, the surface texture and color, etc. The effect is that the display space is very discontinuous, both spatially and in terms of display quality, particularly when compared with traditional HMD–based augmentations.

In fact, spatial discontinuity happens not only because the surfaces are scattered throughout the environment but also because most of the time adjacent surfaces are not connected smoothly, but by corners. These discontinuities create a major problem when virtual objects are to be translated over the visual field. Raskar et al. [22] propose a solution in the case of adjacent wall corners, but it is easy to see that it is impossible to solve this problem in the generic case.

Most of the techniques used in AR for object manipulation and, in particular, for selection and target determination, use elements that assume visual continuity of the user's virtual visual field. Notably, both the "Go–Go" technique [21] and the "Ray–Casting" [15] methods rely heavily on "floating" objects (a hand and a light ray, respectively), that cannot be rendered satisfactorily in projector/camera systems in most environments.

THE AUGMENTED OFFICE SCENARIO

Our interest in object manipulation techniques for projector/camera–based AR environments stems from our past and current research in augmented workplaces. Previously, we employed a steerable projector system [19] to silently notify occupants of e–mail messages; to move desktop content to walls and tables in support of collaborative work; and to create dynamic, reconfigurable wallpaper [13].

Our recent focus is on using augmented reality techniques to help knowledge workers to manipulate, manage and transform information [8]. Previous studies emphasize the importance of spatial organization and visibility of unfiled information for these workers [10, 16]. Although desktop systems have been developed to meet these needs, projected user interfaces appear to hit the "sweet spot" for these kinds of workers, due to their large physical size and ability to act as persistent, peripheral displays.

A number of research projects have explored the use of large displays and projected user interfaces to support individual office work, starting with Bolt's pioneering work on "Put–That–There" [3], and more recently by others, e.g., [13, 14, 29]. Other uses of large, projected user interfaces include informal collaboration [23, 25, 26], and telepresence [23].

A typical use of projected user interfaces in the workplace is the extension of a user's desktop to the surrounding environment. Consider as an example an augmented office where typical computer applications and documents can be displayed on any surface in a room. The ability to display and interact with applications on any surface would provide the flexibility for the user to work on tables or walls, allowing greater opportunities for collaboration. The ability to store documents throughout the environment would allow the user to leverage spatial layout as an aid for organizing the extended "desktop," enable that user to cluster virtual objects near associated physical counterparts, and quickly move files in and out of the primary work area as needed.

One way to achieve this vision without requiring a large number of projectors is to have one projector able to reach the entire workspace at the same time by way of a convex mirror and a second projector able to steer a high–resolution display to a single surface in the environment at a time [19]. This would allow a few documents to be "in focus" at any one time, while many others remain easily accessible.

It is clear that in such an augmented office scenario described above, the user needs explicit control over the location and appearance of the projected 2D objects representing documents and applications. It is important, then, to study how users can effectively manipulate the location and appearance of projected objects from varying distances and on a variety of surfaces.

There are two main classes of techniques for manipulating projected objects. The first utilizes hardware devices, such

as wireless and gyroscopic mice, to move a pointer around in projected space and drag objects as if they existed in a typical desktop interface, e.g., [5, 25, 30]. The second, and the focus of our research, involves enabling computers to recognize and act on natural human interactions—speech or gesture—to accomplish manipulations of projected objects. Multimodal and gesture–based input has been studied closely in the field, e.g., [12, 17], and we believe it offers the greatest flexibility in allowing users to express their intentions without interfering with their normal work practices.

To address the research question of how users might most naturally interact with projected objects, we took a user–centered design approach (UCD) where we solicited intuitive interaction techniques from users, combined them with four interaction techniques that we designed, and performed an evaluation of the intuitiveness and usefulness of them. We begin by discussing the decisions guiding the design of our set of interaction techniques. Then, we present the results of a multi–phase user study and their implications for projected user interface design.

INTERACTION DESIGN

The question of how users should be able to move projected objects around an augmented environment is more complex than it initially appears. There are many different circumstances to consider.

Distance to the Object

The distance of the object from the user is a primary consideration. When an object is within arm's reach, it may be natural for the user to physically touch it in order to manipulate it. Such interactions are easy to detect and intuitive to use [11]. However, it is not always reasonable to expect a user to walk up to an object before they manipulate it, especially since objects can be displayed in unreachable locations.

When an object is out of reach, a common reaction may be to point at it. Unfortunately, detecting where a user is pointing from a distance is difficult to do with any accuracy. These considerations provide strong constraints on the design of gestures that manipulate remote objects. Although it is possible to support distinct types of gesture for different distances, it may be that users think more in terms of a pointing continuum where close proximity pointing blends seamlessly into distance pointing.

User's Spatial Model of Surfaces

An additional consideration is how the user envisions the space around them. We have identified at least two distinct models: *surface–oriented* and *continuous*.

Our environment is composed of distinct surfaces—table tops, walls, cabinet doors and the like—that are embedded in the larger context of the room. When moving objects about on a single surface, it is most likely natural to think of the surface as a continuous entity on which the object can be located.

However, once the user wants to move an object from one surface to another, the situation becomes less clear. Does the user think of these surfaces as distinct surfaces located within some larger context or as one continuous surface with breaks in it? In other words, how does the user interpret the visual discontinuities of projector/camera–based AR?

One can envision using an inherently within–surface method (point and drag) between surfaces by dragging objects across the gaps between them. One can also envision using an inherently between–surface interaction (grab and throw) on the same surface. Is it advantageous to have a single type of interaction for both types of movements?

Indications of Discrete Events

Another consideration is how the user indicates discrete events during a manipulation, such as the end of the object selection phase or the release of a held object. Some alternatives include using a distinct hand shape, such as a grasp; a distinct hand movement, such as a shake or pause; a separate signal, such as a head nod; or a signal from a different modality, such as voice. While different alternatives will no doubt be best suited in different circumstances, some combinations will be more intuitive to the user and more reliable for the computer to recognize than others.

User's Willingness to Move

Another consideration is the user's propensity to move within the space. If an out–of–reach surface is nearby, how inclined is the user to move to that surface to interact with it, especially given that interaction may be more accurate or reliable up close? Does this propensity change if the user is sitting or standing, or depending on the task? This issue seems to be especially relevant in the case of projector/camera–based scenarios because the virtual objects are created on surfaces that are positioned at different distances from the user.

GESTURES FOR OBJECT MANIPULATION

Given these design considerations and our observations of nine pilot participants' interactions with sample projected objects, we designed a set of four gestures for manipulating the location of a projected object.

Point/Touch and Drag

The user touches or points at the object of interest with her index finger at the end of a large–scale arm movement and pauses until the system recognizes the selection gesture. At this point, the object becomes "affixed" to her fingertip, moving along the projected surfaces in the room and following the ray extending outward from her fingertip as she moves her hand. This "dragging" mode continues until she pauses a second time and quickly retracts her hand, indicating completion of the gesture. Notice that in this interaction

Table 1. Notable characteristics of our four gestures for manipulating the location of a projected 2D object.

Gesture	Suitable for manipulating *distant objects?*	Associated spatial model of surfaces	Method of indicating discrete events	Size of required motion	Accuracy of target specification	Implementation concerns
Point/ Touch and Drag	Yes	Continuous	Pauses; retracting to signal end	Large	High to moderate, depending on distance	Requires accurate estimate of where the user is pointing their arm
Grab and Throw	Yes	Surface–oriented	Primarily hand shape	Large	Moderate	Difficult to provide feedback of perceived target location
Pantograph/ Virtual Mouse	Yes	Continuous	Hand shape	Small	Potentially High	Difficult to understand transform between the user's virtual "mouse pad" and the pointer motion
Flick	No	Surface–oriented	Hand shape and location	Small	Low	May require task–, environment–, or user–specific heuristics to determine target location

technique the user has to assume that the discrete surfaces are somehow "connected" to each other.

Grab and Throw

The user selects the object of interest in the same way as before, by pointing at the object with an extended arm. However, instead of continuously directing the object's motion, she makes a quick grabbing or clutching motion with her entire hand and the object disappears from the projection, as if captured in her grasp. She can then quickly turn to face the object's intended target location and "throw" the object by opening her hand while extending her arm in the direction of the desired throw. The system makes a gross estimate of her intended destination and causes the object to appear again at that location.

Pantograph/Virtual Mouse

The user places her open hand on a surface (horizontal or vertical) and pauses. The system then establishes a local coordinate system around the hand and projects a cursor on or near it. From this point until the user completes the gesture, the cursor moves within the global (room) coordinate system in a direction similar to the direction of the user's hand in its local coordinate system, but by a greatly magnified amount. For example, when the user's hand moves left, the pointer moves left from its current position. When the pointer reaches the object to be moved, the user makes a fist, which, in this case, affixes the object to the cursor and enables it to be dragged by subsequent hand movement. When the object reaches the destination, the user opens her hand again, releasing the object and ending the gesture.

Flick

The user places her hand next to a projected object and turns her hand over quickly as if brushing the object away in a particular direction, with a flick of the wrist. The object quickly slides away from the point of contact in the direction of the flick along the projected surface and either continues onto an adjacent surface and comes to a stop, or stops at the edge of the original surface if there is no adjacent surface upon which to continue.

Each of these gestures was chosen because it works well for some class of situations, is feasible to implement with existing technology, and is easy for a user to remember. Table 1 summarizes some of the characteristics of the different gestures.

USER STUDY

We designed a user study to determine the interaction techniques that users with moderate computer skills employ when asked to manipulate projected objects in an office–like environment augmented with projected computer documents and applications. In particular, we wanted to compare the usage of the techniques described above with interaction techniques created by the users.

To simulate the advanced computer vision system required to recognize our set of gestures, we employed a Wizard of Oz technique similar to that presented elsewhere in the HCI and virtual environment literature [6, 7]. We chose this approach because we wanted subjects to freely choose among our set of interaction techniques and those that they created themselves. However, as detailed later, we were

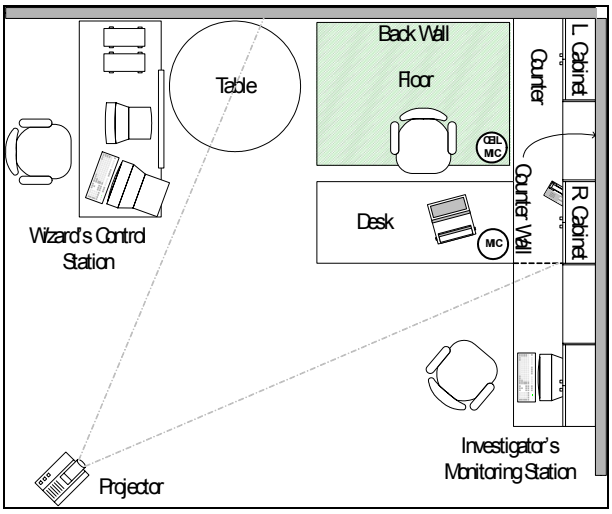

Figure 1. A schematic diagram of the experimental setup, indicating the overall room layout, the projection cone and the interaction surfaces used during the user study.

Figure 2. An overview of the experimental setup (prior to installing the "round table"). The "wizard" is seated at his control station on the left, and a researcher is playing the role of a participant on the right. Note the projected images on the wall, desk, and floor surfaces.

very careful not to disclose the role of the "wizard:" subjects were told that our system could be easily adjusted to accommodate whatever command mode they thought would be appropriate for the task.

To avoid "expert" bias towards our set of commands, we employed a "confederate" scheme where one of the experimenters joins the experiment pretending to be another subject. As detailed later, our confederate introduces the set of gestures defined in earlier as his own, therefore avoiding the bias that could be generated if we, as experimenters, had presented the gestures to the subject as our own.

Participants

We recruited a total of nine participants for our user study. Of those, 6 were female and 3 were male. All used computers to some extent in their jobs but none of their jobs were directly related to computers. The participants' ages ranged from 20 to 61.

Environment

We performed the user study in one corner of a dedicated projection lab (see Figures 1 and 2). We placed several pieces of furniture in our simulated office space to resemble a typical individual office layout. The researcher guiding the participant through the study was seated at a video recording and system monitoring system across the desk from the participant. The "wizard" sat behind the participant as a monitor, which displayed multiple live camera views of the simulated office. (This perspective, together with the "wizard's" knowledge of the participants' intentions, enabled the "wizard" to evoke convincing feedback in response to their interactions. In debriefing, 5 of 9 participants reported that they believed the system was a fully–functional implementation.) Participants were led to believe that the "wizard's" presence was necessary to provide guidance for the (fictitious) computer vision system when it encountered problems.

Eight discrete surfaces fell within the projection cone (see Figure 1). These were:

- the desk at which the participant was initially seated,
- the counter to the participant's left,
- the wall above the counter,
- two cabinets hanging above the counter (referred to as the *left* and *right* cabinets during the study),
- the wall directly behind the participant (referred to as the back wall),
- the floor surface between the participant and the back wall, and
- a round table against the back wall.

Surfaces were arranged to maximize variances in distance from one another and distance from the participant. Participants were outfitted with a lapel microphone and made aware of the location of two stationary room microphones. While we were primarily interested in eliciting a set of gestures, we also did not want to inhibit voice or multimodal interaction.

Procedure

Our experimental protocol consisted of an initial demographic questionnaire, four main user study phases, and an exit survey. No participant took longer than 75 minutes to complete the study. Most needed approximately one hour, including time spent obtaining informed consent before the study and completing a short debriefing session afterwards. Each participant was first seated at the desk and introduced to the projection cone and the surfaces it covered (as described before). Six projected objects were also presented (for example, a QuickTime movie window and a photo of

Albert Einstein) along with phrases that would be used to describe them (e.g., "QuickTime," "Einstein").

Phase 1—Exploratory Phase: In this phase of the study, the participant, seated at the desk, was asked to help "train" the system with gestures and/or voice commands that he/she would subsequently use to manipulate the position of the projected objects. During this phase, the study conductor pointed to a projected object on the desk and simultaneously indicated another area on the desk with a red laser pointer. The participant was then asked: "If you wanted this object to be located near the red laser pointer, how would you communicate that to the system?"

As the participant described a technique, the conductor would elicit step–by–step details and write them on a large sheet of paper mounted on an easel. At the same time, the "wizard" would supposedly enable the computer vision and speech recognition systems to recognize this technique. This step was repeated with different permutations of source object, source location, and target location until the participant's techniques were exhausted. Since we wanted to elicit techniques that would generalize across a wide variety of office work, for the second half of the "training" process, the participant was asked to pick up the phone and imagine an important conversation while manipulating the projected objects so as to temporarily inhibit speech and movement.

Phase 2—Sharing Phase: In this phase, a third researcher (a confederate) was introduced to the participant as a fellow participant who had already trained the system with his manipulation techniques. The confederate's list of techniques was hung next to the participant's and was comprised of our set of gestures described previously: point/touch and drag, grab and throw, pantograph/virtual mouse, and flick.

The participant and confederate were asked to take turns demonstrating and explaining their respective techniques to each other. Each practiced the other's techniques until they were able to demonstrate a reasonable fluency. The confederate then left with instructions to return after the participant finished the following phase of the study.

Phase 3—Structured tasks: For the first part of this phase, both the participant's and confederate's lists of techniques were flipped over so that neither was visible. The conductor then asked the participant to recall as many of the techniques from both lists as possible. The lists were then revealed to the participant and any techniques missed by the participant in the quiz were reviewed and practiced in order to minimize learning and recall differences among participants.

The six projected objects that were introduced at the beginning of the study were re–displayed. As in the exploratory phase, the conductor indicated a source object and pointed to a target location with a laser pointer. The participant was asked to use whichever of the confederate's or their own manipulation techniques they preferred and/or that seemed appropriate to the task at hand to manipulate the object as instructed. Thirteen spatial manipulations were accomplished in this manner.

Phase 4—Unstructured tasks: In this phase, five paintings were projected into the space and introduced to the participant. The paintings were then broken into jigsaw puzzle pieces and scattered around the space. The participant was asked to assemble as many pieces as possible in five minutes using, as before, whichever techniques they liked and/or seemed appropriate to the task at hand. The puzzle pieces moved freely in the space, just as did the projected objects in phases 1 and 3; we provided no automated position "snapping" behavior for correctly–positioned pieces.

After five minutes, the technique lists were again flipped over and the participant was asked to recall as many of the techniques that they could remember, again, to test for recall differences among participants. Finally, they were asked to complete an exit survey.

Results

Phase 1—Exploratory Phase: Most of the subjects (6/9) were able to define three manipulation techniques before they ran out of ideas. All nine subjects defined a voice, point, or touch interaction as their first technique. Voice manipulation was the most frequently defined (6/9), exclusively following a "Move Object–X to Surface–Y" syntax. Pointing manipulations were also dominant; these always involved selecting an object by pointing to it and retracting, then pointing to the target location and retracting (5/9). Touch interactions, the third most frequently defined manipulation, closely followed the pointing syntax, except that both object and surface were literally touched, rather than pointed to (4/9). Another often–suggested manipulation technique involved multimodal use of pointing and voice commands, reminiscent of the traditional "Put–That–There" paradigm [3].

Although subjects were told that the system would be looking for gestures and/or voice, three subjects defined techniques involving the study conductor's laser pointer (point it, turn it on to select the source object, turn it off, and then turn it on again to indicate the target). Table 2 summarizes the analysis of the most popular gestures proposed by the subjects of our study.

Phase 2—Sharing Phase: A number of factors influenced a subject's choice of interaction techniques during the structured task phase. Two primary factors involved distance: the subject's distance from the object to be moved (hereafter, the source location) and the subject's distance from the target location to which the object was to be moved (the target location).

Table 3 shows the different techniques used for selecting the object to be manipulated in the task, sorted by the distance from the subject to the source object. The most salient aspect is that the use of *touching* (highlighted with light grey) *decreases* and the use of *pointing* (highlighted with

Table 2. Notable characteristics of the most popular user-defined gestures for manipulating a projected 2D object.

Gesture	Suitable for manipulating *distant objects?*	Associated spatial model of surfaces	Method of indicating discrete events	Size of required motion	Accuracy of target specification	Implementation concerns
Voice	Yes	Surface-oriented	Speech	None	Low	Fast utterance of control commands may be difficult to parse in context.
Point and Retract	Yes	Surface-Oriented	Retract gesture	Large	High to moderate, depending on distance	Requires accurate estimate of where the user is pointing their arm.
Touch and Retract	No	Surface-oriented	Retract gesture	Large	High	Requires object to be close to user.
Voice and Point	Yes	Surface-oriented	Speech	Large	Low	Requires complex parsing and coordination of the modalities.
Laser Pointing and Retract	Yes	Surface-oriented	Turn on/off	Small	High	Requires cameras monitoring a huge area looking for extremely fast movements.

dark grey) *increases* as the distance between the source location and the subject was increased. When the source location was beyond one arm's length (> 2 feet), touching was no longer used at all. However, when the source location was one foot away from the subject (not under their nose, but easily within reach), pointing was preferred to touching in two out of three instances. These "local points" varied widely. In some cases, the subject's fingertip approached within millimeters of the projected object or its target. In other cases, the fingertip barely made a trajectory away from the subject's body.

Table 4 shows similar results for the techniques used to identify the target location for the manipulated object. Touching was not used when the target location was beyond an arm's reach and pointing was increasingly utilized as the target receded further from the subject's position.

In both cases, one of the subjects (B) exclusively used the laser pointer for selection of sources and to identify target locations. Another subject (C) almost exclusively used the grab and throw gestures. This seems to indicate that some users are likely to have strong preferences for specific techniques.

Table 5 shows the techniques used to select the source object and to specify its location, in the sequence that the subjects performed the experiment. Once a subject chose a technique with which to select a source object (the top line in the pairs), it was extremely likely (89%) that the subject would continue to use the same technique with which to specify a target location. The cases in which the same technique was not used are highlighted in Table 5. Notice that grab/throw pairs are parts of the same technique.

Several variables conjectured to have an influence on technique were not demonstrated to do so. These included the orientation of source and target surfaces, the distance between the source and target surfaces, and the presence of a physical gap between surfaces.

Finally, out a total of 103 complete recorded interactions, in only two instances did subjects stand up to accomplish a manipulation. In one case, an object was projected on the floor and occluded by the subject in his chair. Immediately after completing the interaction (a total of 21 seconds), the subject reseated himself. In the second case, the subject rose for 6 seconds to indicate the far side of the counter as a target location.

Phase 3—Structured tasks: Our subjects overwhelmingly preferred pointing interactions for assembling the jigsaw puzzles. With the exception of one subject who primarily used touching, *83% of all the interactions were accomplished using pointing* for selecting the source object and specifying its target location. Since most of the puzzle pieces were grouped together, the need for large movements between surfaces was minimal. When it was necessary to move a piece a long distance, however, subjects often used the throw gesture (7% of all interactions). Because the puzzle pieces were not named (or even nameable), voice manipulation almost disappeared from this section of the study. When voice was used, however, it was in conjunction with pointing and addressed to the object, rather than to the system. (For example, a subject would say, "No, no, not you. You!" to an object she was trying to select.)

Table 3. Technique used to select the object to be moved.

Source	Dist.	A	B	C	D	E	F	G	H
desk	0	Grab	Laser	Touch	Touch	Point	Voice	Voice	Touch
desk	0	Panto	Laser	Grab	Touch	Point	Touch	Voice	Touch
round tbl	2	Point	Laser	Grab	Touch	Voice	Voi-Poi	Point	Point
round tbl	2	Touch	Laser	Grab	Touch	Touch	Point	Voice	Point
counter	2	Voice	Laser	Grab	Laser	Point	Touch	Point	Voice
floor	3	Point	Laser	Grab	Laser	Point	Point	Voice	Point
floor	3	Voice	Laser	Grab	Laser	Point	Point	Voice	Point
counter	4	Voice	Laser	Grab	Point	Voice	Voice	Voice	Point
back wall	4	Point	Laser	Point	Point	Point	Point	Point	Point
back wall	4	Point	Laser	Grab	Point	Panto	Voi-Poi	Voice	Voice
back wall	5	Point	Laser	Grab	Laser	Point	Point		Point
left cab	5	Voice	Laser	Point	Point	Point	Point	Point	Voice
right cab	5	Voice	Laser	Point	Laser	Point	Point	Point	Point

Table 4. Technique used to specify the target position.

Target	Dist.	A	B	C	D	E	F	G	H
desk	0	Touch	Laser	Throw	Touch	Panto	Touch	Voice	Voice
desk	0	Touch	Laser	Throw	Touch	Point	Touch	Voice	Touch
desk	0	Voice	Laser	Throw	Laser	Point	Touch	Point	Voice
round tbl	2	Touch	Laser	Throw	Laser	Point	Point	Voice	Point
floor	2	Touch	Laser	Throw	Laser	Point	Point		Point
round tbl	2	Panto	Laser	Throw	Touch	Point	Point	Voice	Touch
counter	2	Touch	Laser	Throw	Point	Voice	Touch	Point	Touch
counter	2	Voice	Laser	Point	Laser	Point	Point	Point	Point
to right	3	Voice	Laser	Throw	Laser	Point	Point	Voice	Point
back wall	4	Voice	Laser	Throw	Point	Voice	Voice	Voice	Point
counter wall	5	Throw	Laser	Point	Point	Point	Voice	Voice	Point
right cab	5	Voice	Laser	Point	Point	Point	Point	Point	Voice
higher	6	Point	Laser	Point	Point	Point	Point	Point	Point

The most striking aspect of this phase was the odd combination of spatial *fastidiousness* and physical *inertia* demonstrated by 2/3 of the subjects. They were surprisingly willing to spend a great deal of time using coarse pointing gestures to accomplish fine object movements (spending up to 18 seconds per manipulation) so that they would not have to stand up. (All, however, fully exploited the mobility of their chair to rotate and roll around the limited area.)

Of the three subjects who did stand up during this phase, all were female. Two of these subjects used a touching technique when particularly fine accuracy was required.

Phase 4—Unstructured tasks: Five subjects acknowledged the difficulty of the interactions while sitting but stated a firm preference for not standing regardless of the impact to the time spent on a task or the quality of the work accomplished. Three subjects described themselves as "lazy" to explain their preference. Another expressed her disdain for having to accommodate herself to virtual objects: "I feel like it's not really there, so why would I have to get up and go get it?" All the subjects voiced satisfaction with the interaction techniques that they employed. Three subjects were concerned that the system was not fast enough to track them accurately.

Table 5. Projected output selection techniques and target location identification techniques.

Src/Tgt	D	A	B	C	D	E	F	G	H
desk	0	Grab	Laser	Touch	Touch	Point	Voice	Voice	Touch
count. wall	5	Throw	Laser	Point	Point	Point	Voice	Voice	Point
counter	4	Voice	Laser	Grab	Point	Voice	Voice	Voice	Point
back wall	4	Voice	Laser	Throw	Point	Voice	Voice	Voice	Point
back wall	4	Point	Laser	Point	Point	Point	Point	Point	Point
higher	6	Point	Laser	Point	Point	Point	Point	Point	Point
floor	3	Point	Laser	Grab	Laser	Point	Point	Voice	Point
round tbl	2	Touch	Laser	Throw	Laser	Point	Point	Voice	Point
back wall	5	Point	Laser	Grab	Laser	Point	Point		Point
floor	2	Touch	Laser	Throw	Laser	Point	Point		Point
left cab	5	Voice	Laser	Point	Point	Point	Point	Point	Voice
right cab	5	Voice	Laser	Point	Point	Point	Point	Point	Voice
back wall	4	Point	Laser	Grab	Point	Panto	Voi-Poi	Voice	Voice
desk	0	Touch	Laser	Throw	Touch	Panto	Touch	Voice	Voice
desk	0	Panto	Laser	Grab	Touch	Point	Touch	Voice	Touch
round tbl	2	Panto	Laser	Throw	Touch	Point	Point	Voice	Touch
round tbl	2	Point	Laser	Grab	Touch	Voice	Voi-Poi	Point	Point
counter	2	Touch	Laser	Throw	Point	Voice	Touch	Point	Touch
right cab	5	Voice	Laser	Point	Laser	Point	Point	Point	Point
counter	2	Voice	Laser	Point	Laser	Point	Point	Point	Point
floor	3	Voice	Laser	Grab	Laser	Point	Point	Voice	Point
to right	3	Voice	Laser	Throw	Laser	Point	Point	Voice	Point
round tbl	2	Touch	Laser	Grab	Touch	Touch	Point	Voice	Point
desk	0	Touch	Laser	Throw	Touch	Point	Touch	Voice	Touch
counter	2	Voice	Laser	Grab	Laser	Point	Touch	Point	Voice
desk	0	Voice	Laser	Throw	Laser	Point	Touch	Point	Voice

In the survey at the beginning of the study, eight of the nine subjects revealed some degree of discomfort with being monitored by cameras and microphones in the workplace, even if one could be guaranteed that the signal would not leave their office. At the end of the study, six of these eight subjects experienced a change of attitude and would consent to being monitored for the purposes of an environment like the one being studied.

DISCUSSION AND CONCLUSIONS

In this paper, we examined the role of gesture in interacting with projected user interfaces in an augmented office environment. We hypothesized that four factors seem to play a key role in defining such interfaces: *distance from the target*, *the user's spatial model of surfaces*, *the indications of discrete events*, and *the user's willingness to move*. Based on the different constraints imposed by each of these factors, we designed a set of four gestures for manipulating the location of projected objects.

We discovered a strong initial preference for voice–based commands, followed by point–and–retract and touch–and–retract techniques. However, when confronted with an actual task and a gesture repertoire enlarged by our set of gestures, users tended to abandon voice and more complicated gesture commands and *relied mostly on pointing as the means for object selection*.

Our results also show that distance plays a strong role in the selection of which gesture to use when interacting with projected user interfaces, with participants more often using touch gestures when the projected object or destination is very close to their body, and quickly reverting to point–and–drag gestures when the distance exceeded 2 feet.

Subjects clearly preferred remaining seated and making small body movements. Some mentioned feeling self–conscious making large pointing gestures and most simply insisted on staying in their chair, even when getting up would have made the task easier. Most telling were those participants who asked to use the laser pointer and then used it as they would a TV remote, using small movements to aim it at the projected objects and even smaller movements to turn the laser on and off, indicating discrete events. Motions that can be performed sitting, preferably with motions from the elbow, or even the wrist down, seem to be preferred. However, subjects did not like the pantograph/virtual mouse technique which also enabled them to interact while sitting and using only small movements. We conjecture that a possible reason is the difficulty to map the 2D plane of the virtual mouse onto the collection of disconnected projection surfaces around them.

The results of the study signal that sensing mechanisms constructed to recognize pointing gestures should focus on recognition of small movements. Given the current technological limitations, it is very likely that separate techniques for coarse and fine–grained positioning will be necessary. Although it is feasible to accurately track small body movements and map them to pointing or selection actions, there are some conditions which must be met to make these motions reliable for interaction. In particular, it is usually difficult to detect discrete events in small–scale hand movements. Users make such movements very often in the course of their work but only some of these movements will be salient to a tracking system. In order to use these movements, the sensing system must know *when* to attend to these movements, as well as *what* the user means by the movement and *where* to find the hand. Therefore the problem goes from one of finding new "gestures" for these types of interactions to one of answering *what*, *when* and *where* small hand movements are being used to manipulate the projected objects.

Finally, we observed that many participants demonstrated a general unwillingness to actually touch the physical surfaces, particularly after being exposed to the remote interaction gestures. Possible reasons for this include a desire for consistency in gesture selection (e.g., the participant always preferred to point instead of mixing selection modes or a general unwillingness to touch the surface). This may be related to the behaviors observed by Podlaseck et al. [20], where different surface materials seem to affect the willingness of participants to interact with them.

Based on these studies, a user interface for manipulating documents in an augmented office scenario should consider the use of pointing gestures involving small hand movements and negligible body movement. In many ways, the desired interface may resemble the traditional egocentric manipulation model used in most VR applications. However, implementing such an interface in the context of a projector/camera–based environment poses significant challenges for the sensing mechanisms as discussed above. Whether such a system is currently feasible deserves further consideration; one possible approach is the use of multi–resolution, multi–camera vision systems that track the user's body, limbs, and hands [2].

ACKNOWLEDGEMENTS

This work was completed while the first author was participating in the IBM Summer Co-op Program at the T.J. Watson Research Center in Hawthorne, New York. We would like to thank all of our study participants for their time and their creativity in manipulating our projected objects.

REFERENCES

[1] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S. and MacIntyre, B. Recent Advances in Augmented Reality. *IEEE Computer Graphics and Applications*, *25*, 6 (2001), 24–35.

[2] Bobick, A.F. and Bolles, R.C. The Representation Space Paradigm of Concurrent Evolving Object Descriptions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *14*, 2 (1992), 146–156.

[3] Bolt, R. Put That There: Voice and Gesture at the Graphics Interface. *ACM Computer Graphics, 14*, 3 (1980), 262–270.

[4] Bowman, D.A. and Hodges, L.F. An Evaluation of Techniques for Grabbing and Manipulating Remote Objects in Immersive Virtual Environments. In *Proc. 1997 Symposium on Interactive 3D Graphics*, (1997), 35–38.

[5] Cao, X. and Balakrishnan, R. VisionWand: Interaction Techniques for Large Displays Using a Passive Wand Tracked in 3D. In *Proc. UIST 2003*, ACM Press (2003), 193–202.

[6] Corradini, A. and Cohen, P.R. O the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. In *Proc. of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, Copenhagen, Denmark, (2002).

[7] Dahlbäck, N., Jönsson, A. and Ahrenberg, L. Wizard of Oz Studies: Why and How. In *Proc. IUI 1993*, ACM Press (1993), 193–200.

[8] Drucker, P.F. *Management: Tasks, Responsibilities, Practices.* 1st ed. Harper & Row, New York (1974).

[9] Hinckley, K., Pausch, R., Goble, J.C. and Kassell, N.F. A Survey of Design Issues in Spatial Input. In *Proc. UIST 1994*, ACM Press (1994), 213–222.

[10] Kidd, A. The Marks are on the Knowledge Worker. In *Proc. CHI 1994*, ACM Press (1994), 186–191.

[11] Kjeldsen, R., Levas, A. and Pinhanez, C. Dynamically Reconfigurable Vision–Based User Interfaces. In *Proc. Third International Conference on Vision Systems (ICVS 2003)*, Springer–Verlag (2003), 323–332.

[12] Koons, D.B., Sparrell, C.J. and Thorrison, K.R. Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures. In *Intelligent Multimedia Interfaces*, M. Maybury, ed., AAAI/MIT Press, Menlo Park, California, (1993), 257–276.

[13] Lai, J., Levas, A., Chou, P., Pinhanez, C. and Viveros, M. BlueSpace: Personalizing Workspace through Awareness and Adaptability. *International Journal of Human–Computer Studies, 57*, (2002), 415–428.

[14] MacIntyre, B., Mynatt, E.D., Voida, S., Hansen, K.M., Tullio, J. and Corso, G.M. Support for Multitasking and Task Awareness Using Interactive Peripheral Displays. In *Proc. UIST 2001*, ACM Press, (2001), 41–50.

[15] Mackinlay, J.D., Card, S.K. and Robertson, G.G. Rapid Controlled Movement through a Virtual 3D Workspace. *ACM Computer Graphics*, *24*, 4 (1990), 171–176.

[16] Malone, T.W. How Do People Organize Their Desks? Implications for the Design of Office Information Systems. *ACM Trans. on Office Information Systems*, *1*, 1, (1983) 99–112.

[17] Oviatt, S. Mutual Disambiguation of Recognition Error in a Multimodal Architecture. In *Proc. CHI 1999*, ACM Press (1999), 576–583.

[18] Pingali, G., Pinhanez, C., Levas, A., Kjeldsen, R., Podlaseck, M., Chen, H. and Sukaviriya, N. Steerable Interfaces for Pervasive Computing Spaces. In *Proc. First IEEE International Conference on Pervasive Computing and Communications (PerCom'03)*, IEEE, (2003), 315–322.

[19] Pinhanez, C. The Everywhere Displays Projector: A Device to Create Ubiquitous Graphics Interfaces. In *Proc. 3rd International Conference on Ubiquitous Computing 2001 (UbiComp'01)*, Springer–Verlag (2001), 315–331.

[20] Podlaseck, M., Pinhanez, C., Alvorado, N., Chan, M. and Dejesus, E. On Interfaces Projected onto Real–World Objects. In *Ext. Abstracts CHI 2003*, ACM Press (2003), 802–803.

[21] Poupyrev, I., Weghorst, S., Billinghurst, M. and Ichikawa, T. Egocentric Object Manipulation in Virtual Environments: Empirical Evaluation of Interaction Techniques. *Computer Graphics Forum, EUROGRAPHICS'98 Issue*, *17*, 3 (1998), 41–52.

[22] Raskar, R., van Baar, J., Beardsley, P., Willwacher, T., Rao, S. and Forlines, C. iLamps: Geometrically Aware and Self–Configuring Projectors. *ACM Trans. on Graphics*, *22*, 3, (2003), 809–818.

[23] Raskar, R., Welch, G., Cutts, M., Lake, A. and Stesin, L. The Office of the Future: A Unified Approach to Image–Based Modeling and Spatially Immersive Displays. In *Proc. SIGGRAPH 1998*, ACM Press (1998), 179–188.

[24] Raskar, R., Welch, G., Low, K. and Bandyopadhyay, D. Shader Lamps: Animating Real Objects with Image–Based Illumination. In *Proc. 12th EUROGRAPHICS Workshop on Rendering Techniques*, Springer–Verlag, (2001), 89–102.

[25] Rekimoto, J. and Saitoh, M. Augmented Surfaces: A Spatially Continuous Workspace for Hybrid Computing Environments. In *Proc. of CHI 1999*, ACM Press (1999), 378–385.

[26] Streitz, N.A., Geißler, J., Holmer, T., Konomi, S. Müller–Tomfelde, C., Reischl, W., Rexroth, P., Seitz, P. and Steinmetz, R. i–LAND: An Interactive Landscape for Creativity and Innovation. In *Proc. CHI 1999*, ACM Press, (1999), 120–127.

[27] Ware, C. and Balakrishnan, R. Reaching for Objects in VR Displays: Lag and Frame Rate. *ACM Trans. on Computer–Human Interaction*, *1*, 4, (1994), 331–356.

[28] Ware, C. and Rose, J. Rotating Virtual Objects with Real Handles. *ACM Trans. on Computer–Human Interaction*, *6*, 2, (1999), 162–180.

[29] Wellner, P. Interacting with Paper on the DigitalDesk. *Communications of the ACM*, *36*, 7, (1993), 86–97.

[30] Wilson, A. and Shafer, S. XWand: UI for Intelligent Spaces. In *Proc. CHI 2003*, ACM Press (2003), 545–552.

[31] Wingrave, C.A., Bowman, D.A. and Ramakrishnan, N. Towards Preferences in Virtual Environment Interfaces. In *Proceedings of the EUROGRAPHICS Workshop on Virtual Environments*, (2002), 63–72.