EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2016)
C. E. Catalano and L. De Luca (Editors)

3D in-world Telepresence With Camera-Tracked Gestural Interaction

Erik Malcolm Champion [1], Li Qiang[2], Demetrius Lacet[3] and Andrew Dekker[4]

[1]'MCCA, CIC and AAPI, Curtin University, [2]Shenyang Aerospace University, [3]University Center of João Pessoa–Brazil (UNIPÊ),
[4]ITEE, University of Queensland|

Abstract

While many education institutes use Skype, Google Chat or other commercial video-conferencing applications, these applications are not suitable for presenting architectural or urban design or archaeological information, as they don't integrate the presenter with interactive 3D media. Nor do they allow spatial or component-based interaction controlled by the presenter in a natural and intuitive manner, without needing to sit or stoop over a mouse or keyboard. A third feature that would be very useful is to mirror the presenter's gestures and actions so that the presenter does not have to try to face both audience and screen.

To meet these demands we developed a prototype camera-tracking application using a Kinect camera sensor and multi-camera Unity windows for teleconferencing that required the presentation of interactive 3D content along with the speaker (or an avatar that mirrored the gestures of the speaker). Cheaply available commercial software and hardware but coupled with a large display screen (in this case an 8 meter wide curved screen) allows participants to have their gestures, movements and group behavior fed into the virtual environment either directly or indirectly. Allowing speakers to present 3D virtual worlds remotely located audiences while appearing to be inside virtual worlds has immediate practical uses for teaching and long-distance collaboration.

Categories and Subject Descriptors (according to ACM CCS): H.5.1 [Multimedia Information Systems]. Animations, Artificial, augmented, and virtual realities, Experimentation, Human Factors.

1. Introduction

A more immersive experience in presenting digital archaeological simulations would allow the audience to interact with the presenter *inside* the digitally simulated environment, and without forcing the presenter to use thematically inappropriate peripherals such as mouse or keyboard to navigate around or select objects inside the simulated environment. Participants' gestures, movements and group behavior could be fed into the virtual environment either directly or indirectly in order for presenters to present 3D virtual worlds to remotely located audiences while appearing to be inside those virtual worlds [AM99].

Publications suggest this is of great usefulness to the fields of architecture, urban design and archaeology [AM99] [C03] [PCJ*02] and to learning in general [AG13]. The prototype described here could be configured for a wide range of accessible camera-tracking devices and other interface technology as well as cheaply available commercial software and hardware. Although the initial project was specifically for presenting to remote audiences information about archaeological simulations, it could be deployed in classroom or in general conferences to better spatially or interactively integrate the speaker with the displayed digital content.

2. The Research Problem

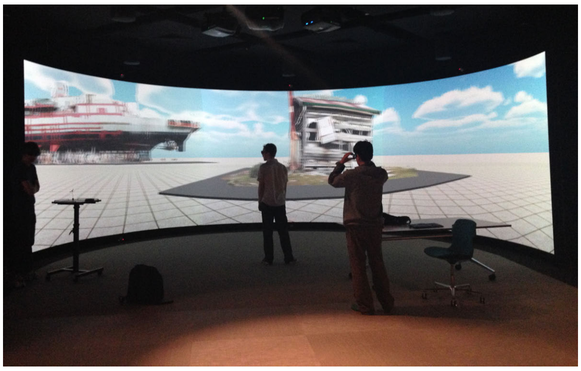

Figure 1: *Cylindrical stereo display, the HIVE, Curtin University.*

Despite the above literature and other related work [ST13a] [ST13b] [GDT*14] [PFS*15], there is so far no effective way to combine 3D models and video conferencing particularly for large scaled cylindrical displays such as the stereo display (Figure 1) that was used for the prototype. This 8 meter wide display is nearly perfectly semi-circular and due to its three-part projection systems could allow several panes of teleconferencing windows, with the avatars on the outside

DOI: 10.2312/gch.20161394

windows able to see each other as well as the presently located person and aspects of the work that they are presenting. The primary software is Unity and MiddleVR (for Unity) but other 3D real-time rendering engines can be employed.

Camera tracking can be deployed via Kinect (Xbox 360 or Kinect One), or via proprietary software (Optitrack) and we have the use of customized wand and helmet-based sensors that can also be added to other types of props or clothing, but the navigation requires some degree of expertise (experience and understanding of 3D space).

An added issue in that in the current setup the sweet spot for viewing (and for stereo vision) is at the centre of the display, where speakers typically stand to present. Either the speaker faces the audience with his or her back to the display (and hence cannot direct navigation in the virtual projected environment), or the speaker faces the display to navigate in the virtual scene and the audience can no longer see their face (and their voice is harder to hear).

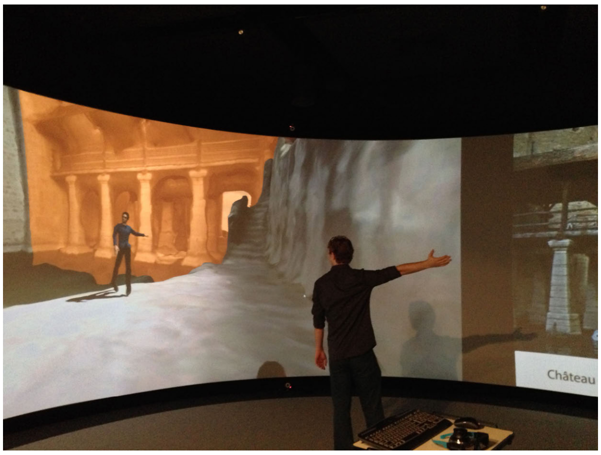

Figure 2: *The avatar mirrors the speaker's tracked gestures, triggering slides by pointing at the relevant object.*

Our hypothesis was that either

- The speaker would prefer to stand to the side of the display and navigate using an avatar in the scene that would control the navigation and display of objects in the environment based on the mirrored gestural motions of the speaker (Figure 2) OR
- The speaker would prefer to be standing to the side of the screen looking between the screen and the audience (i.e. at an angle to both) and their body would be blue-screened into the virtual environment.

For both hypotheses we predicted that the displayed content would appear both more immersive and more intuitive for non-expert users, from the point of view of both speakers, the audience and a physically non-present audience.

3. Significance

The prototype allows presenters of 3D digital environments to insert a streaming live narrator (or a pre-rendered movie of a narrator) into the environment and control aspects of the environment remotely via camera tracking using the Microsoft Kinect. Longer-term this could be extended to help participant movement and eye direction/eye gazing in head mounted displays.

4. Aims

Our aim was to develop and evaluate methods to help communicate the 'inhabited' and spatial sense of archaeolgical and architecturally-related information. Our first objective was to greenscreen the narrator into a 3D environment. Secondly, we wanted to control an avatar in the virtual environment using the speaker's gestures either via a mirrored in-scene avatar (Figure 2) or via a hand icon (Figure 3).

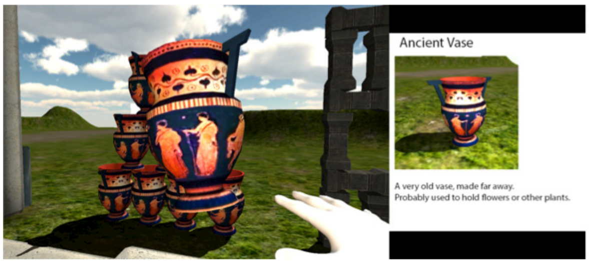
Ancient Vase
A very old vase, made far away.
Probably used to hold flowers or other plants.

Figure 3: *Another option is to simply have a hand which points to objects in the scene, the virtual hand moves and points according to the tracked hand of the speaker.*

The third objective was to trigger slides and movies inside a UNITY environment via speaker finger-pointing. Ideally the speaker could also change the chronology of built scene with gestures (or voice), could alter components or aspects of buildings, move or replace parts or components of the environment. A Leap controller could also be employed to control objects in the environment or navigate between slides, but this constrains the presenter's hand or hands to a small enough area for the Leap controller to detect.

The fourth objective was to better employ the 8 meter wide curved screen (which can also display stereo content) so that participants could communicate with each other without visual obstruction or distraction created by conventional peripherals (such as keyboard or mousse).

4.1 Task I: Pointing and Integrated Slides

The project involved exploring and developing possible methods of connecting motion control to interactive presentations on alternative displays. Displaying research data between academics or to the general public is usually through linear presentations, either timed or stepped through by a presenter. By the use of motion tracking and gestures, presenters can provide a more engaging experience to their audience, as they won't have to rely on prepared static media, timing, or 'mousing' around.

Fields such as archaeology, architecture, or urban design would benefit from being able to take the audience on a non-linear course through what they are presenting, or enable presenters to freely manipulate objects or environments within their presentations. However, game design courses could also employ this prototype to point out and change level design by using natural motions on a large display screen. The prototype could allow a remote presenter to be displayed within their data or virtual environment, allowing them to interact with their data during a presentation, providing a more immersive viewing experience.

There are other possible applications in museums where a display can allow participants to engage in discovering more about an exhibit through motion control. The project aims to investigate possibilities in hardware and software to identify how the different technologies interact and their application, as well as providing a base for future, possibly more specialized, development.

4.2 Task II: Avatar Mirror-Controlled By Speaker

An alternative to the speaker controlling the movement and gestures directly is for their gestures to be tracked and mirrored by a Non Playing Character (NPC) as shown earlier in Figures 2 and 3. By transferring the speaker's gestures and movements to a NPC the audience can concentrate on the objects and events in-scene. Also the speaker can play different roles, constrained only by imagination and technology. We can also add some interesting effects. For example the appearance of the NPC could relate to the gestures or objects visited and triggered by the speaker or by the number of slides that the speaker and audience find and take the time to read.

4.3 Technical Constraints

The commercial company Zigfu gave us access to the Zigfu Development Kit http://zigfu.com/ which allows developers to create cross-platform, motion-controlled apps with the Kinect and other 3D sensors in HTML5/JavaScript, Unity3D and Flash. The ZDK allowed us to easily connect the Kinect up to different types of environments and it also tracked the skeletons of multiple players simultaneously (three people could be tracked but the tracking was unreliable).

One possible limitation is that for developing with Zigfu one also needs to buy Unity Pro. For our prototype we can use Kinect 360 or Kinect One (although the Kinect camera version 2 requires Windows 8 and Direct X11, it is more accurate and powerful than the older Kinect 360). Please also note that while this prototype was developed specifically for a 180 degree, 8-metre diameter cylindrical stereo screen, the prototype can also be used for HMDs (Head Mounted Display), the web and other display devices.

5. FUTURE WORK

Aware that technology can distract from the intended learning objectives [DE15], we plan to review and develop a robust methodology and heuristics for presenting a narrator or conference presenters via a 2D real-time or pre-rendered movie inside a 360- degree panorama or 3D virtual environment, via an HTML webpage or directly inside a virtual environment on a curved or tiled display and on a HMD. An XBOX Kinect version 2 camera is programmed to track gestures that can be fed into both 3D virtual environments and 360-degree panoramas into which movies have been inserted.

Demetrius Lacet created video panoramas [http://www.onzeonze.com.br/blog360/toursaofrancisco/index.html] which can be viewed on Google Cardboard VR or Samsung Gear VR (Figure 4). The case study and example shown was of Brazilian baroque churches and the community was trained to create these panoramas to both develop cultural skills and to increase their awareness of the dangers of vandalism (Figure 5). The original technology of the pre-rendered video and panorama ran on HTML using JavaScript and three.js. Via the Kinect or other sensors, we can remove the background of a presenter in real-time and we would like to next work out how to stream panorama and live avatar together via the Internet, while allowing the presenter (or avatar) to trigger objects in the panorama by pointing at them.

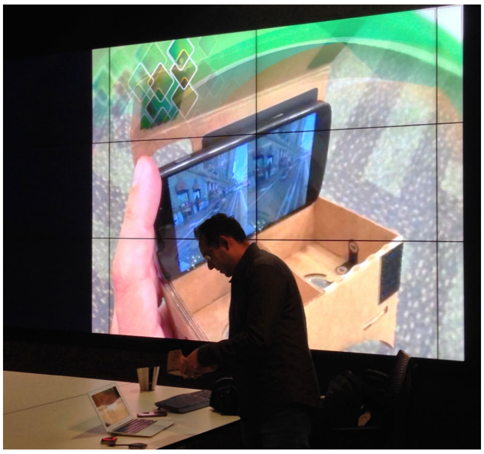

Figure 4: *Low-cost HMDs can run video-panoramas.*

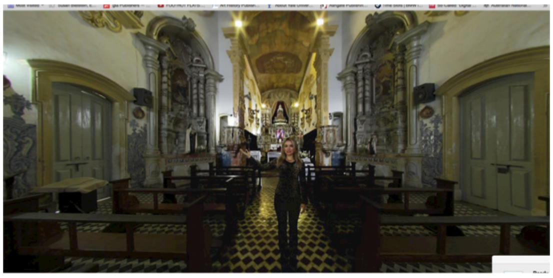

Figure 5: *Interactive 360-degree panorama featuring a filmed presenter (via the web or Head Mounted Displays).*

Motion analysis from the HMDS, biofeedback from low-cost biofeedback devices (*emotiv*, *Neurosky*, *Wild Divine* et al.) and gaze analysis data (*Torbii* and other products) can be incorporated into generic feedback for both audience and presenter. These products (Figure 6) can connect with entertainment technology [Fas12] or mainstream 3D real-time rendering engines such as Unity (used in this project). In 2014 we developed middleware for biofeedback sensors that could be connected to mainstream game engines and we will also see if indirect biofeedback can be incorporated into the background of the presentation or provide feedback to the speaker.

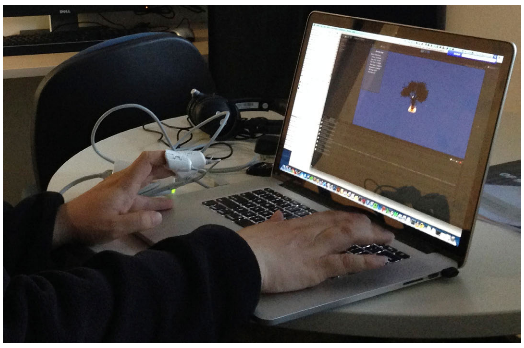

Figure 6: *Middleware for biofeedback, (here the equipment is Wild Divine) finger sensors with Unity.*

6. CONCLUSION

We have discussed the development of a prototype that can be used to communicate both virtual environments and speakers (either separately or in-scene) to remote audiences or to large local audiences. The speaker's gestures can also be mirrored by an in-scene avatar, (thus minimizing potential external distractions and allowing interesting 'slippages' between avatar and presenter). In terms of technical implementation, the current setup uses a large curved display that can also project in stereo, but this application can also be used on conventional desktops. Our next task is to evaluate the usefulness, usability and engagement of this software, focussing on specific design scenarios drawn from architectural and archaeological presentations and classroom teaching.

Also, tracking head movement, gaze direction, postural changes, biofeedback or "thought control" [PMS*13] could allow conference participants, hosts and distantly located narrators the ability to create more immersive conference presentations inside and outside of digital 3D models.

7. References

[AM99] Alvarado, R. G. & Maver, T. Virtual reality in architectural education: Defining possibilities, *ACADIA Quarterly* 18, (1999), 7-9.

[ST13a] Sherstyuk, A. & Treskunov, A. Head tracking for 3D games: Technology evaluation using CryENGINE2 and faceAPI, *Virtual Reality (VR), 2013*, (2013) *IEEE*, 67-68.

[ST13b] Sherstyuk, A. & Treskunov, A. Natural head motion for 3D social games, *Virtual Reality (VR), 2013* (2013), *IEEE*, 69-70.

[GDT*14] Gadanac, D., Dujak, M., Tomic D., & Jercic, D. Kinect-based presenter tracking prototype for videoconferencing, *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014*, (2014), 485-490.

[PMS*13] Powell, C. Munetomo, M., Schlueter, M. & Mizukoshi, M. Towards Thought Control of Next-Generation Wearable Computing Devices, in *Brain and Health Informatics*, ed: Springer, (2013), pp. 427-438.

[C03] Cheng, N-Y. Approaches to design collaboration research, *Automation in construction,* vol. 12, 6 (2003), Elsevier, pp. 715-723.

[PCJ*02] Peng, C., Chang, D. C., Blundell Jones, P. & Lawson, B. Exploring urban history and space online: design of the virtual Sheffield application, *Design Studies,* vol. 23, 5 (2002), pp. 437-453,.

[AG13] Andujar M. & Gilbert, J.E. Let's learn!: enhancing user's engagement levels through passive brain-computer interfaces, *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (2013), pp. 703-708.

[Fas12] Fassbender, E. Use of 'The Elder Scrolls Construction Set' to create a virtual history lesson," *Game Mods: Design Theory and Criticism* (2012), ETC Press, USA, pp. 67-86.

[DE15] De Angeli, D. & O'Neill, E. Transfer of learning between screen-based and gallery-based content: an initial study. SEAHA Conference (2015), University of Bath. http://www.seaha-cdt.ac.uk/conference-programme/

[PFS*15] Papaefthymiou, M., Feng, A., Shapiro, A., & Papagiannakis, G. A fast and robust pipeline for populating mobile AR scenes with gamified virtual characters. In *SIGGRAPH Asia 2015: Mobile Graphics and Interactive Applications, (Nov* 2015), 2, p. 22. ACM.

8. ACKNOWLEDGMENTS

In particular we would like to thank the HIVE at Curtin University, ZIGFU, an iVEC/Pawsey intern grant, and summer intern/programmer Samuel Warnock.