Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task

Eric D. Ragan, Doug A. Bowman, Regis Kopper, *Member, IEEE*, Cheryl Stinson, Siroberto Scerbo, and Ryan P. McMahan, *Member, IEEE*

Abstract—Virtual reality training systems are commonly used in a variety of domains, and it is important to understand how the realism of a training simulation influences training effectiveness. We conducted a controlled experiment to test the effects of display and scenario properties on training effectiveness for a visual scanning task in a simulated urban environment. The experiment varied the levels of field of view and visual complexity during a training phase and then evaluated scanning performance with the simulator's highest levels of fidelity and scene complexity. To assess scanning performance, we measured target detection and adherence to a prescribed strategy. The results show that both field of view and visual complexity significantly affected target detection during training; higher field of view led to better performance and higher visual complexity worsened performance. Additionally, adherence to the prescribed visual scanning strategy during assessment was best when the level of visual complexity during training matched that of the assessment conditions, providing evidence that similar visual complexity was important for learning the technique. The results also demonstrate that task performance during training was not always a sufficient measure of mastery of an instructed technique. That is, if learning a prescribed strategy or skill is the goal of a training exercise, performance in a simulation may not be an appropriate indicator of effectiveness outside of training—evaluation in a more realistic setting may be necessary.

Index Terms—Artificial, augmented, and virtual realities; Graphical user interfaces

1 Introduction

Trainers and educators in a variety of domains, including military [e.g., 1], medicine [e.g., 2], and athletics [e.g., 3], have begun to use virtual reality (VR) systems for task training. This approach was pioneered in the flight simulation community decades ago [4], but now the use of VR has expanded to motor skills training, decision-making / cognitive training, and psychological training in many domains. Common reasons for using VR include the following:

- Complete control over the environment and task stimuli; flexibility
- Repeatability
- Safe simulations of dangerous situations
- Ability to provide high levels of task and environment realism without exorbitant costs
- Ability to "immerse" the trainee in the training environment

Despite its widespread use, however, it is still difficult to say when VR training really works, when VR should be chosen over other training alternatives, and what sorts of VR systems provide the most effective training. In this work, we are focused on the last of these questions. Rephrasing the question, we ask, "How do the characteristics of VR training systems impact the effectiveness of those systems?" In particular, we focus on the effects of the realism, or *fidelity*, of the system.

Fidelity is a general and useful concept for characterizing different VR systems, since a common goal for VR is to provide a high-fidelity experience—one similar to the real world. Using stereoscopic graphics, using head movements to control one's view of the virtual environment, and using photorealistic textures are a few of the many ways that VR systems can provide high fidelity.

For training systems, it is a reasonable belief that higher fidelity will result in greater effectiveness [5]. In other words, it is intuitively better to train in a more realistic simulation of the real-world scenario than to train in a poor facsimile of that scenario. But is this always true, or are there cases where somewhat lower fidelity might be acceptable or even helpful? Is the highest possible level of fidelity required, or can we achieve very similar training effectiveness with lower levels? Previous research has shown that higher overall fidelity is not always necessary or advantageous over lower-fidelity simulations [e.g., 6, 7], and a better approach might be to ensure realism for certain elements of a simulation [8]. The challenge, then, becomes identifying

- *E.D. Ragan is with the Cyber and Information Security Research Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831. E-mail: raganed@ornl.gov.*
- *D.A. Bowman and S. Scerbo are with the Department of Computer Science, Virginia Tech, Blacksburg, VA 24061. E-mail: {bowman, scerbo}@vt.edu.*
- *R. Kopper is with the Pratt School of Engineering, Duke University, Durham, NC 27708. E-mail: regis.kopper@duke.edu.*
- *C. Stinson is with the Precision Nutrition, Toronto, ON M5E1W7, Canada. E-mail: cstinson@vt.edu.*
- *R.P. McMahan is with the Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080. E-mail: rymcmaha@utdallas.edu.*

Manuscript received 11 Apr. 2014; revised 6 Dec. 2014; accepted 29 Dec. 2014. Date of publication 12 Feb. 2015; date of current version 29 May 2015. Recommended for acceptance by J.D Fekete.
For information on obtaining reprints of this article, please send e-mail to: reprints.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2015.2403312

which components need to be realistic to be most beneficial for training effectiveness.

To address this challenge, we must be able to evaluate training effectiveness. The most common and straightforward approach is to look at *training transfer*, which is defined as the degree to which learned skills or knowledge can be applied to another situation [9]. Training effectiveness can be evaluated by assessing task performance after a training program [10]. Thus, evaluations of training simulators are often done by evaluating performance of the corresponding real-world task (i.e., whether success in the training system predicts success in the real world) after training with simulation [e.g., 2, 11].

The goal of the research reported in this paper was to examine the effects of relevant components of fidelity on the effectiveness of a VR training system. The system is designed to train users in the task of visual scanning, a common task in many contexts. For example, military personnel need to visually scan the environment to identify threatening objects or people; factory workers need to scan for defects in products; and sea rescue personnel need to scan for victims in the water. Visual scanning is a type of visual search, with the special requirement that it is important to search the entire scene systematically, ensuring that all target objects are found. Thus, having a well-defined visual scanning strategy is critical.

We chose to study the effects of the training system's *field of view* (FOV) and *visual complexity* for visual scanning tasks. FOV refers to the angular size of the area of the scene that a user can see instantaneously. A wider FOV allows the user to see more of the scene at once and to use peripheral vision, while a narrower FOV may reduce distraction in the periphery and allow the user to focus on the region of interest in the scene. Common VR systems have a wide range of FOVs, from less than 30 degrees (e.g., in some consumer-level head-mounted displays) to 180 degrees or more (the limitation of the human FOV; e.g., in surround-screen displays).

We use the term *visual complexity* to refer to the amount of detail, clutter, and objects in a scene [12]. The level of visual complexity is related to the fidelity of a simulation. Simulations with low fidelity often use simplified geometry and textures, and they may leave out some elements; this results in reduced visual complexity. High-fidelity graphical simulations can better replicate the visual richness and complexity of the real world. Training systems with low visual complexity may provide a scaffold for visual scanning, allowing trainees to learn the proper strategies in a simpler environment; on the other hand, systems with high visual complexity may provide more appropriate prepration for what users will encounter in the real world.

In order to study the effects of these two variables in a controlled way, we employed the mixed reality (MR) simulation approach [13], in which a single high-end VR system is used to simulate systems with lower levels of fidelity. In this experiment, we also used the highest-fidelity condition as a proxy for the real world so that we could study training transfer without loss of experimental control.

The results of our experiment contribute a deep understanding of the effects of FOV and visual complexity on training effectiveness for visual scanning tasks and, as a side benefit, also teach us something about the effects of these variables on raw task performance. More importantly, these results add to the growing body of literature on the effects of various components of fidelity [14], [15], which is needed to enable effective VR system design for training and many other application domains.

2 Background

In this section, we review related literature on the evaluation of VR training systems and the impact of fidelity, and on the understanding of VR fidelity.

2.1 Evaluating VR Training Effectiveness

VR-based training spans a variety of applications, such as flight simulators [16], surgical simulators [2], and medical examination training [17]. Studies have evaluated the effectiveness of VR training systems in different contexts. For medical examination training, Johnsen et al. [17] showed a significant correlation between performance in interview/examination sessions with virtual patients and performance with live patient actors.

As an example for flight simulators, Hart and Battiste [18] studied the effectiveness of simulation training games. The researchers compared flight school performances of participants who trained with a specialized flight-training game or commercial flight simulator game to those who had no additional game training. The results demonstrated how system design can have major impacts on training effectiveness: participants who trained with the specialized game had the highest continuation rates through the flight program, while participants who trained with the commercial flight game had the largest number of non-continuing students.

The effectiveness of VR simulators has also been demonstrated for surgical training, where a number of studies have shown significant gains in transfer of training and transfer effectiveness ratio for participants who trained in a simulator (as opposed to no additional training) before being assessed in real-world surgery [e.g., 2, 19]. Training effectiveness of virtual reality has also been demonstrated in other application areas, including stroke rehabilitation [20], pedestrian safety [21] and post-traumatic stress disorder treatment [22].

In a study of the effects of simulator fidelity on training effectiveness for a bicycle wheel-truing task, Baum et al. [23] compared a line-rendered graphics application with different physical props. Participants performed significantly better with more visually realistic props, but the fidelity of how well the props functioned did not make a difference. In a study with similar goals, Allen et al. [11] tested for effects of simulator fidelity on training transfer using an electromechanical-troubleshooting task. By manipulating the realism of the appearance and functionality of the physical training system, the researchers found evidence of faster problem solving after training with higher-fidelity systems.

Studying training for a real-world maze navigation task, Waller et al. [24] had participants prepare with either real-world navigation, a map of the environment, desktop VR, or immersive VR with a head-tracked HMD. Real-world training was the most effective overall, and immersive VR

was only advantageous over the other non-real conditions after longer periods of training.

2.2 Framework for Evaluating VR Fidelity

The experiment presented in this paper is one of many possible experiments on the effects of fidelity in VR systems. We believe this to be a fundamental question in the field of VR since one of the goals of much VR research is to increase the level of fidelity. Ivan Sutherland presented this vision for VR in his seminal paper "The Ultimate Display" [25], which described a display system that was indistinguishable from the real world. Research and development on such topics as high-resolution imaging [26], photorealistic computer graphics [27], and infinite walking through virtual environments [28] all point to the desire for greater fidelity. It is critical, then, to understand what effects these ever-increasing levels of fidelity will have on task performance, presence, satisfaction, acceptance, engagement, training transfer, and other outcomes. Even if we assume that higher levels of fidelity are usually better than lower levels, there is still a cost-benefit question to consider.

To study fidelity's effects, we must have a clear understanding of what fidelity is. Although we and others have been performing such studies for many years [e.g., 11, 29, 30], we have done so with an evolving understanding and with evolving terminology (e.g., compare [14] and [30]). Recently, we developed a more systematic framework to understand, describe, and evaluate fidelity in VR systems [31]. We present an updated outline of this framework here as a secondary contribution of this paper and to provide a foundation for future experiments on VR fidelity.

Consider the flow of information that occurs when a user interacts with a simulation. First, the user likely uses a piece of hardware or a tracked body part as an input device to generate some type of data. That data is then interpreted by software as some meaningful effect, which the simulation decides how to handle based on the physics and rules of the virtual world and the model data. Software then renders a representation of the current state of the simulated scenario, which is then displayed to the user through a hardware device.

This loop allows us to define and separate three types of fidelity in VR systems. We associate the realism of the input devices and interpretation software with *interaction fidelity*, the objective degree of exactness with which real-world interactions are reproduced in an interactive system. Similarly, we associate the verisimilitude of the displayed output with *display fidelity*, the objective degree of exactness with which real-world sensory stimuli are reproduced by a display system (note that display fidelity has also been referred to as *immersion*—see [32] for more details). Lastly, we refer to the realism of the simulated scenario and the associated model data as *scenario fidelity*, which we define as the objective degree of exactness with which behaviors, rules, and object properties are reproduced in a simulation as compared to the real or intended experience. The levels of fidelity for the interaction, display, and scenario categories can, in most cases, be assessed independently, and the combination of the three levels determines the overall realism of the simulation.

2.3 The Effects of Visual Fidelity and Complexity

Substantial research efforts have sought to evaluate the effects of fidelity in VR. Some examples of visual components of display fidelity include stereoscopy (the display of different images for each eye, providing additional depth cues), display resolution, FOV, field of regard (FOR; the range of the VE that can be viewed with physical head and body rotation), and refresh rate. Evaluating different components of display fidelity independently enables the understanding of what aspects of fidelity cause a benefit for particular applications. For example, in a previous study, we evaluated the effects of head tracking, stereoscopy, and FOR for a spatial judgment task [30]. The study found that performance was significantly better with head tracking or a wide FOR, and an interaction effect showed faster task completion when head tracking was coupled with stereoscopy.

Existing research has also provided evidence about the effects of varying visual complexity and FOV on search tasks. Lessels and Ruddle [33] investigated the effects of FOV (unrestricted vs. $20^\circ \times 16^\circ$) for a task involving navigation and searching in the real world. The study found no significant differences for performance metrics, though FOV did influence the types of search strategies used by participants. A second experiment evaluated the same search task in a virtual environment with two levels of visual fidelity (i.e., realistic textures and flat shading) and two travel techniques. The results showed that a constrained forward-only travel technique significantly outperformed unrestricted movement, and high-fidelity visuals led to significantly faster performance. These results suggest that, for a visual search task in a cluttered environment, it may be better to have lower interaction fidelity and higher visual realism.

Also related to visual search, a study by Pausch et al. [34] compared a tracked HMD to a non-tracked HMD with reduced FOV for a visual search task, with the results showing that participants more quickly determined the absence of targets with the head tracking and greater FOV. Looking at another search task, Lee et al. [35] used the MR simulation approach to study differences in visual realism for virtual and augmented reality. Their study found minimal effects of visual realism on task performance, but the authors explain that this may have been a side effect of the high difficulty of the task.

For a different study that involved finding data patterns in statistical analysis tasks, Arns et al. [36] compared a desktop display with a four-screen CAVE-like display with stereo and higher FOV. Results showed faster performance with the CAVE conditions.

Other studies have considered the effects of display fidelity and visual complexity on tasks involving spatial perception. Bacim et al. [37] studied different combinations of visual clutter and display fidelity for several spatial inspection tasks. The study found that higher display fidelity (in this case, the addition of head tracking, stereoscopy, and display screens) was beneficial for spatial judgments regardless of the level of visual clutter. In other work, Mania et al. [38] found that lower visual complexity (i.e., flat-shading, as compared to radiosity rendering) led to better spatial awareness of objects in a 3D environment. From other research, evidence indicates that visual realism is not

a factor in the known problem of distance underestimation in virtual environments [39], but FOV was shown to significantly affect distance estimation [40]. Studies have also shown that limiting FOV can reduce the speed and accuracy in maneuvering through a real-world obstacle course [41] and reduce the underestimation of perceived image motion [42].

In an initial investigation with visual scanning tasks in virtual environments, Kopper et al. [43] evaluated the effects of horizontal FOV and amplified head rotations. The study found that a narrow horizontal FOV of 30 degrees led to significantly worse performance than higher levels of 52 and 102 degrees in a visual scanning task similar to the one presented in this paper. The study did not find a significant difference in performance between the medium and high levels of FOV. This may have been due to the fact that vertical FOV was constant at a high level for all trials. In the study presented in this paper, the aspect ratio of the display was kept constant, such that both the vertical and horizontal FOV varied consistently.

Overall, these studies suggest that limited FOV can have negative effects on visuospatial perception and search, providing reason to expect a similar effect when training for visual scanning. The effects of visual complexity on training effectiveness are less clear. Reduced complexity may simplify training, allowing better task performance during training and helping trainees to focus on learning strategies. On the other hand, training in conditions less like the real conditions where the skills are needed might not adequately prepare trainees for the real tasks. Our study investigates the effects of FOV and visual complexity together in VR training systems.

3 Method

The primary goal of our experiment was to study the effects of fidelity on training effectiveness of a VR training system for an ecologically valid visual scanning task. The experiment measures how different levels of the FOV and visual complexity of the scenario affect performance and training transfer for a visual scanning task. Our design follows the assumption that the purpose of the training is to prepare for a real-world scenario that would have high visual complexity and unrestricted FOV. To this end, participants trained in a VR system with a given combination of the FOV and complexity levels. Then, for a controlled comparison, they performed the task in a high-fidelity VR scenario with high visual complexity and high FOV (i.e., as close to the assumed real-world conditions that the simulator could provide).

3.1 Hypotheses and Approach

We studied how the variables affect: 1) how well a given visual scanning strategy can be learned, and 2) target detection rate on the scanning task. The overarching hypothesis was that training in a system that is more similar to the intended simulated scenario would be more beneficial for training effectiveness. On a more specific level, our experiment tested the following hypotheses:

H1. Training with higher FOV will improve target detection in a later high-fidelity scenario more than training with a lower FOV.

H2. Training with higher FOV will lead to better adherence to the prescribed visual scanning strategy in a later high-fidelity scenario.

H3. Higher FOV will lead to better target detection during a scanning task.

H4. Training with higher visual complexity will lead to better target detection in a later high-fidelity scenario with high complexity.

H5. Training with higher visual complexity will lead to better adherence to the prescribed visual scanning strategy in a later high-fidelity scenario with high complexity.

H6. Higher visual complexity will lead to worse target detection during a scanning task.

In addition, to help investigate whether performance in a simulator might predict performance in a real-world setting, we tested hypotheses about the correlation between training performance and performance in the following high-fidelity scenario.

H7. Target detection performance in a training environment will be significantly correlated with performance in a later high-fidelity scenario.

H8. Target detection performance in a training environment will be significantly correlated with correct use of visual scanning strategy during a later high-fidelity scenario.

To study these effects in a controlled way, we employed the MR simulation approach [13], which we have used in many prior experiments [e.g., 30, 32, 44]. MR simulation is an evaluation methodology that studies mixed reality systems (including VR and augmented reality) using a single high-fidelity VR system to simulate systems and experimental conditions with equal or lower levels of fidelity. Systematically studying the effects of fidelity using MR simulation, rather than comparing different MR technologies, provides knowledge of the effects of individual design components. MR simulation studies have also been shown to produce valid results [44], although there have been exceptions [45].

In order to evaluate training effectiveness, our experiment contained three phases. The *instruction phase* was used to familiarize participants with the visual scanning task and the environment, and to teach them a prescribed scanning strategy. In the *training phase*, participants performed the visual scanning task multiple times in a particular condition (combination of FOV and level of visual complexity). In the *assessment phase*, participants performed the visual scanning task multiple times in the highest-fidelity condition.

3.2 Apparatus

An nVis SX111[1] head-mounted display (HMD) was used for the simulation. This HMD features dual displays (one per eye), each with a resolution of 1280 $\times$ 1024 pixels and a 50 degree binocular overlap. The total horizontal FOV of the HMD is 102 degree, and the total vertical FOV is 64 degree. The total weight of the HMD is 1.3 kg. Head-tracked viewing (orientation only) was enabled with a wired Intersense IS-900 tracker[2] on the HMD.

1. http://nvisinc.com/product.php?id = 48
2. http://www.intersense.com/pages/20/14

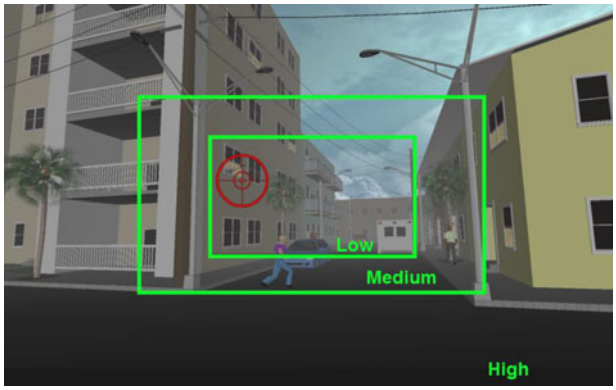

Fig. 1. Representation of the three levels of FOV.

Participants used a wireless tracked IS-900 wand controller in the dominant hand. The wand was tracked so that participants could point at objects in the environment. Pointing position was shown with a virtual crosshair, and participants used the wand's trigger button to indicate targets in a search task. Participants could freely turn their heads and bodies.

The software for the experiment was written using the Vizard Virtual Reality Toolkit by WorldViz,[3] with plugins to interface with the IS-900 and SX111 HMD. The application ran on a Microsoft Windows XP workstation with an Intel Core2 660 CPU at 2.40 GHz and 2 GB of RAM. The frame rate was approximately 50 frames per second for all conditions.

3.3 Experimental Design

The experiment followed a 3×3 between-participants design with FOV and visual complexity as the independent variables. This led to nine possible conditions, and each participant performed the experiment in one condition.

For the FOV variable, both horizontal and vertical FOV were varied together to maintain the aspect ratio of the maximum FOV supported by the SX111 HMD ($102^\circ \times 64^\circ$). FOV was varied in three levels: high ($102^\circ \times 64^\circ$; 120.41° diagonal), medium ($52^\circ \times 32.63^\circ$; 80.44° diagonal), and low ($30^\circ \times 18.82^\circ$; 35.81° diagonal). The medium and low FOV levels were chosen to simulate those of mid- and low-end commercial head-mounted displays. Fig. 1 shows how the three levels of FOV affected the view of the environment. To control the medium and low levels, the FOV was limited by virtual black blinders.

Visual complexity was also varied in three levels: high, medium, and low. The level of complexity was controlled by changing several components, including model-based factors and rendering factors (distance fog and skybox). The highest level of complexity had distance-based fog, a cloudy and detailed skybox, additional objects, more-detailed geometry, and more realistic texturing than the lowest level of complexity. The medium level of complexity was a balance between the high and low levels. Fig. 2 shows the three levels of visual complexity.

As dependent variables, we measured target detection and adherence to the scanning strategy. Target detection was measured for both training and assessment trials. Adherence to the scanning strategy was assessed by subjective ratings of how closely participants' visual scanning techniques followed the technique that they were trained to use (see Section 3.8 for further explanation of strategy ratings). Because we were primarily interested in studying the transfer of the scanning strategy, strategy was evaluated only during the assessment trials.

3. http://www.worldviz.com/products/vizard

Fig. 2. Screen shots of the three levels of visual complexity. The top image shows low realism, the middle shows the medium level, and the bottom image shows the highest level of complexity.

3.4 Visual Scanning Task

We consulted with experts to choose a single-user training task that was relevant to real-world activities and that was a reasonable target for a training system. In particular, we focused on the military domain. We found that it is common for military personnel to drive through urban streets to visually search for signs of dangerous activity and threatening individuals. This critical task requires great attention to detail and focus. We therefore chose *visually scanning an urban environment for threats* as the training task for our study.

Fig. 3. Example of a one-sided street used in the experiment. This image was taken from an out-of-simulation render to provide a clear overview of a street model.

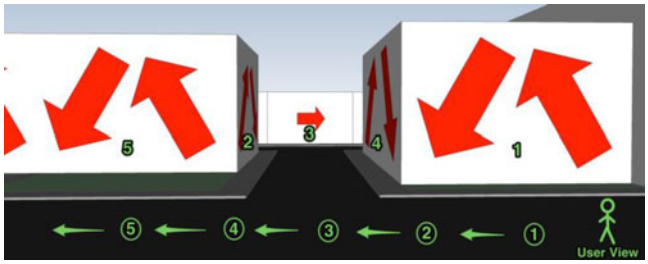

Fig. 5. Simplified view of a street intersection annotated to demonstration the prescribed scanning order. Building faces are simplified as white boxes. The circled numbers at the bottom of the image show the direction of the automatic movement down the street. The number labels on the building faces show which face the user should be scanning when the user is at the corresponding circled number along the street.

We designed the task so that participants had to search virtual city streets (see Fig. 3). During each trial each participant was moved automatically down a single street at a steady rate of 11.67 miles per hour (18.78 kilometers per hour). Aside from the motion of the viewer, the scene was static; the objects of the virtual scene were not animated. The virtual streets included simple models of people, and the targets for the search task were any people holding firearms. Fig. 4 shows examples of the target and non-target models. Due to the variety of colors of character models, buildings, and background objects, all character models had to be inspected in order to determine whether they were targets.

Participants were told to scan the right side of the street to find the targets (that is, participants did not need to turn more than 90 degrees to the left or right). We informed participants that there were between 12 and 18 targets in each trial (in fact, each trial had exactly 15 targets, but we concealed this fact to motivate participants to scan throughout the entire trial). We instructed participants to scan the environment using a particular strategy (described below) and to indicate each target found by pressing a button on a hand-held controller.

Our consultation with experts in the field revealed no standardized protocol for visual scanning in urban environments. Therefore, we developed our own prescribed visual scanning strategy. Our strategy is not necessarily the best method for scanning urban environments, but we confirmed with military experts that it was reasonable and would likely work well.

The basic concept of the visual scanning strategy is for users to use vertical head movements to scan building faces with sweeping up-and-down motions as they move down the street. Fig. 5 shows the general scanning directions with red arrows on simple, non-textured buildings. Because participants moved down the street from their right to their left side, the strategy's default scanning pattern had participants scan front-facing surfaces from the right side to the left as they swept up and down.

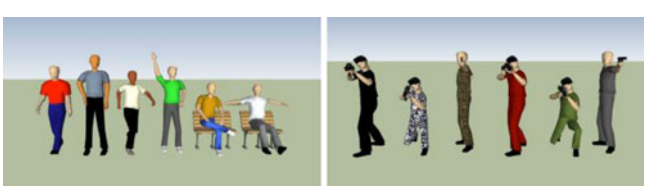

Fig. 4. Examples of virtual human models from the visual scanning task. The left image shows non-targets. The image on the right shows target models holding firearms.

The strategy changed slightly when participants approached the intersecting (perpendicular) side streets or alleys. Fig. 5 shows the order that building faces were to be scanned (the white boxes represent buildings). The image shows a view looking straight down an intersecting side street. Note that movement along the main street would be from the right side to the left in the figure. Fig. 2 (bottom) shows a similar view of an intersecting street but with a detailed street model. We trained participants to scan intersections by beginning with the left-most face of the intersecting street (i.e., the face labeled with 2 in Fig. 5—the first face that would be visible when moving from the right to the left), then by looking down through the intersection and sweeping across the furthest surface from the main street (i.e., surface 3 in Fig. 5). Finally, the intersection scan finished by sweeping the remaining side (i.e., the right side, or surface 4 in Fig. 5) of the intersecting street. This strategy affords a strong perspective of the intersection because it allows viewing of building faces as soon as they are visible.

After participants had passed by the intersection or alley, they resumed the right-to-left, vertical scanning pattern of buildings along the main street. The strategy training also instructed participants to avoid looking too far ahead (down the street in the direction of movement) or too far behind them (where they came from).

Since we did not use eye tracking but wanted to keep track of the visual scanning strategy, we instructed participants to point the crosshair where they were looking, meaning that the location of the crosshair would match the current point of gaze. While this pointing method does not provide a perfect measure of gaze, the method was appropriate for our evaluation training transfer. That is, we trained participants to use a specific scanning technique, and pointing with the crosshair was a component of that technique. Consequently, crosshair movement provided an effective indicator of strategy adherence.

3.5 Environment

In this study, participants were automatically moved straight down an urban street environment. Participants scanned only one side of the street (because the view was controlled with head tracking, participants could physically turn 180 degree to look at the opposite side of the street,

TABLE 1
Breakdown of Street Models Created with Different Levels of Visual Complexity

Level of Visual Complexity	Instruction Models	Training Models	Assessment Models
Low	5	15	0
Medium	5	15	0
High	5	15	5
Total	**15**	**45**	**5**

which was empty). Each street was 800 feet (243.84 m) long and had exactly three side streets, although the locations of the side streets varied between models.

Different street models were created so that 1) each participant could complete multiple task trials and 2) different models fit the three levels of visual complexity. A total of 65 street models were created. All participants saw 25 street models throughout the study, but the models that participants scanned during the instruction and training phases depended on the level of visual complexity in the given experimental condition. Since the assessment was always done in the highest-fidelity condition, all participants scanned the same high-complexity street models in the assessment phase. Table 1 shows the breakdown of street models, and the following sections describe the model designs for the instruction, training, and assessment phases of the experiment.

3.5.1 Instruction Models

During the instruction phase, each participant went through five instruction trials corresponding to the assigned level of visual complexity (therefore, there were a total of 15 instruction models). All five of the street models for each condition featured the same geometry and street layout, but environmental features were added incrementally as the instruction progressed. Additional details about the progression through the instruction phase are described in Section 3.6.

3.5.2 Training Models

During the training phase, each participant went through 15 trials with street models corresponding to the assigned visual complexity condition (therefore, there were a total of 45 training models). Instead of creating 15 unique layouts for each condition, we created three base layouts, with five variations of each having different building color and texture. People, vehicles, plants, other elements, and 15 targets were distributed throughout each of the models for that condition. The targets were dispersed so there were always five at street level, five in windows or on balconies, and five on building rooftops.

3.5.3 Assessment Models

During the assessment phase, each participant went through five trials in the highest-fidelity condition. All participants used the same five high-complexity assessment models. The assessment models featured unique street layouts but used the buildings from the training models. The textures and colors of the buildings were changed, and the locations of people, vehicles, plants, and other elements varied among models. The 15 targets were dispersed throughout the models according to the same structure as the training models—five targets at street level, five in windows or on balconies, and five on top of rooftops.

TABLE 2
Differences between Levels of Visual Complexity for Models

Street Model Details	Low Complexity	Medium Complexity	High Complexity
Number of targets	15	15	15
Number of side streets	3	3	3
Street length	800 feet	800 feet	800 feet
Number of alleys	0	3-4	5-8
Depth of side streets and alleys	50 feet	75 feet	100 feet
Building complexity	Flat faced and at street level (no recessed buildings), flat textures, no balconies	Combination of flat and complex textures, some recessed buildings, some balconies	All complex textures, many recessed/varied shaped buildings, many balconies
Vehicles	10-14	15-24	20-29
People (non-targets)	11-17	16-32	24-39
Sky	Solid blue	Textured blue with some clouds	Textured with many clouds
Plants	No plants	Some plants	Many plants
Additional elements	No street lights, power lines, benches, dumpsters, or patio furniture	Some street lights, power lines, benches, dumpsters, and patio furniture	Many street lights, power lines, benches, dumpsters, and patio furniture

3.5.4 Three Levels of Visual Complexity

Since the level of visual complexity was varied between participants, we needed three separate groups of models for low, medium, and high levels. Fig. 2 shows representative screenshots of the different levels of complexity.

We developed the high-complexity models first and then developed the medium- and low-complexity models by simplifying the high-complexity versions. Thus, each set of three models shared a similar street layout and building architecture, and the ordering of three levels of complexity was guaranteed for each set. Side streets were always in the same places, and the overall skyline (building height and layout) was comparable (but not identical) between the three models in each set. Variations between the models required modifications to the width/depth of some buildings and removal or merging of others.

The people, vehicles, plants, and other elements were also systematically simplified from the initial high complexity models. Details on the exact differences between the three levels of visual complexity are show in Table 2.

3.6 Procedure

The study was approved as required by the Institutional Review Board at our university. Upon arrival, participants were given an informed consent form to read and sign. They then completed a background questionnaire to provide basic information about education and experience with technology. After that, they were given an Ishihara Color Test [46] to detect color blindness. Color-blind participants were dismissed.

Participants were then briefed on the environment and task. We showed them images (shown in Fig. 4) to help explain which models represented targets (people with firearms) and which were non-targets. They were then shown a

diagram of the scanning strategy they needed to use to sweep the environment (similar to Fig. 5). Participants were instructed to follow their gaze with the crosshair and to try to stick to the visual scanning strategy at all times.

After participants acknowledged that they understood the task and scanning strategy, they were introduced to the HMD and guided through five instruction trials. These trials were displayed at the level of FOV and visual complexity for the assigned experimental condition. In the first instruction trial, buildings were textured with arrows representing the scanning strategy (see Fig. 5), and an automatic moving spotlight guided the participant's eyes to demonstrate the strategy. Additionally, the first trial was paused periodically to give the experimenter time to slowly explain the scanning strategy in action.

The second trial still used the spotlight guide, but used the standard building textures instead of arrows. For the third trial, the spotlight scaffold was removed and additional objects were added (but no targets were present).

In the fourth instruction environment, targets were added. The participant viewed an automatically moving ideal scanning trial, which stopped at each target to ensure the participant saw it. Participants practiced clicking the trigger to indicate when they identified a target. The fifth instruction model was the same as the fourth but with objects and targets in different locations. This trial allowed the participant to practice scanning and identifying targets in the same conditions that would be used in the following training trials.

After the last instruction trial, the experimenter immediately scored target detection and strategy performances with the participant and provided feedback. Throughout the instruction series, the experimenter watched the participant's performance and provided critique to encourage participants to follow the strategy and align the crosshair with gaze direction. Participants were then given a five-minute break to conclude the instruction phase.

After the break, participants performed 15 training trials with the same combination of FOV and visual complexity level as in the instruction phase. After each training trial, participants reviewed the trial and received performance feedback to help them improve their adherence to the prescribed strategy. Participants were asked to watch a replay of the trial in the HMD. The experimenter reviewed the trial with the participant at the same time (using a separate monitor). The replays paused at each point where the trigger was clicked, and the experimenter would determine whether or not a target was correctly identified. The experimenter could manipulate the angle and zoom of the environment when necessary so that both experimenter and participant could determine whether the identified elements were in fact characters with firearms. The experimenter provided feedback on how well the participant was following the prescribed strategy and made recommendations for improvement (if necessary). At the end of the replay, the experimenter provided the participant with a performance summary of the number of targets found and the number missed.

Participants had a five-minute break after the seventh training trial and another five-minute break after the final training trial. Finally, participants performed five assessment trials in the condition with the highest FOV and visual complexity. During the assessment phase, replays were not reviewed and the experimenter did not provide feedback on the participant's performance or strategy.

Participant sessions took approximately 90 minutes.

3.7 Participants

We recruited a total of 51 participants, but six did not complete the entire experiment either because of simulator sickness effects or dismissal due to color blindness. Thus, 45 participants completed the study (five per each of the nine conditions). All but one were students; 13 were graduate students, 30 were undergraduates, and one did not specify. Students were from a variety of disciplines—the most common of which were computer science (13) and psychology (10). Participant age ranged from 18 to 37 years, with a median age of 21. Seventeen participants were female. The majority (all but six) of participants reported that they had experience with video game systems that used motion tracking. Thirty-two participants reported playing first-person shooter video games.

3.8 Assessment of Scanning Strategy

To study the transfer of the prescribed scanning strategy to the assessment environment, we developed scoring criteria to measure how closely participants' scanning techniques in the assessment trials followed the prescribed technique, and independent raters scored each assessment trial's adherence to the strategy. Trials were recorded from the participant's point of view. Because participants were instructed to move the crosshair to follow their gazes, the movement of the crosshair made it possible to observe their scanning patterns.

Though the criteria for strategy analysis was well defined, perception of how well participants adhered to the strategy was still somewhat subjective. Thus, scanning strategies were analyzed by a team of three raters who each reviewed all five assessment trials for all 45 participants. The entire list of 225 assessment trials was randomly ordered (with different orderings for each rater), and an anonymized identification code was assigned to each trial. Because all assessment trials used the high-complexity models with the highest FOV, the raters had no information about which conditions the participants had trained with. One of the raters was a member of the research team who had not overseen the experimental trials and had no knowledge of the viewing order. The other two raters were external to the research team.

3.8.1 Rating Procedure

Prior to scoring the assessment trials, all raters went through a training session to demonstrate the prescribed scanning strategy. First, to demonstrate the technique, the session included the explanation that all participants went through at the start of the experiment. Next, raters were instructed on how to score trials using trial playback software and paper scoring sheets. The playback software allowed raters to view the anonymized trials, pause playback, rewind playback, and choose between real-time and half-time playback speeds. Scoring sheets showed the building layouts for each of the five models used in the

assessment trials, showing outlines of the building faces that were to be scanned.

Strategies for each assessment trial were scored in two ways: component surface scoring and summary scoring. For component surface scoring, raters provided a strategy score (with values from 0 to 3) for each individual surface (i.e., face of a building). A score of 0 meant that the surface had not been scanned at all, as judged by the position of the crosshair. A score of 1 indicated minimal scanning coverage of a surface, but not in adherence to the instructed strategy. A score of 2 meant a reasonable level of surface scanning while following the prescribed strategy, while a score of 3 indicated that the surface was scanned in perfect accordance to the instructed strategy. Total surface scores could then be calculated for each street model by summing the scores for the individual faces. Thus, this method provided a metric for strategy adherence that took each individual scanning surface into account.

The second method of scoring was summary scoring, which assigned a holistic rating of the overall quality of the strategy used over the entire trial (a single street model). Values for summary scores ranged from 1 to 10 (inclusive) as a single number corresponding to how well the participant's strategy followed the instructed strategy.

The scoring sheets provided locations for raters to record both summary scores and component surface scores. Once raters understood the scoring criteria, they viewed examples of fabricated trials that demonstrated different levels of adherence to the instructed strategy. These trials allowed for practice using the playback software and scoring the trials, and a member of the research team was present to answer any questions about the process or scoring. Following the practice, raters viewed and scored the participant assessment trials. To account for the possibility of raters adjusting their scoring sensitivities with more exposure to trials, the batch of all trials included five extra trials at the beginning of the set. These first trials provided additional practice and gave raters a chance to establish a baseline for the subjective component of the strategy scoring. Raters then scored the 225 trials in their given random orders.

3.8.2 Inter-Rater Reliability

Due to the subjective nature of the strategy scoring, we tested for inter-rater reliability to check consistency of ratings. We judged the individual surface and component ratings to be ordinal measures due to the possibility of subjective interpretations between score values. For our analysis, it was important that raters were consistent in the assignment of high or low scores (relative to each rater), but the raters did not have to agree in terms of exact score values (i.e., we were not concerned with inter-rater agreement). To this end, we used Spearman correlations to judge inter-rater consistency (following the rationale provided by Stemler and Tsai [47]), and we tested for correlations among the three combinations of the three raters (as done by others, such as [48]) for all scored trials ($n = 225$). All correlations were significant with $p < 0.001$ (Spearman's ρ values ranged between 0.5 and 0.9). These results show high inter-rater reliability for both component surface scoring and summary scoring.

We also tested for intraclass correlation (ICC) among raters using two-way mixed averages measures for consistency, following Shrout and Fleiss [49]. The test yielded ICC (3, 3) = 0.868, showing strong reliability (note that 0.8 is often used as a high standard for reliability; see [50] for further explanation).

4 Results

We tested for the effects of FOV and visual complexity on both target detection and scanning strategy performance. We tested for effects due to FOV and visual separately, and we also tested for interactions between the two variables. Only significant effects are reported for ANOVA tests and posthoc analyses. For all statistical tests, $n = 45$.

4.1 Target Detection Results

Detection performance on the scanning task depended on the correct identification of targets and the number of false identifications. Note that target detection was assessed separately from scanning strategy ratings. We present the hit detection rate (the percentage of correct identifications out of the total number of targets) and error rate (the percentage of false-positive identifications of non-target characters out of the total number of non-target characters). Detection was analyzed separately for training trials (with the experimental levels of FOV and visual complexity) and for the assessment trials (all having the highest levels of FOV and complexity).

Hit rate data were judged to be normally distributed, with the results of Shapiro-Wilk tests detecting no evidence to the contrary, and Levene's tests showing homogeneity of variance across conditions. Thus, two-way independent factorial ANOVA tests were used for statistical analyses of the effects of FOV and visual complexity on hit rate. In contrast, false-positive rates were positively skewed, so the data were transformed with the square root function to meet the assumptions of two-way factorial ANOVAs.

4.1.1 Detection Performance in Training Phase

The overall hit rate during the training phase had $M = 63.43$ and $SD = 17.86$. As expected, overall target detection rate significantly improved as the training progressed (significant Pearson's correlation yielded $r = 0.56$ and $p = 0.016$).

Fig. 6 shows training detection means and standard error broken down by FOV and visual complexity. The ANOVA found a significant effect of FOV on target detection in the training phase, with $F(2, 36) = 10.58$, $p < 0.001$, and $\eta_p^2 = 0.37$. Bonferroni-corrected post-hoc tests showed high FOV was significantly better than low with $p < 0.001$ and Cohen's $d = 0.84$, and medium was significantly better than low with $p < 0.01$ and $d = 0.62$.

The ANOVA for hit rate also found a significant effect of visual complexity, with $F(2, 36) = 57.62$, $p < 0.0001$, and $\eta_p^2 = 0.76$. Bonferroni-corrected post-hoc tests showed significant differences between all levels of complexity with $p < 0.001$, with lower levels better than higher levels. Effect sizes were notably large, with Cohen's $d = 3.03$ between low and high complexity, $d = 1.56$ between low and medium, and $d = 1.99$ between medium and high. Errors (i.e., false positives) were more common in conditions with higher visual complexity due to larger numbers of

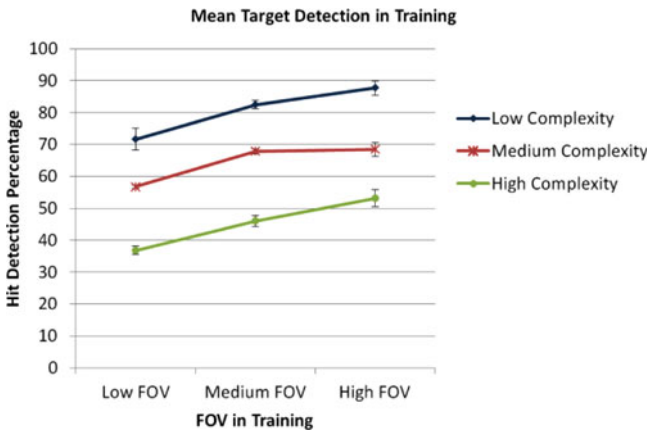

Fig. 6 Mean target detection performance scores in training. Error bars show standard error.

non-targets (see Table 2), and the total error count with high complexity was greater than with or medium.

To account for differents number of non-targets in conditions, we tested error rate (i.e., percentage of errors out of total non-targets). Error rates were low across all conditions (percentage was M = 1.16 with SD = 1.11). The ANOVA for error rates was significant for FOV with F(2, 36) = 3.32, p = 0.047, and $\eta_p^2 = 0.16$. The post-hoc Bonferroni tests only found high FOV (M = 0.82, SD = 1.00) had significantly lower error rate than medium FOV (M = 1.48, SD = 1.20) with p = 0.043 and d = 0.33. The ANOVA also detected a significant effect for complexity with F(2, 36) = 4.00, p = 0.027, and $\eta_p^2 = 0.18$. The post-hoc test found high complexity (M = 1.60, SD = 1.07) had significantly worse error rate than the medium level (M = 0.83, SD = 0.78) with p = 0.040 and d = 0.81.

4.1.2 Detection Performance in Assessment Phase

After training with the assigned combination of FOV and visual complexity, the assessment phase always had high FOV and high complexity for all participants. Overall hit detection rate was M = 40.71 and SD = 9.05 during assessment. We tested for effects of different levels of FOV and visual complexity used in training on detection performance during the assessment trials. The ANOVA for assessment hit detection rate did not detect significant effects for FOV, visual complexity, or the interaction between the two. Similarly, no significant effects were found for error rate during assessment. Error rates were low (overall, M = 1.16 and SD = 1.11).

These results suggest that the differences in experimental training conditions did not, in fact, cause any differences in target detection performance during the assessment trials. Though the different levels of visual complexity did significantly affect scanning strategies (see Section 4.2), these differences were not detectable by considering performance alone in the assessment trials. To further test this result, we conducted a one-tailed Pearson's correlation test between training performance and assessment performance scores. The test did not find a significant correlation, yielding r = 0.200 and p = 0.094.

4.2 Strategy Transfer

To produce the final strategy metrics, we summed the scores for the three raters and calculated the percentages of

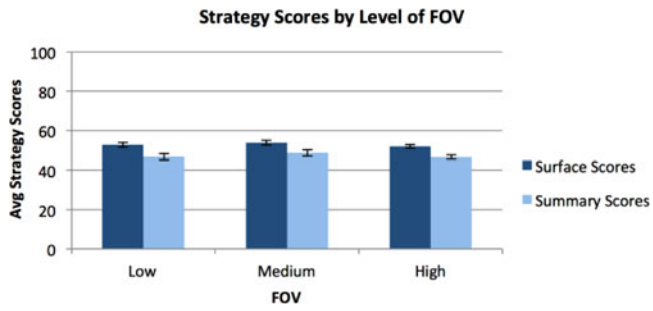

Fig. 7. Mean strategy scores from assessment trials by varying levels of training FOV. Error bars show standard error.

the maximum possible scores. We analyzed the effects of FOV and visual complexity on both types of strategy scores using two-way independent factorial ANOVA tests. We note that the experimental design satisfied the assumptions for parametric testing. Both surface scores and summary scores met the conditions of normality and homogeneity of variance (by Shapiro-Wilk and Levene's tests).

Fig. 7 shows strategy scores by FOV conditions. The ANOVAs failed to detect a significant main effect of FOV on strategy summary scores or surface scores. Also, the test did not detect a significant interaction between FOV and visual complexity.

Fig. 8 shows strategy scores broken down by level of visual complexity. Strategy adherence was better for participants who trained with higher complexity. The ANOVA found a significant effect of visual complexity on surface scores, with F(2, 36) = 6.076, p = 0.005, and $\eta_p^2 = 0.252$. The Bonferroni-corrected post-hoc test only showed high complexity to be significantly better than low complexity (p = 0.005) with Cohen's d = 1.22, showing a large effect. The ANOVA for strategy summary scores also yielded a significant main effect of complexity on strategy, with F(2, 36) = 5.44, p = 0.009, and $\eta_p^2 = 0.232$. Post-hoc Bonferroni t-tests showed high complexity to have significantly better performance than low (p = 0.015, d = 1.07), and high was significantly better than medium (p = 0.030, d = 0.94).

We also analyzed the effects of the independent variables on strategy surface scores and found similar results as with the summary scores.

We also tested for correlations between target detection performance during training and strategy adherence during assessment. Two-tailed Pearson correlations indicated significant negative correlations between training scores and strategies for both surface scores (r = -0.414 and p = 0.005)

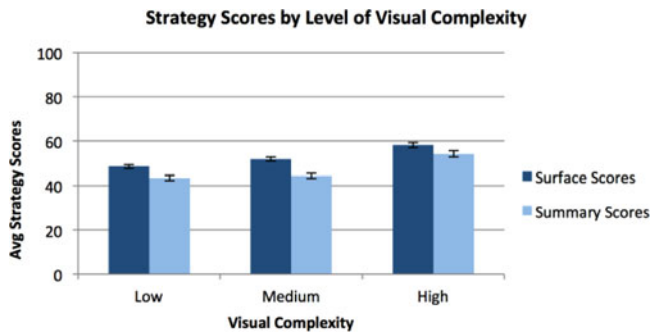

Fig. 8. Mean strategy scores from assessment trials by level of visual complexity in training. Error bars show standard error.

and summary scores ($r = -0.457$ and $p = 0.002$). Participants who found more targets during training demonstrated worse strategies in the assessment phase.

Additionally, we found that strategy ratings failed to predict target detection in the assessment. A one-tailed Pearson's correlation test between strategy surface scores and assessment performance scores did not find a significant correlation, with $r = 0.057$ and $p = 0.356$. Likewise, assessment performance was not significantly correlated with strategy summary scores.

To account for the influence of FOV and complexity, we also compared both types of strategy scores to ranked detection results (i.e., ranked by performance within condition) with Spearman correlations. Again, there was no evidence of correlation.

5 Discussion

The experiment provided interesting insight into the effects of FOV and visual complexity on VR training system effectiveness and resulted in some unexpected findings.

5.1 Effects of FOV

The level of FOV used during training did not have a significant effect on either assessment target detection or assessment strategy usage, so we did not find evidence to support H1 or H2. We did find a highly significant effect of FOV on detection performance during training, with higher FOVs leading to better training trial detection, which supports H3.

Taken together, these results show that while FOV can have a measurable effect on task performance, the size of the FOV during training does not appear to affect strategy learning or training transfer. We believe that FOV affected detection performance during training because a wider FOV allowed users to look ahead, anticipate upcoming parts of the environment, and plan the visual scanning pattern. It may also be that the wider FOV allowed users to notice targets in the periphery and modify the visual scanning pattern to catch them.

It is not clear from our results why the FOV of the training system had no measurable effect on training transfer. Participants who trained with different FOV levels had approximately the same detection performance and strategy transfer scores during the assessment. It is possible that FOV had multiple competing effects. For example, training with a narrow FOV may have helped users focus on the task and the correct strategy, but the much wider view in the assessment environment may have distracted the users, negating these gains. Alternatively, it could be that training with a wide FOV made the training task easier, such that users did not focus enough mental effort on the training, resulting in lower-than-expected scores during assessment. Finally, it may be that our assessment trials were too difficult, washing out any effects of training (more on this below). Future work is needed to examine some of these hypotheses.

5.2 Effects of Visual Complexity

We did not find a significant effect of visual complexity on target detection performance in the assessment phase of the experiment, so H4 was not supported. However, the analysis did find a significant effect of complexity on both strategy transfer and training task performance, supporting hypotheses H5 and H6, respectively.

The ultimate goal of any task-training system is to improve real-world task performance, so we might be tempted to take the lack of support for H4 (effect of visual complexity on assessment target detection) as an indication that the level of visual complexity in the training system is not critical for training transfer. However, we see a more nuanced picture when combining the results for H5 and H6 (effects of visual complexity on strategy adherence and on detection performance in training, respectively). During the training trials, participants scored much higher with the lower levels of complexity; the simpler the environment was, the easier it was to pick out the targets. In the assessment trials, on the other hand, participants who trained with the low and medium levels of complexity demonstrated the worst use of the prescribed visual scanning strategy. We speculate that these participants were not forced to work hard in the training phase—they could score well without following the prescribed strategy, so they did not learn the strategy very well despite constant reinforcement of the strategy by the experimenter. Pure performance is not the only factor of importance to training system designers; learning of correct procedures, strategies, and skills (which are assumed to be critical for good real-world task performance) is also essential.

Thus, our results indicate that training systems for visual scanning and similar tasks should, when possible, use a level of visual complexity that is as close to the real environment as possible in order to ensure good transfer. Strengthening this result is the fact that the posthoc analysis of strategy results did not show a significant benefit of the moderate level of complexity over the low level, which further demonstrates the importance of matching the visual complexity of the training to that of the actual scenario. In this study, a moderate increase of complexity was not sufficient for significantly improving transfer outcomes.

However, visual complexity in our study was determined by an amalgam of factors (e.g., textures, fog, number of objects, geometric complexity). It is probable that different tasks are affected differently by various factors, which is why it is important to investigate the effects and differences under varying circumstances. For the type of visual scanning task used in this study, it could be interesting to further separate elements of visual complex for a deeper understanding of which elements might be most important for the design of training systems.

5.3 Correlations between Training and Assessment Task Performances

A common assumption in training system design is that training system task performance can be used to predict real-world performance. That is, if a trainee performs well in the training system after a certain amount of training, it is assumed that the trainee is well trained and will perform well on the corresponding real-world task. We tested these assumptions about our VR training conditions by considering the high-fidelity assessment phase to serve as the role of the "real" scenario that participants were training for. We correlated training system performance with assessment

performance. Surprisingly, we found no significant correlation. Hence, H7 was not supported.

Combined with the results showing no effects of FOV or visual complexity on assessment detection performance, this suggests that other factors besides the training condition determined participants' performance in the assessment trials. We might speculate that only the FOV and visual complexity of the assessment condition (both at the highest level) determined performance, rather than the FOV and complexity of the training environment. This would be consistent with our finding that the training condition had significant effects on training performance.

We note that the level of performance scores in the assessment were in the 35 to 40 percent range, which is quite a bit lower than the training performance scores, which varied between 42 and 80 percent depending on the condition. It may be that it was simply difficult to perform the visual scanning task in the high visual complexity condition and that this dominated the effects of training condition during the assessment trials. Based on the significant performance improvement over the progression of training trials, we do not attribute the worse assessment performance to fatigue.

In the end, this study does not allow us to draw any conclusions about assessment performance (which was our proxy for real-world task performance) because there are no significant effects of the independent variables on assessment detection performance and no correlations of other measures with assessment detection performance.

On the other hand, we designed our VR training system to teach a specific visual scanning strategy and found that visual complexity had a significant effect on strategy transfer. But we also found that detection performance during the training phase was inversely correlated with strategy use during assessment, showing that training performance was not enough to judge mastery of the scanning strategy. Thus, we reject H8. To explain the observed result, we hypothesize that if participants performed well during training without using the prescribed strategy, then they lacked the incentive to revise the chosen strategy. This result demonstrates that training performance is not always a sufficient indicator of technique. If learning a prescribed strategy (or procedure or skill) is the goal of a training exercise, performance may not be an appropriate indicator of effectiveness.

Additionally, the fact that we did not detect a correlation between strategy adherence and target detection suggests the possibility that the prescribed strategy was not optimal for the scanning task. This possibility should have little bearing on the training transfer results, as the important outcome was that participants were following the instructed method. But poor appropriateness of the scanning strategy for the task could explain deviation from the strategy.

We still think that the sweeping strategy is generally appropriate for the task; the method promoted thorough and consistent scanning of building faces as soon as they were available. However, it is possible that detection performance may have sometimes benefitted from "greedy" deviations from the instructed strategy. For example, we sometimes observed participants departing from the instructed scanning pattern to inspect character models as soon as they were noticed, rather than at the expected point in the sweeping pattern. Despite receiving feedback in the instruction phase about prioritizing strategy adherence, participants may have opted to deviate from the prescribed strategy if they felt they were more successful with other methods. This could be an alternative explanation for the lack of correlation between detection performance and strategy.

6 Conclusions and Future Work

VR training is broadly applicable to a wide variety of domains, so it is important to understand the effects of various VR system characteristics on training effectiveness. In particular, we focus on the effects of fidelity, a fundamental property impacting many design decisions for VR systems. The experiment reported in this paper studied the impact of FOV and visual complexity in the context of a visual scanning task. Both factors influenced task performance during training, and visual complexity of the training condition significantly affected participants' learning of the prescribed scanning strategy.

This research contributes a better understanding of the influence of display fidelity and visual realism in VR training system design. In particular, we conclude that designers of VR training systems should use a high level of visual realism for tasks that involve visual scanning or visual search in visually complex environments. Additionally, we contribute evidence that the direct measurement of learning is a better measure of training effectiveness than raw task performance.

Our future work will continue to explore how the fidelity of VR training systems impacts training effectiveness. We have identified several open questions about FOV and visual complexity above. We will also examine other components of fidelity for visual scanning tasks. For example, an upcoming experiment will investigate whether amplifying the user's head rotations in a training system leads to disorientation and negative training transfer. Finally, we will look at other variants of visual search tasks and at other categories of training tasks.

Acknowledgments

The authors would like to thank Tobias Höllerer, Tao Ni, and Peter Squire for their help and support of this research. They also thank Jim Templeman for the ideas for the visual scanning task, and the anonymous raters who reviewed hundreds of scanning trials. This research was funded by the Immersive Sciences program in the Office of Naval Research. This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05–00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] M. Zyda, "From visual simulation to virtual reality to games," *IEEE Comput.*, vol. 38, no. 9, pp. 25–32, Sep. 2005.

[2] N. E. Seymour, "VR to OR: A review of the evidence that virtual reality simulation improves operating room performance," *World J. Surg.*, vol. 32, no. 2, pp. 182–188, 2008.

[3] R. M. Sorrentino, R. Levy, L. Katz, and X. Peng, "Virtual visualization: Preparation for the olympic games long-track speed skating," *Int. J. Comput. Sci. Sport*, vol. 4, p. 40, 2005.
[4] J. M. Rolfe and K. J. Staples, *Flight Simulation*. New York, NY, USA: Cambridge Univ. Press, 1988.
[5] R. T. Hays and M. J. Singer, "Simulation fidelity as an organizing concept," in *Simulation Fidelity in Training System Design, Recent Research in Psychology*, R. T. Hays and M. J. Singer, Eds. New York, NY, USA: Springer, 1989, pp. 47–75.
[6] W. F. Moroney and B. W. Moroney, "Flight simulation," in *Handbook of Aviation Human Factors*, J. A. Wise, V. D. Hopkin and D. J. Garland, Eds. Mahwak, NJ, USA: Lawrence Erlbaum, 1999, pp. 355–388.
[7] W. Schneider, "Training high-performance skills: Fallacies and guidelines," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 27, no. 3, pp. 285–300, 1985.
[8] D. Druckman and R. A. Bjork, *Learning, Remembering, Believing: Enhancing Human Performance*. New York, NY, USA: National Academies Press, 1994.
[9] M. K. Singley, *The Transfer of Cognitive Skill*. Cambridge, MA, USA: Harvard Univ. Press, 1989.
[10] R. T. Hays and M. J. Singer, "Training effectiveness evaluation," in *Simulation Fidelity in Training System Design, Recent Research in Psychology*. New York, NY, USA: Springer, 1989, pp. 112–159.
[11] J. A. Allen, R. T. Hays, and L. C. Buffardi, "Maintenance training simulator fidelity and individual differences in transfer of training," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 28, no. 5, pp. 497–509, 1986.
[12] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper, "Identifying the perceptual dimensions of visual complexity of scenes," in *Proc. 26th Annu. Meeting Cogn. Sci. Soc.*, 2004, pp. 101–106.
[13] D. A. Bowman, R. P. McMahan, C. Stinson, E. D. Ragan, S. Scerbo, T. Höllerer, C. Lee, and R. Kopper, "Evaluating effectiveness in virtual environments with MR simulation," presented at the Interservice/Ind. Training, Simul. Edu. Conf., Orlando, FL, USA, 2012.
[14] D. A. Bowman and R. P. McMahan, "Virtual Reality: How much immersion is enough?" *IEEE Comput.*, vol. 40, no. 7, pp. 36–43, Jul. 2007.
[15] D. A. Bowman, R. P. McMahan, and E. D. Ragan, "Questioning naturalism in 3D user interfaces," *Commun. ACM*, vol. 55, no. 9, pp. 78–88, 2012.
[16] R. T. Hays, J. W. Jacobs, C. Prince, and E. Salas, "Flight simulator training effectiveness: A meta-analysis," *Military Psychol.*, vol. 4, no. 2, pp. 63–74, 1992.
[17] K. Johnsen, A. Raij, A. Stevens, D. S. Lind, and B. Lok, "The validity of a virtual human experience for interpersonal skills education," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 1049–1058.
[18] S. G. Hart and V. Battiste, "Field test of video game trainer," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 1992, vol. 36, pp. 1291–1295.
[19] B. Dunkin, G. Adrales, K. Apelgren, and J. Mellinger, "Surgical simulation: A current review," *Surgical Endoscopy*, vol. 21, no. 3, pp. 357–366, 2007.
[20] G. Saposnik, M. Mamdani, M. Bayley, K. Thorpe, J. Hall, L. Cohen, and R. Teasell, "Effectiveness of virtual reality exercises in stroke rehabilitation (EVREST): Rationale, design, and protocol of a pilot randomized clinical trial assessing the Wii gaming system," *Int. J. Stroke*, vol. 5, no. 1, pp. 47–51, 2010.
[21] J. McComas, M. MacKay, and J. Pivik, "Effectiveness of virtual reality for teaching pedestrian safety," *Cyber Psychology Behavior*, vol. 5, no. 3, pp. 185–190, 2002.
[22] G. M. Reger, K. M. Holloway, C. Candy, B. O. Rothbaum, J. Difede, A. A. Rizzo, and G. A. Gahm, "Effectiveness of virtual reality exposure therapy for active duty soldiers in a military mental health clinic," *J. Traumatic Stress*, vol. 24, no. 1, pp. 93–96, 2011.
[23] D. R. Baum, S. L. Riedel, R. Hays, and A. Mirabella, "Training effectiveness as a function of training device fidelity," Honeywell Syst. Res. Center, Minneapolis, MN, USA, Tech. Rep. 82SRC37, 1982.
[24] D. Waller, E. Hunt, and D. Knapp, "The transfer of spatial knowledge in virtual environment training," *Presence: Teleoperators Virtual Environ.*, vol. 7, no. 2, pp. 129–143, 1998.
[25] I. E. Sutherland, "The ultimate display," in *Proc. Congr. Internation Federation Inf. Process. Congress*, 1965, pp. 505–508.
[26] V. H. Hiep, R. Keriven, P. Labatut, and J. P. Pons, "Towards high-resolution large-scale multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1430–1437.
[27] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, and P. Debevec, "The digital emily project: Achieving a photorealistic digital actor," *IEEE Comput. Graph. Appl.*, vol. 30, no. 4, pp. 20–31, Jul./Aug. 2010.
[28] E. A. Suma, D. M. Krum, and M. Bolas, "Redirected walking in mixed reality training applications," in *Human Walking in Virtual Environments*, F. Steinicke, Y. Visell, J. Campos, and A. Lécuyer, Eds. New York, NY, USA: Springer, 2013, pp. 319–331.
[29] K. Mania, T. Troscianko, R. Hawkes, and A. Chalmers, "Fidelity metrics for virtual environment simulations based on spatial memory awareness states," *Presence: Teleoperators Virtual Environments*, vol. 12, no. 3, pp. 296–310, 2003.
[30] E. D. Ragan, R. Kopper, P. Schuchardt, and D. A. Bowman, "Studying the effects of stereo, head tracking, and field of regard on a small-scale spatial judgment task," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 5, pp. 886–896, May 2013.
[31] R. P. McMahan, "Exploring the effects of higher-fidelity display and interaction for virtual reality games," PhD dissertation, Dept. Comput. Sci., Virginia Tech, Blacksburg, VA, USA, 2011.
[32] R. P. McMahan, D. A. Bowman, D. J. Zielinski, and R. B. Brady, "Evaluating display fidelity and interaction fidelity in a virtual reality game," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 626–633, Apr. 2012.
[33] S. Lessels and R. A. Ruddle, "Movement around real and virtual cluttered environments," *Presence: Teleoperators Virtual Environ.*, vol. 14, no. 5, pp. 580–596, 2005.
[34] R. Pausch, D. Proffitt, and G. Williams, "Quantifying immersion in virtual reality," in *Proc. 24th Annu. Conf. Comp. Graph. Interactive Techn.*, 1997, pp. 13–18.
[35] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman, "The effects of visual realism on search tasks in mixed reality simulation," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 4, pp. 547–556, Apr. 2013.
[36] L. Arns, D. Cook, and C. Cruz-Neira, "The benefits of statistical visualization in an immersive environment," *in Proc. IEEE Virtual Reality*, Houston, TX, USA, 1999, pp. 88–95.
[37] F. Bacim, E. Ragan, S. Scerbo, N. F. Polys, M. Setareh, and B. D. Jones, "The effects of display fidelity, visual complexity, and task scope on spatial understanding of 3D graphs," in *Proc. Graph. Interface Conf.*, 2013, pp. 25–32.
[38] K. Mania, D. Wooldridge, M. Coxon, and A. Robinson, "The effect of visual and interaction fidelity on spatial cognition in immersive virtual environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 3, pp. 396–404, May/Jun. 2006.
[39] W. B. Thompson, P. Willemsen, A. A. Gooch, S. H. Creem-Regehr, J. M. Loomis, and A. C. Beall, "Does the quality of the computer graphics matter when judging distances in visually immersive environments?" *Presence: Teleoperators Virtual Environ.*, vol. 13, no. 5, pp. 560–571, 2004.
[40] P. B. Kline and B. G. Witmer, "Distance perception in virtual environments: Effects of field of view and surface texture at near distances," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 1996, vol. 40, pp. 1112–1116.
[41] A. Toet, S. E. Jansen, and N. J. Delleman, "Effects of field-of-view restrictions on speed and accuracy of manoeuvring 1, 2, 3," *Perceptual Motor Skills*, vol. 105, no. 3f, pp. 1245–1256, 2007.
[42] K. L. Yeung and L. Li, "Effect of the field of view on perceiving world-referenced image motion during concurrent head movements," *Displays*, vol. 34, pp. 165–170, 2012.
[43] R. Kopper, C. Stinson, and D. Bowman, "Towards an understanding of the effects of amplified head rotations," in *Proc. EEE VR Workshop Perceptual Illusions Virtual Environ.*, 2011, pp. 10–15.
[44] C. Lee, S. Bonebrake, T. Höllerer, and D. A. Bowman, "The role of latency in the validity of AR simulation," in *Proc. IEEE Virtual Reality Conf.*, 2010, pp. 11–18.
[45] B. Laha, D. A. Bowman, and J. D. Schiffbauer, "Validation of the MR simulation approach for evaluating the effects of immersion on visual analysis of volume data," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 4, pp. 529–538, Apr. 2013.
[46] S. Ishihara, *A Series of Plates Designed as a Test for Colour-Blindness*. Tokyo, Japan: Kanehara Shuppan, 1960.
[47] S. E. Stemler and J. Tsai, "Best practices in interrater reliability: Three common approaches," in *Best Practices in Quantitative Methods*, J. Osborne, Ed. Newbury Park, CA, USA: Sage Publications, 2008, pp. 29–49.

[48] J. M. LeBreton, J. R. Burgess, R. B. Kaiser, E. K. Atchley, and L. R. James, "The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar?," *Org. Res. Methods*, vol. 6, no. 1, pp. 80–128, 2003.
[49] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bull.*, vol. 86, no. 2, pp. 420–428, 1979.
[50] J. Kottner, L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner, "Guidelines for reporting reliability and agreement studies (GRRAS) were proposed," *Int. J. Nursing Studies*, vol. 48, no. 6, pp. 661–671, 2011.

Eric D. Ragan received the PhD degree in computer science from Virginia Tech. He is a research scientist at Oak Ridge National Laboratory. He is a member of the IEEE Computer Society.

Doug A. Bowman received the PhD degree in computer science from the Georgia Institute of Technology. He is a professor of computer science and director of the Center for Human-Computer Interaction at Virginia Tech. He is a member of the IEEE Computer Society.

Regis Kopper received the PhD degree in computer science from Virginia Tech. He is a research scientist and director of the Duke immersive Virtual Environment at Duke University. He is a member of the IEEE and the IEEE Computer Society.

Cheryl Stinson received the MS degree in computer science from Virginia Tech. She is a web developer at Precision Nutrition.

Siroberto Scerbo is currently working toward the PhD degree in the Department of Computer Science and the Center for Human-Computer Interaction at Virginia Tech.

Ryan P. McMahan received the PhD degree in computer science from Virginia Tech. He is an assistant professor of computer science and of arts & technology at the University of Texas at Dallas. He is a member of the IEEE and the IEEE Computer Society.