A Two Visual Systems Approach to Understanding Voice and Gestural Interaction

BARRY A. PO, BRIAN D. FISHER, AND KELLOGG S. BOOTH

Department of Computer Science, University of British Columbia
201-2366 Main Mall, Vancouver, B.C., Canada V6T 1Z4

Email: {po, fisher, ksbooth}@cs.ubc.ca
Phone: +1 (604) 822-8990
Fax: +1 (604) 822-8989

Abstract:

It is important to consider the physiological and behavioral mechanisms that allow users to physically interact with virtual environments. Inspired by a neuroanatomical model of perception and action known as the two visual systems hypothesis, we conducted a study with two controlled experiments to compare four different kinds of spatial interaction: (1) voice-based input, (2) pointing with a visual cursor, (3) pointing without a visual cursor, and (4) pointing with a time-lagged visual cursor. Consistent with the two visual systems hypothesis, we found that voice-based input and pointing with a cursor were *less* robust to a display illusion known as the induced Roelofs Effect than pointing without a cursor or even pointing with a lagged cursor. The implications of these findings are discussed, with an emphasis on how the two visual systems model can be used to understand the basis for voice and gestural interactions that support spatial target selection in large-screen and immersive environments.

Keywords:

two visual systems, pointing, cursors, visual feedback, voice input, visual illusions

Abbreviations:

VR, virtual reality; HCI, human-computer interaction; ANOVA, analysis of variance; CAD, computer aided design

1. Introduction

An important part of virtual reality (VR) research is the development of voice and gestural interaction. Ivan Sutherland, who is credited with first envisioning VR-style technology, noted that the muscles of the hands and arms allowed for such dextrous movement that gesture was a "natural choice" for computer control [1]. This was later reinforced by Bolt, whose classic "Put-That-There" demonstration explored the multimodal convergence of voice and gesture for spatial information systems [2]. Further implications of allowing people to use

speech and gesture to interact in ways most natural to them have been discussed by Oviatt and Cohen [3]. Today, there are many examples of VR interaction techniques that are based on voice, gesture, or the combination of both. Our focus is on understanding voice and gestural interaction for specifying spatial information, or the locations on an information display at which objects of interest are or will be displayed. We refer to this as spatial *target selection*, emphasizing the task of specifying the location of a target.

There are many assumptions made about what elements should be present in an interaction technique to support reliable voice and gestural interaction for target selection. For example, the inclusion of graphical feedback in the form of visual cursors for pointing is typical throughout the history of VR. Other graphical elements in addition to cursors that have been used to support interaction techniques include frames, scales, and shadows (in 3D environments). These graphical elements are used because of our intuition about the value of visual feedback and visual context. Especially with large-screen VR environments, the only graphical elements present are those within the display itself. The lack of contextual graphical elements in the surrounding environment could mean that graphical elements in the display take on stronger and sometimes unanticipated roles. With a renewed interest in multimodal interaction techniques for virtual environments, it is important to characterize the types of graphical elements that are present in a display and to understand their impact on voice and gestural input because this could have serious implications for multimodal techniques that are developed in the future.

Our approach to understanding the impact of graphical elements starts with a knowledge of the basic physiological and cognitive factors that allow people to use voice and gesture in virtual environments. In cognitive psychology and neuroscience, advances have been made in understanding how the functional architecture of the human brain allows people to perceive relevant aspects of the surrounding world and how this leads them to interact with the physical world in a useful, meaningful way. In particular, a neuroanatomical model known as the *two visual systems hypothesis* provides a glimpse of the complex relationship between visual perception and human motor movement [4, 5, 6, 7]. The hypothesis suggests that voice and gesture rely on different mental representations for the processing of visual information, and this could mean that user performance might

differ in the presence or absence of particular graphical elements when spatial tasks are performed.

We can use the two visual systems model to predict how target selection through voice or gestural pointing can be influenced by the presence or absence of two types of graphical elements: visible cursors and frames. We report the results of a study with two controlled experiments designed to test our predictions by contrasting the performance of users under four different interaction conditions: (1) voice-based input, (2) pointing with a visual cursor, (3) pointing without a visual cursor, and (4) pointing with a lagged visual cursor. By having users interact in an immersive virtual environment in the presence of graphical frames, we expected a visual illusion known as the *induced Roelofs Effect* to occur with certain kinds of input [8]. We found that pointing without a visual cursor and even pointing with a lagged visual cursor could outperform both voice-based input and pointing with reliable visual feedback under these conditions. Our findings are discussed in the context of how the two visual systems hypothesis can help build a better understanding of VR interactions that rely on voice and gesture.

2. Background and Related Work

Before describing the experimental approach we have taken, we review the fundamentals of the two visual systems hypothesis, which arose from a number of studies described in more detail later in this section [4, 5, 6]. The hypothesis proposes that two distinct mental representations of visual space are simultaneously generated when visual information is transduced at the neuroanatomical level. Figure 1 outlines the functional division between these two visual representations.

<- [Figure 1 goes about here] ->

The *ventral stream*, also known as the *cognitive stream* of visual processing, generates an allocentric, or world-relative view, of visually-perceived objects in the surrounding environment. The *dorsal stream*, also known as the *sensorimotor stream* of visual processing, concurrently generates an egocentric, or body-relative, view of these same objects. These streams are believed to have evolved independently from the biological need to use visual information to accomplish different tasks. The ventral stream is primarily responsible for enabling active object identification and parsing complex visual scenes, including the visual

perception of physical object properties such as color and shape (a "what" system) and the relative positions of objects in the scene. The dorsal stream is primarily responsible for enabling visually-guided motor movements, especially for physical actions that take place within peripersonal space, such as pointing, reaching, and grasping (a "how" system). It is less concerned with the relative positions of objects.

The world-relative view provided by the ventral stream is known to be susceptible to difficulties when dealing with egocentric judgments, sometimes manifested as "visual illusions." In contrast, the egocentric view provided by the dorsal stream is known to be robust against such ambiguities. Because the ventral stream was specialized to process the physical, but not spatial, characteristics of visual information (sometimes called "vision for perception"), it failed to evolve a robust mechanism for dealing with spatial ambiguities. Knowing what kinds of visual features influence the ventral and dorsal representations should be useful for predicting performance differences in target selection in VR applications between cognitively-based interactions, such as voice-based spatial target selection, versus motor-based interactions, such as pointing and other gestural techniques.

The original motivation for the two visual systems model came from Trevarthen's examination of split-brain monkeys and Schneider's work on modular retinal projections [4, 9]. Later work by Ungerleider and Mishkin characterized the ventral and dorsal streams of visual processing as respectively the "what" and "where" representations of visual space [10]. Milner and Goodale reported several experiments with a patient D.F., who exhibited the phenomenon of blindsight, or the inability to report visual awareness while retaining the ability to physically interact with the visual world [5]. Their studies further characterized the ventral and dorsal streams as "what" and "how" representations of visual space.

A large body of experimental evidence with healthy subjects supports the two visual systems model. Early work by Bridgeman, Lewis, Heit, and Nagle showed that subjects were not aware of visual displacements timed to occur in the middle of a visual saccade, a phenomenon known as saccadic suppression [11]. However, they also found that subjects were always able to accurately point at displaced targets, regardless of whether they were able to report the displacement.

Later work by Bridgeman, Kirch, and Sperling used the visual illusion of apparent motion, or the appearance that visual targets are moving in the presence of a displaced background, and showed that motor movements toward these targets remained accurate despite cognitive reports of apparent target movement [12]. A more recent study by Bridgeman, Peery, and Anand used the induced Roelofs Effect to study the two visual systems and found differences in cognitive versus sensorimotor forms of report for spatial target selection in a (real) physical environment [6, 8].

There is still vigorous debate in cognitive psychology about the exact mechanisms that underlie the two visual systems hypothesis and how researchers should interpret related evidence [13, 14, 15]. However, as a means of understanding the physiological basis for voice and gestural interaction, the two visual systems model holds considerable value. In the general domain of human-computer interaction (HCI), we have recently shown how the direct linkage between the two visual systems and the upper and lower visual fields of the human eye leads to a measurable impact on mouse and touchscreen selection performance on large graphical displays [16]. In the two experiments presented here, we show how voice and pointing interactions can be influenced by the two visual systems.

Visual Illusions and the Induced Roelofs Effect

Visual illusions are frequently used in experimental psychology to test the limits of the human visual experience but they are avoided for activities requiring visual perception in everyday life [17]. In VR, developers are often taught to consider the impact of visual illusions although these are more often mentioned in the context of theoretical study rather than applied practice.

<- [Figure 2 goes about here] ->

The two visual systems hypothesis provides an interesting context for studying visual illusions and their implications for VR. Recent psychophysical experiments testing the two visual systems hypothesis have involved the use of a visual illusion known as the induced Roelofs Effect, which has graphical elements that are similar to the ones seen in many kinds of VR graphical display applications [6, 7, 8]. Figure 2 illustrates the perceptual effects of this visual illusion. The induced Roelofs Effect is best described as a systematic bias in the

perceived location of targets presented within a surrounding rectangular frame. This rectangular frame is either symmetrically centered on the physical midline of a viewer or is asymmetrically offset by some distance to the left or right of the viewer. When the frame is centered, there is no perceptual bias; presented targets are consistently perceived in their correct locations. However, when the frame is offset to the left, targets within the frame are systematically perceived as being further to the right of the viewer than they really are. Likewise, when the frame is offset to the right, targets within the frame are systematically perceived as being further to the left of the viewer than they really are.

The rectangular frame is similar to bordering elements, such as virtual window frames or the physical walls of a CAVE, which provide visual context in a graphical display. The targets within the frame are likewise similar to the icons and interactive elements such as buttons, menu items, or other objects in a virtual world. Thus, the "basic" visual illusion of the induced Roelofs Effect offers an opportunity to study the influence of asymmetric frames in VR and how they could be a source for unintentional errors of execution in item selection and target acquisition tasks in immersive virtual environments.

The two controlled experiments presented here use an instance of the induced Roelofs Effect to understand how the two visual systems influence voice-based and gestural spatial target selection in an immersive virtual environment. Each experiment examines four different interaction techniques for specifying target location and measures the presence or absence of the systematic errors predicted by the induced Roelofs Effect.

Although the graphical elements in the induced Roelofs Effect are not a full VR application, they are representative components of many VR applications, as we have just noted. There are several advantages to using this basic stimulus over a more complex display. First, the simpler display removes many possible confounding display factors that could cause performance differences between the four interaction techniques. Second, using this kind of display allows us to design experiments with basic interaction tasks that are representative of those used in many VR systems. Third, using this kind of visual display allows us to understand how even the most basic of visual elements can impact user performance in VR applications.

3. **Predictions of User Performance**

The classic predictions associated with the two visual systems hypothesis suggest that purely cognitive forms of spatial interaction are susceptible to the perceptual biases of visual illusions while purely motor forms of interaction are robust against these biases. The hypothesis states that these performance differences are a direct result of different mental representations of visual space. Based on this, we formulated the following hypotheses:

- *Voice-based input* is an inherently cognitive form of interaction that solely depends on the ventral stream of visual perception because no direct, physical movement is required for the response. Thus, this kind of interaction will be most susceptible to perceptual ambiguities such as the induced Roelofs Effect.
- *Pointing without visual feedback* (i.e. without any visible graphical cursor) is an inherently motor form of interaction that solely depends on the dorsal stream of visual perception because there is a direct, physical movement required for response and there is no reliable way to make cognitive corrections to initial pointing movements. Thus, this kind of interaction will be unaffected by perceptual ambiguities such as the induced Roelofs Effect.

This type of prediction is central to most of the experimental work on the two visual systems in cognitive psychology. We extended these by making interpretations consistent with the two visual systems hypothesis to predict user performance with other interaction techniques. We formulated two other experimental hypotheses:

- *Pointing with visual feedback* (i.e. with a visible graphical cursor) engages the ventral stream of visual perception, thereby making it dependent on the cognitive representation of visual space. Thus, the presence of visual feedback means that "closed loop" interactions will be susceptible to perceptual ambiguities such the induced Roelofs Effect.
- *Pointing with lagged visual feedback* (i.e. with a temporally-delayed graphical cursor) could engage either the ventral or dorsal streams of visual perception based on the interaction strategy employed by the user. Users who disregard the visual feedback effectively make the pointing interaction an "open loop" interaction, like pointing without visual

feedback. Users who depend on the visual feedback effectively make the pointing interaction a "closed loop" interaction, like pointing with visual feedback. Thus, the presence of lagged visual feedback could cause some users to be susceptible to perceptual ambiguities such as the induced Roelofs Effect, while other users might not be affected.

These hypotheses may seem counterintuitive. They predict that voice-based interaction will be more susceptible to perceptual errors than will other kinds of physical interaction and they predict that pointing performance will be poorer *with* visual feedback present than when it is absent. Moreover, our hypotheses predict that a lag in displaying the visual cursor might actually *improve* performance compared to a non-lagged visual cursor. In contrast to these predictions, most VR applications assume that reliable feedback in the form of a visual cursor is necessary for pointing; emerging multimodal techniques assume that voice-based techniques are not susceptible to errors induced by the presence of graphical frames; and it is almost universally assumed that the presence of lag in visual cursors is detrimental to performance.

4. Experiment One

Our first experiment in the study was an initial test of our theoretical predictions. We devised a simple target acquisition task where vocal localization and spatial gesture interaction were equally feasible methods of interaction. In this experiment, subjects completed four blocks of trials that required them to select presented targets from fixed positions. Each block of trials used a different interaction mechanism for selection. One block used voice-based input and three blocks used pointing under varying levels of visual feedback. The experimental method was described in an earlier report that focused on the individual differences between subjects [7], but we summarize it again here for comparison with the second experiment.

Methods

This experiment used a within-subjects experimental design with four distinct conditions. Every subject attended a single, individual session lasting between sixty and ninety minutes where they completed all four conditions. Each condition was characterized by the use of a specific interaction technique and

consisted of one block of 54 trials. Every subject completed four blocks (one for each condition) for a total of 216 trials. The conditions were identified as follows:

- *(1) Voice-based input*. A multiple-choice voice protocol was used for target selection. All subjects began their sessions with this interaction method for target selection. No physical pointing interactions occurred in this condition.

The three remaining conditions used a continuous spatial pointing interaction for target selection. A handheld pointer (i.e. a pointing stylus) was used. These conditions differed from one another by the kind of visual feedback provided during trials.

- *(2) Pointing without visual feedback*. No tracking cursor was visible during this condition, meaning that subjects were effectively "blind" to their pointing movements during this block of trials.
- *(3) Pointing with visual feedback*. A tracking cursor was visible during trial pointing. The cursor was a graphical crosshair similar to the kind of visual feedback often used in interactive desktop and VR environments.
- *(4) Pointing with lagged visual feedback*. A small, temporally-delayed tracking crosshair was present during pointing. This was done by adding a one-half second lag to the cursor used in the pointing with visual feedback condition. Our intention was to identify the potential influence of lag on the two visual systems rather than to simulate the response lag seen in typical virtual environments. All subjects finished their sessions with this condition.

Conditions (2) and (3) were counterbalanced such that half of the subjects started with one condition before the other while the other half were presented with the conditions in the reverse order. This was done to keep the number of required subjects to a minimum compared to the number required for a fully counter-balanced study. Based on the literature we were confident that the voice-based condition would exhibit an effect, so we used it as a "baseline" condition. We kept the lagged cursor condition last because we were less confident about it and did not want to confound the comparison between the other two pointing conditions even though early pilot studies had suggested that our predictions were valid.

Subjects

Fourteen subjects between the ages of 17 to 31 were recruited for this study. Seven of the participants were male and seven of the participants were female. All had normal or corrected-to-normal vision. Twelve were right-handed and two were left-handed.

Apparatus

<- [Figure 3 goes about here] ->

Figure 3 presents a photograph of the VR display setting used in this study. The environment consisted of a three-screen, wide-angle projection surface, though only the center surface was used for the study. The active display was forward-projected and it had physical dimensions of 275 cm by 215 cm. Subjects were seated at a distance of 250 cm to avoid projector occlusion effects. With the exception of illumination from the projector and display surface, all light sources were extinguished. During the experiment, an experimenter was always present to facilitate session progression. A PC workstation and custom software were used to render trials and record quantitative data.

A large, wooden table was constructed and positioned directly in front of subjects to obscure viewing of their hands and arms during sessions. The table had dimensions of 120 cm by 95 cm by 80 cm (width, depth, and height). These table dimensions ensured that all subjects would have enough space to make free distal pointing movements. By requiring subjects to keep their limbs underneath the table throughout the experiment so they could not see where their hands were pointing, we were able to strictly control their perception of visual information to only that provided by the display. Subjects still had proprioceptive information about their physical pointing actions.

A Polhemus Fastrak was used for pointing in this experiment. Prior to each session, the Fastrak was calibrated using software and all sources of metallic interference were kept away from the transmitter and sensors. Subjects wore a head-tracker and a pair of active stereo glasses even though no stereoscopic imagery was presented to them during this study. These requirements were made to ensure compatibility with future studies that might involve binocular perception. A stylus pointer attachment for the Polhemus Fastrak was held in the dominant hand of subjects during the pointing conditions. The form factor of the

stylus was similar to that of a laser pointer or tracking wand, making it an ideal interaction device for experimental purposes. With the head-tracker and stylus operating simultaneously, the Fastrak achieved an update rate of 60 Hz.

Procedure

Trials consisted of a one-second presentation of a single, red circular target surrounded by a green rectangular frame on a black background (see Figure 2). The circular targets were 0.5 degrees of visual angle in diameter and could appear in one of three locations, either centered on the subject, or offset to the left or right of center by 1.5 degrees of visual angle. The rectangular frames were 21.0 degrees in width by 9.0 degrees in height, with a thickness of one degree. The frames were either centered on the screen relative to a projection of the participant's mid-line, or they were offset to the left or right of centre by 4.0 degrees.

After one second, the target and frame vanished, leaving only the black background. Either immediately, or after a four-second delay, subjects were instructed to indicate the position of the now-extinguished target using the interaction technique specified for the condition they were completing. This "extinguish and point" design was used to ensure a fair assessment of performance between all of the tested interaction conditions. If the target remained present on the display, the pointing without visual feedback condition would be at a severe disadvantage compared to the other interaction conditions and we might not be able to learn anything about the differences between interaction techniques relative to the two visual systems.

These trial parameters resulted in eighteen trial types: three target positions, three frame positions, and two response delays. Each trial type was repeated three times, yielding a total of 54 trials per condition. Trials were randomized prior to presentation in each condition such that no two consecutive trials in a condition had the same target position and frame position.

Voice Condition

Voice-based input was simulated for this experiment in a Wizard-of-Oz fashion. This meant that subjects were told to indicate target positions by providing vocal commands to the display software. In reality, the experimenter monitored subject responses and manually entered the responses into the

recording software by hand. Once participants were told to respond, they did so by providing a vocal command in the form of one of five possible choices: "Far Left," "Left," "Center," "Right," and "Far Right." Each choice corresponded to the potentially perceived position of a given target position. Although there were only three possible target positions in the actual trial set, the induced Roelofs Effect could have made it appear as though targets were at a greater eccentricity than they really were. In these instances, the "Far Left" and "Far Right" responses allowed subjects to respond appropriately to their perception of target positions.

Pointing Conditions

Pointing interactions were accomplished with the use of the Polhemus Fastrak and attached stylus. Responses were made by aiming the stylus at the display like a laser pointer. Once participants were satisfied with where they were aiming, they maintained their aim for approximately two seconds until an audio confirmation was provided. This kind of "point-and-dwell" response was required to avoid the inaccurate responses due to the pen-drop phenomenon that is common with button presses on input devices not grounded on a surface [18].

Training

Subjects were provided with instructions at the beginning of the session and prior to the start of each experimental condition. Each block of 54 trials was preceded by a minimum of fifteen practice trials where subjects were offered the opportunity to familiarize themselves with the response protocol for that particular condition. Practice trials were presented in roughly the same fashion as actual experimental trials with the exception that the rectangular frame remained fixed in a centered position.

Results

Subject data were analyzed using statistical techniques consistent with those used in psychophysical analysis and other similar human factors experiments [16, 19]. Our primary analysis consisted of a series of two-way analyses of variance (ANOVAs) with repeated measures over independent factors of target position and frame position for each subject, condition, and response delay. These two-

way ANOVAs allowed us to test for the presence or absence of the induced Roelofs Effect in each subject and each interaction technique, which in turn allowed us to verify our theoretical predictions. While we present a new aggregate analysis of our data here, further analysis in the context of individual differences is available in our earlier report [7].

If a participant were biased by the induced Roelofs Effect with a particular kind of interaction, it would be visible in the ANOVA as a statistically significant main effect of frame position with an accompanying main effect of target position, uncomplicated by higher-order interaction effects. In terms of our experimental predictions, we expected a high proportion of our subjects to show main effects of frame position in the voice-based input and pointing with visual feedback conditions. Fewer were expected to show such effects in pointing with lagged visual feedback. We expected few, if any, to show main effects of frame position in pointing without visual feedback.

We used a dependent measure of subject response that differed depending on the condition being evaluated. For the voice-based input condition, subject responses were provided categorically, as one of the five previously-described vocal choices. For the pointing conditions, subject responses were provided in a continuous fashion, recorded as the position on the graphical display where a line projected along the major axis of the stylus would intersect the screen. Since there were only horizontal variations in target and frame positions across trials, only the horizontal or *x*-axis of participant response was used in our analysis.

In the discussion that follows, we adopt standard practice in experimental psychophysics and only report the smallest F-values for significant results and the largest F-values for non-significant results. Across all conditions, all fourteen subjects had statistically significant main effects of target position [$F(2, 18) > 4.214$, $p < 0.032$]. This indicated that their responses were highly consistent and highly reliable and thus provided us with assurance that subjects were completing the verbal and pointing tasks as instructed.

With respect to main effects of frame position for each subject, we found significant differences in susceptibility depending on the interaction technique, indicating that some interaction techniques were indeed more susceptible to the induced Roelofs Effect than others. The following results are sorted from "most susceptible" to "least susceptible."

- *Voice-based input:* ten of the fourteen subjects had significant main effects of frame position [$F(2, 18) > 4.460$, $p < 0.027$], indicating that a majority of them were biased by the induced Roelofs Effect.
- *Pointing with visual feedback:* eight of the fourteen subjects had significant main effects of frame position [$F(2, 18) > 3.850$, $p < 0.05$]. A chi-square test against voice-based input showed no significant difference in the number of subjects affected by the illusion [$(1, N = 14) = 1.4$, $p = 0.237$].
- *Pointing with lagged visual feedback:* six of the fourteen participants had main effects of frame position [$F(2, 18) > 4.280$, $p < 0.030$]. A chi-square test against voice-based input confirmed there was a significant difference in the number of subjects affected by the illusion [$(1, N = 14) = 5.6$, $p = 0.018$].
- *Pointing without visual feedback:* only four of the fourteen participants had main effects of frame position [$F(2, 18) > 5.650$, $p < 0.013$], indicating that a majority of them were *not* biased by the induced Roelofs Effect. A chi-square test against voice-based input confirmed there was a significant difference in the number of subjects affected by the illusion [$(1, N = 14) = 12.6$, $p < 0.001$].

For the voice-based input condition, the measured effect size for those with main effects of frame position was exactly one target position, or 1.5 degrees of visual angle. For the three pointing conditions, the measured effect size for those with corresponding main effects was approximately half of a target position, or 0.75 degrees of visual angle. Interestingly, we found no consistent differences for any of the response delay analyses, which was a parameter included for compatibility with several two visual systems experiments reported previously in the cognitive psychology literature [6, 12].

The subject data and analysis from this first experiment supported our theoretical predictions. Within the framework of the two visual systems hypothesis, we can interpret our results as evidence that each of these interaction techniques predominantly draws upon different representations of visual space. Voice-based input and pointing with visual feedback appear to draw from the world-relative ventral representation of space while pointing without visual

feedback and pointing with lagged visual feedback are more likely to draw from the egocentric dorsal representation of space.

5. Experiment Two

Although our first experiment yielded interesting results, the possibility remained that this was not actually due to the two visual systems but to the presence of uncontrolled variables in the experimental design. First, the lack of full counterbalancing meant that subjects could have been influenced by order effects. Second, the use of discrete target positions meant that subjects could have resorted to perceptual memory to simply recall where targets were. Third, it was unclear why some subjects were still susceptible to the induced Roelofs Effect with the supposedly dorsal-driven physical pointing interaction conditions. A fourth problem, which was a consequence of the choice of discrete targets, was that we could not have a uniform measure of target selection error across all four interaction techniques because the categorical responses for the voice-based input could not be reliably scaled to match those in the pointing conditions.

We decided that it was important to conduct a second experiment with more stringent controls to address the weaknesses that were present in the first experiment and to allow more direct comparison of voice-based input and pointing. Many of these changes were the result of suggestions or comments on our preliminary report of Experiment 1 [7]. We also took this opportunity to simplify the experimental design. In this experiment, only immediate responses were made by subjects because there was no effect of response delay (immediate versus a four second delay) in Experiment 1.

Methods

As in the first experiment, we used a completely within-subjects design where subjects were asked to complete four blocks of target selection trials (one voice-based and three pointing) in a single, individual session. However, this experiment included a full counterbalancing of experimental conditions, which subsequently led to the testing of a much larger group of subjects. Moreover, this experiment presented targets over a continuous range instead of having them appear at fixed positions as in the previous study, allowing us to make direct comparisons of target selection accuracy across all four conditions. In addition,

the voice and pointing interaction protocols used in this experiment were substantially improved, which led to more rapid and more accurate measures of subject responses within individual trials.

Subjects

Twenty-four new subjects (sixteen male and eight female) participated in this experiment. Their ages ranged from 18 to 31 years. All were right-handed and had normal or corrected-to-normal vision.

Apparatus and Procedure

Similar to the first experiment, subjects were instructed to specify the location of targets surrounded by rectangular frames that were presented for exactly one second. The four kinds of interaction used in the first experiment were once again tested here. Most of the graphical display parameters remained the same, except that the circular targets were one degree of visual angle in diameter (double the size of the targets in Experiment 1) and could appear anywhere along an eight-degree horizontal range from the center of the display (up to four degrees to the left or right of center).

In the voice-based input condition, subjects specified their perceived location of targets on a nine-point scale, with the verbal indication "one" being furthest to the left and the verbal indication "nine" being furthest to the right. The choice of a particular point along the scale was effectively a subject's closest estimation of a target's presented position. Subjects were told to use whole numbers in their responses and that fractional values would not be accepted even though targets could be at non-integral positions.

During pointing trials, a passive stylus for the Polhemus Fastrak was held in the dominant (right-hand) of subjects while a standard mouse was held in the non-dominant (left-hand) of subjects. The two-handed stylus and mouse setup allowed subjects to point at the display with one hand while pressing a mouse button to confirm their final pointing position. This was used as a much cleaner alternative to the "point-and-dwell" mechanism used in the first experiment.

Unlike the first experiment, where subjects either responded immediately or after a four-second delay, subjects were instructed to always respond immediately after the presented target and frame had vanished. Trials were repeated twelve times for each of the three possible frame positions (non-offset, offset left, and

offset right) for a total of 48 trials per condition (4 blocks × 48 trials = 192 trials per subject). Trials were randomized in each condition so that targets were uniformly-distributed across the eight-degree horizontal visual range.

Results

<- [Figure 4 goes about here] ->

Like the first experiment, the statistical analysis of subject data was primarily meant to identify the presence or absence of the induced Roelofs Effect. However, the refined design of this second experiment also allowed us to better characterize the individual differences between subjects and interaction conditions. A global one-way ANOVA was performed on each of the four interaction conditions to analyze responses to manipulations of frame position across all subjects. This analysis was complemented by individual one-way ANOVAs for each subject in each of the four interaction conditions, consistent with the analysis used in the first experiment. Figure 4 provides a summary of the primary analysis, measuring the effect size of the induced Roelofs Effect across the varying frame positions and different interaction types tested in this study. The statistics below are again sorted from "most susceptible" to "least susceptible."

- *Voice-based input:* we found a statistically significant main effect of frame position overall for voice-based responses [$F(2, 766) = 252.85$, $p < 0.001$]. Sixteen of the twenty-four subjects were individually found to have significant main effects of frame position [$F(2, 30) > 3.35$, $p < 0.049$].
- *Pointing with visual feedback:* we found an overall significant main effect of frame position for pointing with a visual cursor [$F(2, 766) = 27.91$, $p < 0.01$]. Fourteen of the twenty-four subjects had individually significant main effects of frame position [$F(2, 30) > 3.80$, $p < 0.034$].
- *Pointing with lagged visual feedback:* we found no overall significant main effect of frame position for pointing with a lagged cursor [$F(2, 766) = 2.26$, $p = 0.105$]. None of the subjects had any significant main effects of frame position [$F(2, 30) < 1.82$, $p > 0.180$].
- *Pointing without visual feedback:* we found no overall significant main effect of frame position when pointing without any cursor [$F(2, 766) =$

1.29, $p = 0.277$]. None of the subjects had any significant main effects of frame position [$F(2, 30) < 0.88$, $p > 0.425$].

Our results in this experiment were consistent with our theoretical predictions and with the results found in the previous experiment. The additional controls added into this experiment give us confidence that our results were not just random variations but are evidence that the presence of certain kinds of graphical information can bias visual processing, and in turn user performance, in systematic, predictable ways. The use of continuous rather than discrete targets in all four conditions allowed us to directly compare performance across the four conditions by measuring effect size, as illustrated in Figure 4.

6. Discussion

The two experiments presented here suggest that the two visual systems hypothesis has relevance for VR and an understanding of voice and gestural interaction. By showing how user performance can be manipulated in this context, several lessons can be learned.

- *The relationship between visual perception and motor action is important for VR*. Many virtual environments have display form factors so large that the only perceptual cues provided by the environment are the ones that are being provided by the rendered graphics. Unlike desktop settings, where there are numerous kinds of perceptual framing cues like the physical edges of a monitor, immersive displays must often contend with having fewer contextual cues, leading to an increased chance that perceptually ambiguous effects such as visual illusions could have a substantial impact on user performance in VR if elements in the display masquerade as frames.
- *Perceptual judgments are not necessarily the same as motor judgments*. The ability to judge object sizes and spatial location is extremely important for certain kinds of tasks in VR, including computer-aided design (CAD) and design reviews in engineering. For systems that are safety-critical, it is possible to guard against illusory biases by using interactions that draw upon the dorsal representation of visual space rather than the ventral representation.

- *Voice interaction is more reliant on cognition than is gestural interaction*. The two visual systems model tells us that voice relies on "vision for perception" while gesture relies on "vision for action." This suggests that voice interaction might benefit from avoiding perceptual ambiguities in the physical structure of display information (i.e. ambiguities in visual cues such as color or texture) while gestural interaction might benefit from visual cues that lend themselves to motor manipulation.
- *Even basic graphical elements can have a profound impact on visually-guided interactions*. Considering how simple graphical elements like rendered frames and visual cursors can bias user performance, seemingly "obvious" design choices such as the inclusion of a tracking cursor or the presence of contextual asymmetry should be carefully assessed. This could be especially important when graphical elements are placed in the context of a much more complex display and cognitive resources are limited. Our experimental results suggest that minimization of visual information could be used to learn how to make interaction less demanding for users.

In both experiments, participants' data were analyzed completely within-subjects because we wanted a measure of the proportion of participants who were significantly affected by the induced Roelofs Effect under the various methods of input. This is consistent with statistical practice in experimental psychophysics and with more recent innovations in statistical inference that downplay the role of simple null hypothesis tests as the only criterion for experimental success [6]. Vicente and Torenvliet observe that averaging results across participants can be statistically misleading, yet this is often the only technique used to determine whether a particular phenomenon is practically significant [19]. They further indicate that an analysis of participants individually, as was done here, is a robust alternative method of statistical inference that can yield more information about practical significance than a single null hypothesis test alone. We chose to expose our results in this fashion to offer readers more information with which to make their own judgments about whether the effects presented here are practically significant for their own applications.

With respect to the use of voice and gesture in VR, our results should not be interpreted to suggest that one kind of interaction is uniformly better than another.

Rather, these experiments should suggest to researchers and developers that challenging one's intuition about the traditional use of visual information is especially important for VR systems that intend to succeed in real-world applications. The ability to combine voice and gesture into a virtual environment has advantages in many situations, but it is important to understand the elements that make these interactions useful and it is equally important to understand how their use changes the way visual items are perceived and processed to make decisions about presented information on graphical displays.

7. Conclusion

We have presented two controlled experiments that compared user performance with four different interaction techniques in a virtual environment. Our results characterized how the presence or absence of certain visual cues, such as tracking cursors and asymmetric frames, can influence voice and gestural interaction for spatial target selection. The induced Roelofs Effect is only one of a broad class of visual illusions that may behave similary; it serves as a test case.Our results demonstrate that a two visual systems approach to interaction can help researchers better understand the relationship between basic graphical elements and their impact on multimodal interaction in VR applications. The same may be true for other visual illusions, suggesting a number of avenues for future research.

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Sutherland IE (1965) The ultimate display. In Proceedings of the IFIP Congress 65: 506-508.
2. Bolt RA (1980) "Put-that-there": Voice and gesture at the graphics interface. In Proceedings of the 7th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH): 262-270.
3. Oviatt SL and Cohen PR (2000) Multimodal interfaces that process what comes naturally. In Communications of the ACM 43(3): 45-53.
4. Trevarthen CB (1968) Two mechanisms of vision in primates. In Psychologische Forschung 31: 299-337.
5. Milner AD and Goodale MA (1995) The Visual Brain in Action., Oxford Psychology Series 27. New York: Oxford University Press.

6. Bridgeman B, Peery S, and Anand S (1997) Interaction of cognitive and sensorimotor maps of visual space. In Perception and Psychophysics 59(3): 456-469.
7. Po BA, Fisher BD, and Booth KS (2003) Pointing and visual feedback for spatial interaction in large-screen display environments. In Proceedings of the 3rd International Symposium on Smart Graphics: 22-38.
8. Roelofs C (1935) Optische localisation (Optical localization). In Archiv für Augenheilkunde 109: 395-415.
9. Schneider GE (1969) Two visual systems: Brain mechanisms for localization and discrimination are dissociated by tectal and cortical lesions. In Science 163: 895-902.
10. Ungerleider LG, and Mishkin W (1982) Analysis of Visual Behaviour. Cambridge, MA: MIT Press.
11. Bridgeman B, Lewis S Heit, G, and Nagle W (1979) Relation between cognitive and motor-oriented systems of visual position perception. In Experimental Psychology: Human Perception and Performance 5: 692-700.
12. Bridgeman B, Kirch M, and Sperling, A (1981) Segregation of cognitive and motor aspects of visual function using induced motion. In Perception and Psychophysics 29(4): 336-342.
13. Michaels, CF (2000) Information, perception, and action: What should ecological psychologists learn from Milner and Goodale (1995)? In Ecological Psychology 12(3): 241-258.
14. Kerzel D, Hommel B, and Bekkering H (2001) A Simon effect induced by motion and location: Evidence for a direct linkage of cognitive and motor maps. In Perception and Psychophysics 63(5): 862-874.
15. Bridgeman B, Dassonville P, Bala J, and Thiem P (2003) What is stored in the sensorimotor visual system: Map or egocentric calibration? In Proceedings of the 3rd Annual Meeting of the Vision Sciences Society: 10.
16. Po, BA, Fisher BD, and Booth KS (2004). Mouse and touchscreen selection in the upper and lower visual fields. In Proceedings of the ACM Conference on Human Factors in Computing (CHI): 359-366.
17. Luckiesh M (1965) Visual Illusions: Their Causes, Characteristics, and Applications. New York: Dover Publications.
18. Myers BA, Bhatnagar, R, Nichols J, Peck CH, Kong D, Miller R, and Long AC (2002) Interacting a distance: Measuring the performance of laser pointers and other devices. In Proceedings of the ACM Conference on Human Factors in Computing (CHI): 33-40.
19. Vicente KJ and Torenvliet GL (2000) The Earth is spherical ($p < 0.05$): Alternative methods of statistical inference. Theoretical Issues in Ergonomic Science 1(3): 248-271.

Figure Legends

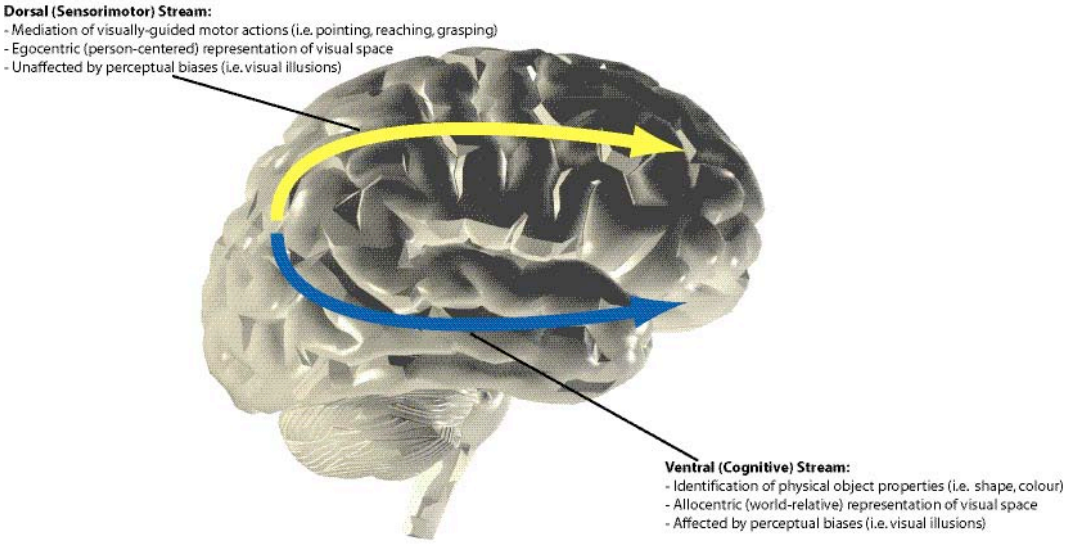

Figure 1. A general overview of the two visual systems hypothesis. Two separate mental representations of visual space are generated when processing visual information. The ventral stream (lower arrow) is specialized for the identification of physical object properties and maintains a world-relative view of space, while the dorsal stream (upper arrow) is specialized for the guidance of visually-based motor movements and maintains an egocentric view of space.

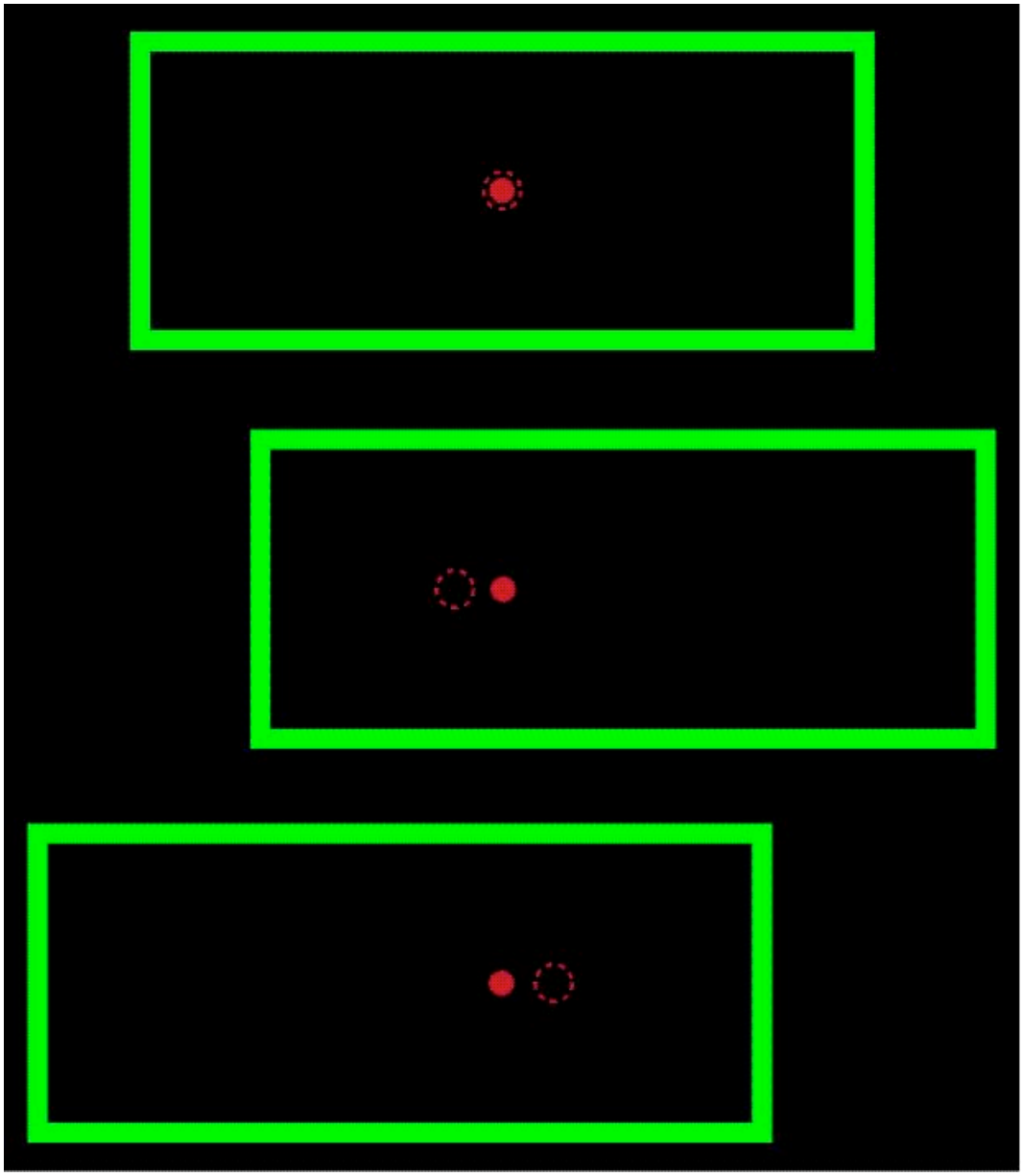

Figure 2. The induced Roelofs Effect. When targets are surrounded by an offset rectangular frame, targets appear more to the left or right of center than they really are. In the figure, solid circles represent actual target positions while dashed circles represent perceived positions.

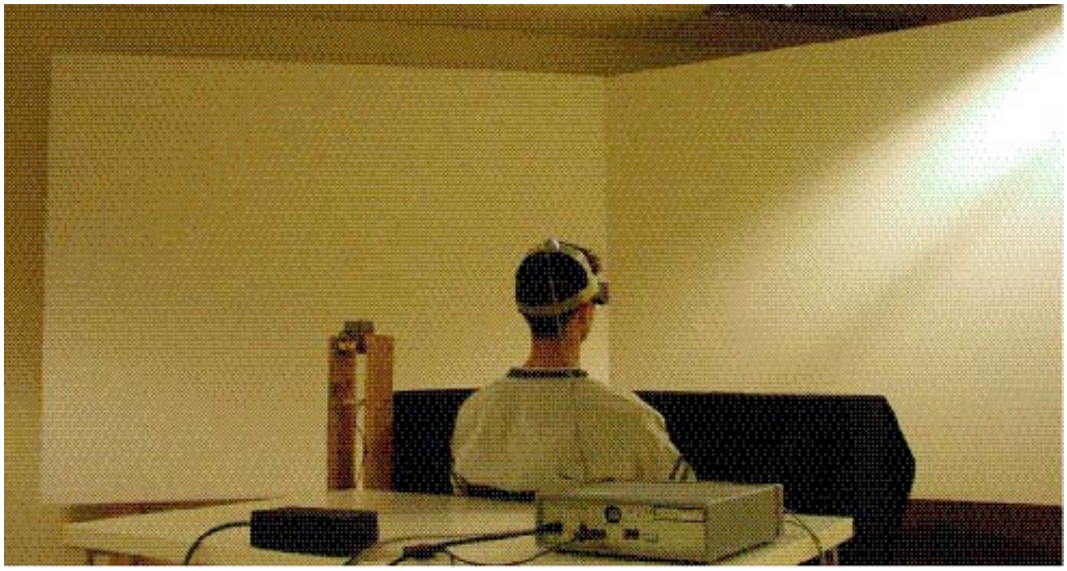

Figure 3. The VR display used for both experiments. Subjects were centered and seated before a three-screen, wide-angle display. Only the center display was used in this study. Spatial interaction was provided by a Polhemus Fastrak with an attached stylus and head-tracker. Arms and hands were kept underneath a large wooden table at all times. During sessions, all ambient light was extinguished and an experimenter was always present.

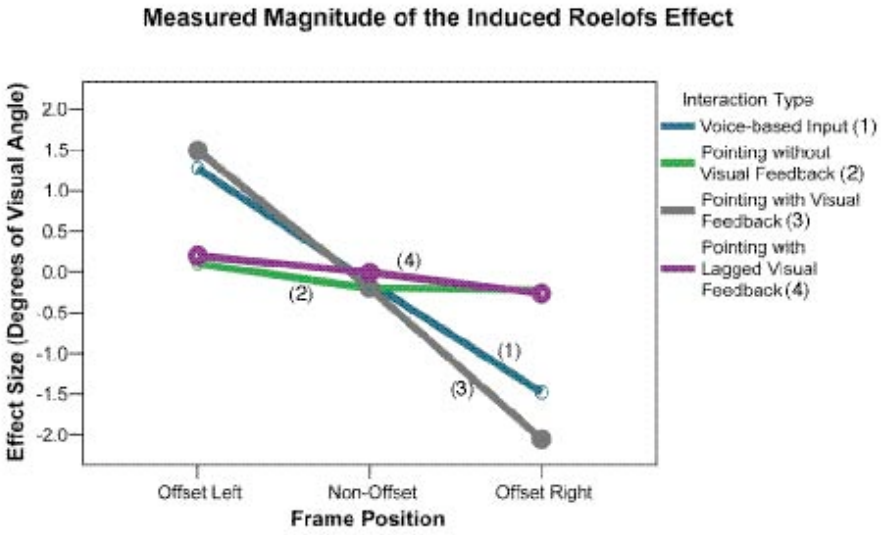

Figure 4. Measured effect size of the induced Roelofs Effect across the different interaction types and varying frame positions in Experiment Two. Effect size indicates the degree to which subject responses deviated from actual target positions. Negative effect sizes indicate response deviations to the left while positive effect sizes indication response deviations to the right, measured in degrees of visual angle. The steep slopes associated with voice-based input and pointing with visual feedback indicate these interaction types were highly susceptible to the induced Roelofs Effect. The corresponding horizontal slopes associated with pointing with lagged visual feedback and pointing without visual feedback indicate these interaction types were highly robust against the induced Roelofs Effect.