An Evaluation of Immersive Displays for Virtual Human Experiences

Kyle Johnsen[1], Benjamin Lok[2]
Department of Computer and Information Science and Engineering
University of Florida

ABSTRACT

This paper compares a large-screen display to a non-stereo head-mounted display (HMD) for a virtual human (VH) experience. As VH experiences are increasingly being applied to training, it is important to understand the effect of immersive displays on user interaction with VHs. Results are reported from a user study (n=27) of 10 minute human-VH interactions in a VH experience which allows medical students to practice communication skills with VH patients. Results showed that student self-ratings of empathy, a critical doctor-patient communication skill, were significantly higher in the HMD; however, when compared to observations of student behavior, students using the large-screen display were able to more accurately reflect on their use of empathy. More work is necessary to understand why the HMD inhibits students' ability to self-reflect on their use of empathy.

Keywords: virtual humans, embodied agents, display comparison, medical education, immersive virtual environments

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism Virtual Reality. H.5.2 [Information Systems and Presentation]: User Interfaces – Evaluation

1 INTRODUCTION

Rapidly gaining in popularity, ***virtual human (VH) experiences*** are virtual reality systems that focus on providing immersive social experiences. VH experiences are being used for many applications including military leadership training [1], social phobia treatment[2], and health professions training [3-5]. In VH experiences, the user's primary task is to interact with one or more VHs. We compare two very different immersive visual displays, a head-mounted display (HMD) and a large-screen display. The choice of these two displays was made because they 1) are commonly used in immersive virtual reality, 2) are widely available, 3) have similar infrastructure and monetary costs, and most importantly 4) enable close-up, life-size VHs. Results are reported from a user study, which evaluated the displays' impact on user perception and behavior in the VH experience.

Display systems may influence users based on their characteristics. Physical display characteristics (e.g. size, weight, wires, and tracking equipment) affect what the user can do in the experience. Intrinsic display characteristics (e.g. resolution, field-of-view, brightness, head-tracking, and stereoscopy) affect what the user perceives. There are many examples in the general virtual reality literature where task performance is impacted by display choice, e.g. [6-9]. In a VH experience, the impact tends to be behavioral or perceptual. Zanbaka et. al. point out that the smaller field-of-view of an HMD allows users to easily ignore a VH by turning their heads slightly away from the VH [10]. When the VH is embedded into the real-world on a large-screen display, they are part of the users' larger field-of-view. Also, immersion has a powerful behavioral effect on user behavior in H-VH interactions. In [11], an immersive display amplified user anxiety when speaking to an audience of VHs.

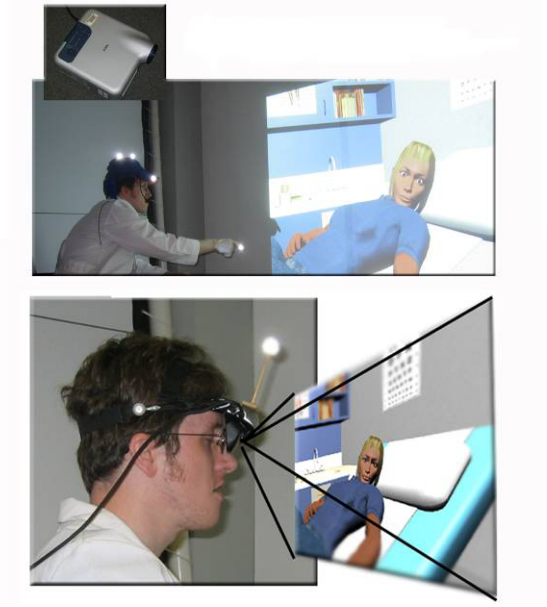

Figure 1. The user interacts with a close, life-size VH using either a large-screen display (top) or HMD (bottom)

This work brings immersive displays research together with an important real-world application, using a VH experience for medical communication skills training. For conducting medical interviews, medical students must develop strong doctor-patient communication skills. Communication skills include the sequence and type of questions, listening, and expressing empathy. A VH experience was constructed that allows students to practice and be evaluated on these skills, called the Interpersonal Simulator (IPS) [4]. The experience has been used by hundreds of medical students, and has been validated for doctor-patient communication skills education [12]. The motivation behind this work is to understand if the choice of immersive visual display can impact students' perception and/or behavior in the VH experience.

[1]kjohnsen@cise.ufl.edu [2]lok@cise.ufl.edu

2 THE INTERPERSONAL SIMULATOR (IPS)

The IPS combines immersive displays, tracking, natural speech processing, and realistic VHs to produce an engaging VH experience. Three networked computers were used to run the IPS during the study. One ran speech recognition and understanding (Intel Pentium 4, 3.4 GHz), another performed rendering and animation (Intel Core 2, 2 GHz, NVidia GeForce 7900), and the last was used for the optical tracking system (Intel Core 2, 2 GHz). The frame rate for the IPS was consistently greater than 60 frames per second.

2.1 Immersive Displays

The IPS can be switched between two immersive displays, a non-stereo head-mounted display (HMD), and a fish-tank projection display (FTPD). **Note**: *Ideally, both of these displays would support stereo viewing; however, the FTPD does not support stereo at this time. Thus the HMD was only used in mono mode (the same image for both eyes). However, subjectively the difference between stereo and mono for the HMD is small for the interview task.*

The FTPD is a type of large-screen display; "fish-tank" because the displayed image is rendered from the perspective of the user as in [13]. To produce the large screen it uses a projector. The projector for the study was an NEC LT260. The resolution was set to 800x600 to match the native resolution of the HMD. In this way, the virtual world and the VH are brought to the user, rather than the user being brought to the VH as in the HMD.

The user's head position is tracked using a custom built optical tracker similar to the one described in [14]. The update rate for this tracker is 60 Hz. The tracking accuracy is approximately 1 cm for the examination room space. The jitter is less than 0.5 mm, and the precision is variable depending on marker size, but is less than 1 mm in practice. The end-to-end latency is less than 100ms.

Features for this type of display relative to the HMD are:

- The user is not encumbered by wires or the display
- The field-of-view is on average larger than the HMD (60-100° diagonal)
- The user can see their own body
- The user can interact with real-world objects easily (e.g note taking)

The HMD used was an eMagin Z800 3D Visor. By tracking the orientation and position of the HMD and coupling them with rendered viewing direction and position, the user is given the impression that they are in the virtual world, and can look in any direction. The Z800 is low-cost (<$1500 USD) for a reasonable field-of-view (40° diagonal) and pixel density (800x600). The Z800 also includes an integrated 3 degree of freedom orientation sensor, which has a 33 Hz update rate, 1 degree accuracy, and little subjectively noticeable jitter. As with the FTPD, the position of the user's head was tracked using the optical tracker.

Features for this display relative to the FTPD are:

- The display is always in front of the user—the user can look in any direction and the virtual world will be visible
- The user is isolated from the real environment—although not necessarily an advantage, it can be if the real environment contains visual distractions.

2.2 Speech Processing

In face-to-face interactions, speaking is the natural method of communicating information. In the IPS, a wireless headset microphone is used for speech input. Speech recognition software, The VH understands natural spoken language, so the user does not need to learn a command set. Also, the user is not given a list of options to choose from; they must come up with questions on their own.

Dragon Naturally Speaking 9.0 (www.nuance.com) translates the speech into text. The translated text is displayed to the user to reduce user frustration when the system does not recognize speech accurately. The accuracy of the speech recognition varies widely based on the user, microphone, and environmental conditions; however, within the limited domain of medical interview skills training, the average speech accuracy has been approximately 60%, with some users achieving as high as 90%, and some users perceiving as high as 100% [15].

2.3 Virtual Humans

The VH bodies are modeled and rigged (binding of the skeleton to the model) using Autodesk's Maya 8.0. The head model and facial expressions are created using Di-o-matic's Facial Studio. The Object-oriented Graphics Rendering Engine (OGRE) is used to render and animate the VHs and surrounding scene. The VHs' voices are the recorded audio of hired actors.

Each VH has a script that defines its responses to user input. The two VHs used in this study were DIANA (Digital Animated Avatar) and EDNA (Elderly Diana). DIANA simulates a patient who is complaining of abdominal pain. EDNA simulates a patient who is worried because she found a mass in her breast.

3 USER STUDY

A within-subjects comparison of the displays in a pilot study with four third-year medical students suggested that students preferred the HMD over the FTPD. Reasons included a feeling of closeness to the patient in the HMD and a feeling of being *inside* the virtual examination room. T*o* follow up on results from the pilot study, a larger (n=27) user study[1] was conducted to quantify the difference between the displays. This study was designed with display system (HMD and FTPD) as a between-subjects variable, as opposed to a within-subjects variable in the pilot study. This was to remove any bias that may have resulted from user preference for one display system over the other. The goal for the study was to determine if there are significant perceptual, behavioral, or performance differences in the VH experience under different immersive display conditions.

3.1 Population

For the study, students were actively recruited by the experimenters from the student population of the Medical College of Georgia. Twenty seven students volunteered to participate in the study. Twenty three (13 male, 10 female) of the students were in their third year of medical school and four (2 male, 2 female) were in their 1st year of physician's assistant school. Each student was compensated $20 (USD) for their time. Group assignment was pseudo-random to ensure that each group had a similar gender distribution (HMD-8M, 5F and FTPD-7M, 7F).

[1] This study was conducted alongside studies on racial diversity, physiological measures, and post-experience review. The post-experience visualization study does not conflict with this study because students did not know they were visualizing their experiences, and all visualization took place after measures were taken. Physiological measures were taken for all students in the same manner. We assume this method is independent from the display system. The racial diversity study directly conflicts with this study, but was controlled for by group assignment. No significant interaction effects were observed between display and race on any measures.

3.2 Procedure and Measures

Pre-experience: The study was conducted in the surgical skills laboratory at the Medical College of Georgia over the course of four days. Students began the experiment by filling out background surveys used to collect demographics, game-playing and prior patient interview experience. Students also self-rated their skill on various dimensions (e.g. expressing empathy, conducting a thorough interview) of the medical interview.

Experience: Students were then outfitted with a wireless microphone, and performed a one minute acoustic adjustment to improve speech recognition performance. Students in the HMD condition were then instructed to put on the HMD and shown how to adjust it. Students in the FTPD condition were instructed to put on the hat used for tracking the student's head (the HMD was tracked in the HMD condition). A two minute tutorial was then conducted by the experimenter where the student was taught how to identify when speech recognition or speech understanding errors were preventing the patient from understanding the student's speech. The student was then given an opportunity to practice interviewing DIANA, and ask the experimenter questions, until they felt comfortable with the system. Immediately upon completion of the practice interview, the student filled out a paper survey which asked for their diagnosis and treatment plan for DIANA. The student then re-entered the examination room and performed an unassisted patient interview of EDNA. A ten minute time limit was enforced for this interview.

Post-experience: Following the interviews, the student was escorted back to the survey area to fill out the post-experience surveys. The post-experience surveys included:

- The Maastricht Assessment of Simulated Patients (MaSP), -a validated survey of simulated patient quality [16].
- The SUS presence survey [17]
- A copresence survey derived from [18]
- A self-evaluation survey used to rate their interviews designed by medical faculty at the Medical College of Georgia.

Grading and Behavioral coding: Students were graded on their differential diagnosis and treatment plan by a third year medical resident for quality on a scale from 1 (very poor) to 5 (very good). Behavioral data were collected for each interview in the form of transcripts and video. The transcripts were used to determine if there were global differences in what students said to the VH. To give a local view of behavior, video of students' responses to the predefined moments in the interviews described below were coded.

- **Sneeze** - EDNA sneezed two minutes after each student began to interview her. We qualified each response as none, conscious, or unconscious based on the amount of time students took to respond. Both verbal (e.g. "bless you") and non-verbal (e.g. head-jerk) responses are considered. A conscious response takes a few seconds. An unconscious response is immediate.
- **Empathetic Opportunity** - EDNA asks in the interview "could this be cancer?" Four external observers quantified each student's response to this empathetic opportunity on a scale from 1 (not empathetic at all) to 7 (very empathetic).

4 Results

Data was primarily analyzed using ANOVAs with display condition as the main factor and controlling for gender. Gender was included in all analyses because it has been shown to have a strong effect in other VH studies [19]. Covariables, such as speech accuracy and self-skill ratings were included in the analyses when a significant correlation was observed with the dependent variable.

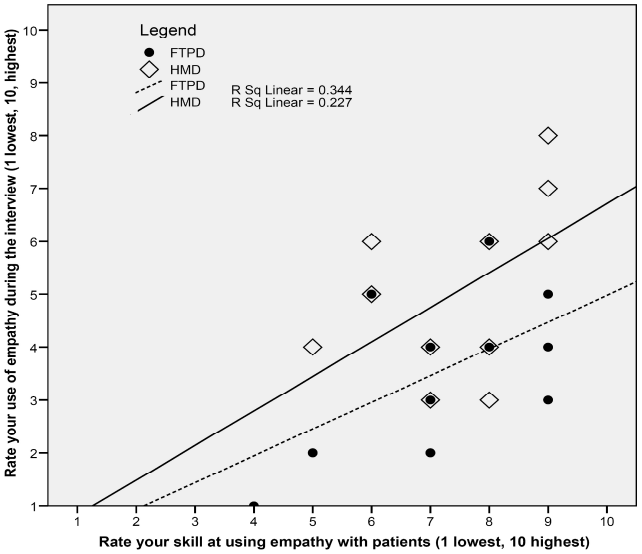

Figure 2. Plot of students' self-rating of empathy rating prior to their interview versus students' self-rating of empathy after the interview

Background: There was a significant difference in backgrounds between the HMD and FTPD groups on the number of years in medical school. This was because the 4 physician's assistant students were all assigned to the FTPD group (an unfortunate mistake in study design). However, analyses conducted with and without these students did not change the significance of any results.

Presence and Copresence: There only survey item with a significant effect ($p<.05$) of display condition was , "I had a sense of 'being there' in the virtual examination room". The HMD group ($M=4.92$, $SD=1.498$) had a higher score than the FTPD group ($M=3.57$, $SD=1.651$). An explanation for this is that the HMD brings the user into the virtual examination room with the VH patient. The FTPD brings the VH patient into the real examination room.

Patient Quality: The only observed difference on the MaSP survey was that the FTPD group felt that "the appearance of the patient fit the role" better than the HMD group. This difference can be explained in that the HMD group could see the patient at a higher spatial resolution (given the lower field-of-view of the HMD). Speech understanding turned out to be the dominant variable in patient ratings, with a significant effect ($p<.01$). This was mostly evident in the overall rating of the patient ($r(25)=-.358$, $p=.07$)) and the rating of the system as a "worthwhile educational experience" ($r(25)=-.439$, $p<.05$).

Self-Ratings: When pre-interview skill ratings were included as a covariable, a significant ($p<.05$) effect of display system was found in students' self-rating of their empathy after their interviews The scatter plot shown in Figure 2 shows a clear trend that students in the HMD group ($M=5.15$, $SD=1.82$) rated their use of empathy higher than students in the FTPD group ($M=3.64$, $SD=1.34$).

Global Behavior: We did not observe an effect of display condition on global behavioral measures taken from the transcript data. Both display system groups conducted similar interviews of EDNA, taking many conversational turns ($M=52.44$, $SD=20.47$) during the length of their interviews ($M=395$s, $SD=195$s). The majority of time, students asked the patient questions, accounting

for 64% of all speech. Repeated (exactly the same words) and rephrased (different words, same meaning) utterances accounted for 23%. The remaining utterances (13%) consisted of a relatively equal distribution of informative statements ("I'm going to ask you some questions now"), confirmations ("okay"), summarizations ("you found a mass in your breast and you're really nervous about it", and empathetic remarks ("I understand that you're scared").

Local Behavior: There was a trend (p=.09) towards a difference in how students respond (verbally or non-verbally) to EDNA sneezing. Out of 13 students in the HMD group, 11 students in the HMD group had no response, 1 had a delayed response, and 1 had an immediate response. Out of 14 students in the FTPD group, 7 had no response, 4 had a delayed response, and 3 had an immediate response. Contrary to the general lack of responses to EDNA's sneeze, most students (10 out of 13 in the HMD group, 9 out of 14 in the FTPD group) responded to the empathetic opportunity. Furthermore, analysis of the results from the external coders' ratings of the empathetic opportunity (reliability = .92) did not find a significant difference between the HMD (M=3.25, SD=1.80) and FTPD (M=3.44, SD=1.91) groups.

Performance: No difference was observed in the grades given to students for their diagnosis or treatment plans for either patient. This is expected given the similar global behavior observed.

5 Discussion and Future Work

In this work, we sought to identify the impact of two immersive visual display systems, a large-screen display (FTPD) and a non-stereo head-mounted display (HMD), on user perception, behavior and performance in a VH experience.

The primary finding for this study is that there is a significant difference in students' perception of their own empathy. Relative to their pre-interview empathetic skills rating, all students rated themselves the same or lower after the experiment ($p < .001$). This effect was greater in the FTPD condition (M=3.71 points lower, SD=1.32) than the HMD condition (M=2.46 points lower, SD=1.46). Correlating the students' rating of their empathy to the coders' rating of the students' empathy, it was found that the HMD group had no correlation ($r(11)= 0.000$), but the FTPD group had a medium correlation ($r(12)=.46$, $p=.08$). It appears that students in the FTPD condition were able to more accurately and consistently reflect on their own interviews. One hypothesis is that the novelty and encumbering nature (weight and wires) of the HMD is an attention draw away from the VH, causing some students to inaccurately reflect upon their interview, and rate themselves higher than they actually deserved. This is alarming from an interview skills training perspective. Medical students must be able to accurately self-reflect on their interviews because they currently receive little feedback from instructors. Therefore, the FTPD is likely superior to the HMD for the current domain.

Overall, user behavior was powerful and natural (especially during the empathetic opportunity) in both immersive display groups. An important line of future work is to compare the *level* of visual immersion for VH experiences. A monitor display is available for little or no cost to every medical student. Significant educational benefit must be found to justify the added infrastructure and monetary cost of using immersive displays.

6 Acknowledgements

D. Scott Lind, MD and Adeline Deladisma, MD performed recruiting of participants and provided the space for the study. Andrew Raij, Brent Rossen, and Aaron Kotranza helped conduct and analyze the results. Funding was providing by a grant from the Medical College of Georgia.

References

[1] R. Hill, J. Gratch, S. Marsella, J. Rickel, W. Swartout, and D. Traum. Virtual Humans in the Mission Rehearsal Exercise System. *Künstliche Intelligenz,* vol. 4, pp. 5-10, 2003.

[2] D. Pertaub, M. Slater, and C. Barker. An experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence: Teleoperators & Virtual Environments,* vol. 11, pp. 68-78, 2002.

[3] R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West. The Virtual Standardized Patient: Simulated Patient-Practitioner Dialog for Patient Interview Training. *Studies in Health Technology and Informatics,* vol. 70, pp. 133-138, 2000.

[4] K. Johnsen, R. Dickerson, A. Raij, B. Lok, J. Jackson, M. Shin, J. Hernandez, A. Stevens, and D. S. Lind. Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills. in Proc. of *IEEE Virtual Reality*, 2005, pp. 179-186, 324.

[5] M. Ponder, B. Herbelin, T. Molet, S. Scherteneib, B. Ulicny, G. Papagiannakis, N. Magnenat-Thalmann, and D. Thalmann. Interactive Scenario Immersion: Health Emergency Decision Training in the JUST Project,. in Proc. of *VRMHR 2002*, 2002.

[6] R. Pausch, M. A. Shackelford, and D. Proffitt. A user study comparing head-mounted and stationary displays. in Proc. of *IEEE Symposium on Virtual Reality*, 1993, pp. 41-45.

[7] D. Bowman, A. Datey, Y. Ryu, U. Farooq, and O. Vasnaik. Empirical Comparison of Human Behavior and Performance with Different Display Devices for Virtual Environments. in Proc. of *Human Factors and Ergonomics Society*, 2002, pp. 2134-2138.

[8] J. E. Swan, Gabbard, II, J. L. Hix, D. Schulman, and R. Kim. A comparative study of user performance in a map-based virtual environment. *IEEE Virtual Reality,* pp. 259-266, 2003.

[9] E. Patrick, D. Cosgrove, A. Slavkovic, J. A. Rode, T. Verratti, and G. Chiselko. Using a large projection screen as an alternative to head-mounted displays for virtual environments. in Proc. of *ACM SIGCHI*, 2000, pp. 478-485.

[10] C. A. Zanbaka, A. C. Ulinski, P. Goolkasian, and L. F. Hodges, Social responses to virtual humans: implications for future interface design. in *ACM SIGCHI*, 2007.

[11] M. Slater, D. P. Pertaub, and A. Steed. Public speaking in virtual reality: facing an audience of avatars. *IEEE Computer Graphics and Applications,* vol. 19, pp. 6-9, 1999.

[12] K. Johnsen, A. Raij, A. Stevens, D. S. Lind, and B. Lok. The Validity of a Virtual Human System for Interpersonal Skills Education. in Proc. of *ACM SIGCHI*, 2007.

[13] C. Ware, K. Arthur, and K. S. Booth. Fish tank virtual reality. in Proc. of *ACM SIGCHI*, 1993, pp. 37-42.

[14] B. Schwald. A Tracking Algorithm for Rigid Point-Based Marker Models. in Proc. of *WSCG'2005*, 2005.

[15] R. Dickerson, K. Johnsen, A. Raij, B. Lok, J. Hernandez, A. Stevens, and D. S. Lind. Evaluating a Script-Based Approach to Simulating Patient-Doctor Interaction. in Proc. of *Proceedings of SCS 2005 International Conference on Human-Computer Interface Advances for Modeling and Simulating*, 2005, pp. 79-84.

[16] L. A. Wind, J. v. Dalen, A. M. M. Muijtjens, and J.-J. Rethans. Assessing Simulated Patients in an Educational Setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education,* vol. 38, pp. 39-44, 2004.

[17] M. Usoh, E. Catena, S. Arman, and M. Slater. Using Presence Questionnaires in Reality. *Presence: Teleoperators & Virtual Environments,* vol. 9, pp. 497-503, 2000.

[18] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators & Virtual Environments,* vol. 14, pp. 379-393, 2005.

[19] C. Zanbaka, P. Goolkasian, and L. Hodges. Can a Virtual Cat Persuade You?: The Role of Gender and Realism in Speaker Persuasiveness. in Proc. of *ACM SIGCHI*, 2006, pp. 1153 - 1162.