Interactions in a Human-Scale Immersive Environment: the *CRAIVE*-Lab

Gyanendra Sharma
Department of Computer Science
Rensselaer Polytechnic Institute
sharmg3@rpi.edu

Jonas Braasch
School of Architecture
Rensselaer Polytechnic Institute
braasj@rpi.edu

Richard J. Radke
Department of Electrical,
Computer, and Systems
Engineering
Rensselaer Polytechnic Institute
rjradke@ecse.rpi.edu

Abstract

We describe interfaces and visualizations in the CRAIVE (Collaborative Research Augmented Immersive Virtual Environment) Lab, an interactive human scale immersive environment at Rensselaer Polytechnic Institute. We describe the physical infrastructure and software architecture of the CRAIVE-Lab, and present two immersive scenarios within it. The first is “person following”, which allows a person walking inside the immersive space to be tracked by simple objects on the screen. This was implemented as a proof of concept of the overall system, which includes visual tracking from an overhead array of cameras, communication of the tracking results, and large-scale projection and visualization. The second “smart presentation” scenario features multimedia on the screen that reacts to the position of a person walking around the environment by playing or pausing automatically, and additionally supports real-time speech-to-text transcription. Our goal is to continue research in natural human interactions in this large environment, without requiring user-worn devices for tracking or speech recording.

Author Keywords

Interaction; Immersive Environment; Person Tracking; Multimedia

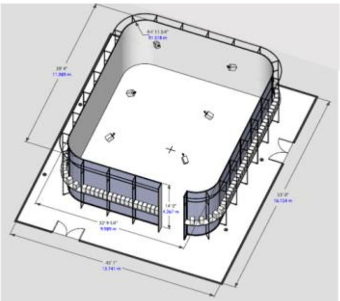

Figure 1. The physical and infrastructural specifications of the CRAIVE-Lab. The lab is equipped with 8 HD video projectors and 128 loudspeakers behind a 5m tall, acoustically transparent micro-perforated screen. 6 downward-pointed cameras are mounted in the ceiling of the space for continuous tracking visual tracking of participants. Figure courtesy J. Parkman Carter.

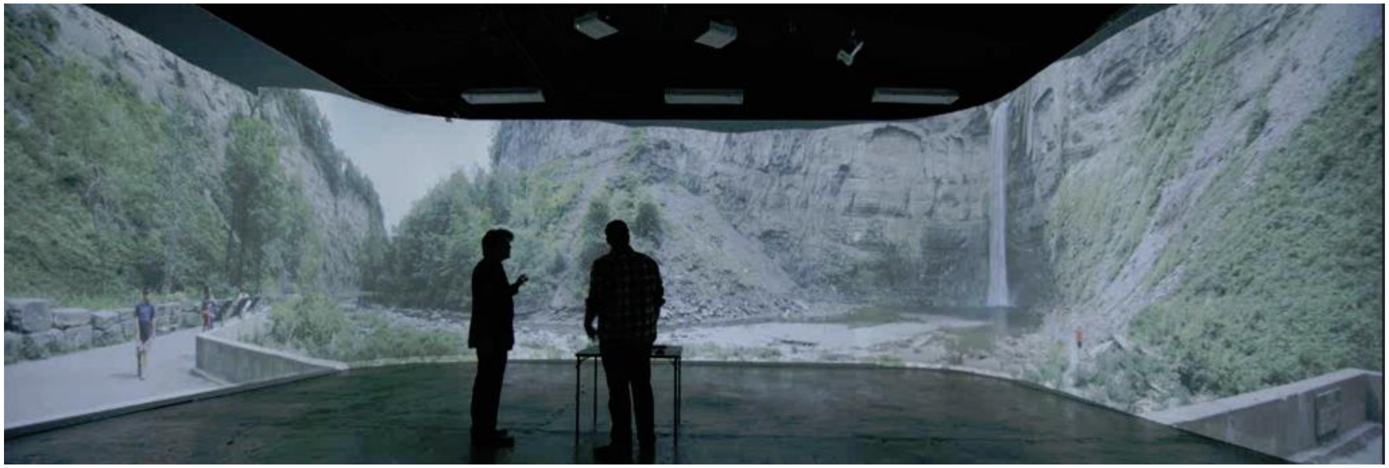

Figure 2. Snapshot of the CRAIVE lab with a panoramic image on display.

ACM Classification Keywords

H.5.m. Information interfaces and presentation

Introduction

Large-scale display technologies have several interesting applications including the presentation of high-resolution imagery for geospatial, scientific visualization, and command-and-control scenarios [7]. When large displays enclose a space to create an immersive environment, it is not viable for users to interact with the displays using traditional interfaces like a mouse or keyboard. The scale of the environment makes it difficult to carry out even basic tasks such as finding the cursor or clicking on a certain part of the screen. In addition, even if visualizations, imagery or multimedia can be efficiently presented on the large screen, exploring them without using invasive technologies such as wearable sensors can be difficult. In this paper, we present the CRAIVE (Collaborative Research Augmented Immersive Virtual Environment) Lab, a human-scale, occupant-aware immersive environment at Rensselaer Polytechnic Institute. Our current work focuses on creating interactive interfaces in the CRAIVE-Lab by automatically detecting the spatial and temporal positions of the participants. Ultimately, this tracking will enable personalized communication between the smart environment and each user through mobile devices and handheld surfaces.

The CRAIVE-Lab, illustrated in Figures 1 and 2, has a 360-degree front-projected screen equipped with 8 1200x1920 resolution projectors along with a network of 6 overhead 1280x960 resolution cameras used for visual tracking of participants. The floor space enclosed within the screen is rectangular with curved corners and has a length of approximately 12 meters and width of 10 meters. The screen is approximately 5 meters in height, and has an effective resolution of 1200x14500 pixels. The projectors are positioned to avoid casting user shadows onto the screen unless they are very near to it.

Here, we describe several user interfaces to the CRAIVE-Lab that require neither wearable sensors nor direct interaction with the system through a mouse or keyboard. We use computer vision algorithms for user tracking along with the Websocket message passing interface to realize an occupant-aware environment. For example, in the "smart presentation" scenario, a user simply has to stand close to a multimedia clip in the physical space, activating the clip by their presence and automatically transcribing their speech in the appropriate location.

The remainder of the paper discusses the integral building blocks of the CRAIVE-Lab system, including visual tracking of participants from above, screen and floor coordinate system calibration, and system architecture, resulting in real-time systems for "person following" and "smart presentation" scenarios.

Related Work

Human scale immersive interactive environments have recently become a research focus in industry, government, and academia. Natural interactions in the context of these environments is a critical issue. Our approach focuses on natural and non-invasive techniques for building occupant-aware spaces. Zabulis et al. [11] explored the problem in a similar way, supporting natural interaction in a large screen display environment.

Bradel et al. [3] studied how multiple users can collaborate and interact using large high resolution displays, though they were not in the context of an immersive environment. However, it is useful to understand how large displays affect collaboration, which is an important issue for our system. North and North [8] studied user experiences based on virtual reality and immersive environments, investigating factors that affect the sense of presence and immersion. One of the major contributions of this study is the conclusion that immersive systems provide users with a heightened sense of presence. This was one of the motivations for the work presented in this paper: to create interfaces that aid in creating meaningful immersive experiences.

Ardito et al. [1] presented a survey related to interactions with large displays. While the testbed for the survey involved simple large screen displays instead of extremely large walls as in our system, several of the surveyed interaction techniques could be used for future interfaces in the CRAIVE-Lab. Interaction techniques involving handheld devices, e.g., [4, 5, 6, 10], are also well within the scope of our future work.

System Architecture

Figure 3 illustrates the software architecture of the CRAIVE-Lab system. The 6 overhead cameras provide a live feed to the tracking algorithm, which returns the global (x, y) floor coordinate for each user inside the CRAIVE-Lab. Using a Websocket client, the coordinates are sent to a Nodejs server hosted in the same node (i.e., computer). This server in turn communicates via Websocket to the HTML client hosted in the projector node. User-worn lapel microphones communicate

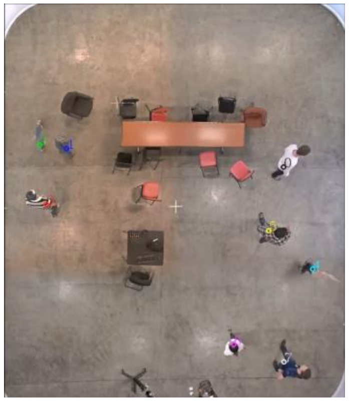

Figure 4. A snapshot of the live visual tracking system in the CRAIVE-Lab. A unique ID (colored number) is assigned to each occupant, which allows the system to keep track of all participants in the environment.

directly with the projector node, using a web client to transcribe speech to text using the Google speech API.

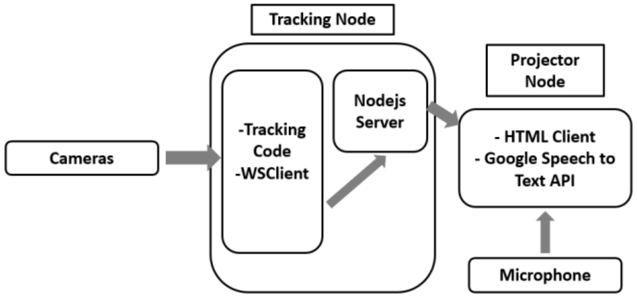

Figure 3. Block diagram of the overall system architecture.

Tracking

The CRAIVE-Lab visual tracking algorithm operates in real-time on each of the 6 cameras, fusing the results into target positions in a global coordinate frame, as illustrated in Figure 4. The six cameras provide full coverage of the CRAIVE-Lab floor, allowing for continuous tracking across any region in the room.

Users are detected inside the space using a background subtraction method. A multi-threaded approach is used to track people in individual cameras. At each moment, the individual camera feeds are combined into a mosaic covering the entire space with proper alignment of the overlapping regions between neighboring cameras. This alignment involves camera-to-camera homographies relating the ground planes in each viewpoint, calibrated at system setup using feature correspondences in the overlap regions between fields of view. In this way, the local tracking coordinates returned by each camera are mapped into global coordinates for the floor.

Calibration

To use spatial and temporal motion of a user in the space to drive the interactive visualization on the screen, a mapping between floor and screen coordinates is required. The mapping should enable visual cues on the screen to aid the user in intuitively translating where he/she is on the floor to a location on the screen.

The first issue is determining a unified screen coordinate system, taking into account that the display is generated from 8 overlapping projected images. We use the PixelWarp software package [5] to handle the 8-projector display as a virtual 14500-pixel-wide single image. The calibration process involves both geometric warps and color blends in the overlap regions, so the projection is virtually seamless to participants. The projector-to-projector mappings were created using a meticulous manual calibration process when the system was commissioned.

We determined the mapping from floor coordinates to screen coordinates by positioning a user at 200-pixel (roughly 0.7 m) intervals at a fixed distance from the display wall along its entire 14500-pixel-wide extent, and recording the corresponding (x, y) locations in floor coordinates. This mapping is generally linear, but must take into account the curved regions at the corners of the projection screen. Ultimately, the user’s floor x coordinate drives the mapping to Screens 1 and 3, and their y coordinate drives the mapping to Screens 2 and 4, as illustrated in Figures 5 and 6.

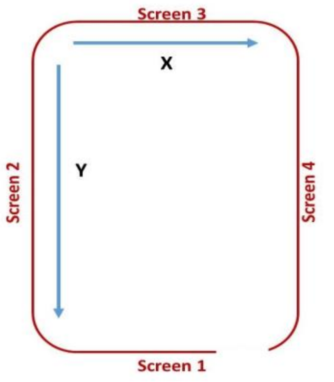

Figure 6. Illustration of the CRAIVE-Lab screen and floor space. The origin of the floor coordinate system is at the upper left corner and coordinates increase as indicated by the arrows. Note that the screen numbers simply indicate locations of content for the purposes of this paper; the screen itself is a single long display surface.

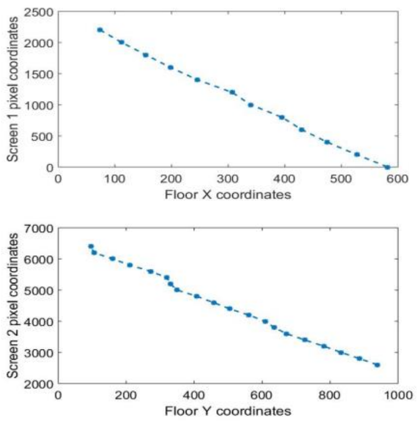

Figure 5. Floor coordinates vs screen pixel coordinates for Screens 1 and 2. Screens 3 and 4 follow similar patterns.

Person Following

As shown in Figure 7, we aim to provide a visual cue to a user walking inside the immersive space. Our decision about the visualization artifact was made with the goal of providing unambiguous feedback about the user's spatial location. For this purpose, we took a "crosshair"-like approach for visualization. A user walking around the space will be followed by four filled circles, one on each screen, such that connecting opposite circles pinpoints the user in the interior of the room. This is illustrated from overhead in Figure 8.

Hence, all four sides of the environment provide feedback to the user regarding his/her location in the space. However, in order to avoid ambiguity as well as to enhance occupant awareness, the circles are resized dynamically. That is, we maintain an inversely proportional relationship between the user's distance to a screen and the circle size on that screen (i.e., the circle on the screen closest to the user is the largest). Also, in order to reduce noise from the tracking algorithm, a median filter is applied across 20 consecutive frames so that the circles on the screens do not have too much fluctuation.

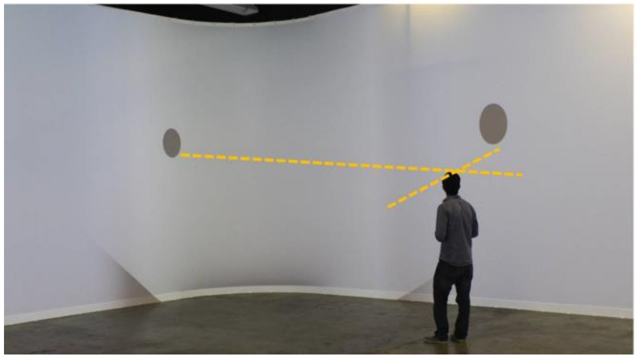

Figure 7. Snapshot of "person following" in the CRAIVE-Lab, with lines added for illustration. Circles with dynamic sizes and locations across the screen are used as visual cues for a user walking around the space and act as "crosshairs". A video link is available at https://youtu.be/o6cF7_Gyz3Y.

Smart Presentation/Multimedia

We next present a more advanced occupant-aware scenario that uses the same underlying tracking algorithm. This design extends the "person following" idea to show how the existing technology can be applied for intuitive, interactive presentations. Since we know the tracking coordinates and their corresponding mapping to screen coordinates as discussed in the previous sections, we designed a multimedia presentation system controlled by the spatial location of the user. The idea is to have several large videos at different positions along the screen. Each video plays

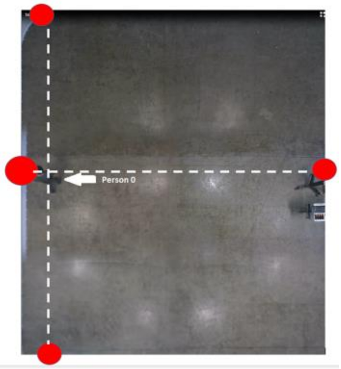

Figure 8. Four circles pinpoint the user's location in the room interior. Here, a mosaiced overhead view of the space is shown with a person in the left side. Four circles on each of the sides of the screen are automatically drawn, such that the virtual dotted lines intersect at the user's position. The screen to which the person is closest has the largest circle.

only when the user is right in front of the video and sufficiently close to the screen. In addition, the user wears a lapel microphone, allowing him/her to talk about the presentation/multimedia; this speech is automatically transcribed and displayed above the corresponding video. Any comments made by the user while outside the mapped floor region of the multimedia content is ignored. This allows the user to move around the space while controlling the content on the screen without making any conventional active interactions such as mouse clicks. An illustration of this idea is shown in Figure 9.

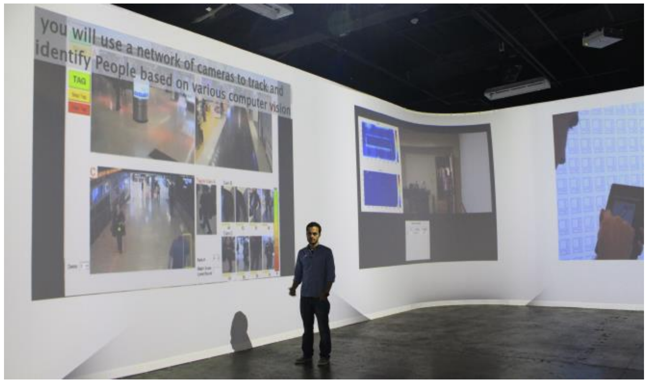
you will use a network of cameras to track and
identify People based on various computer vision
TAG

Figure 9. Snapshot of the "smart presentation". Here, the user has moved to the first video on the left. All other videos are paused as the user is not in the mapped floor region of either of the two videos, while the first one is playing along with the audio transcription appearing above it. A video link is available at: https://youtu.be/VbXMB97HwvM

A prototype of this system was implemented in a separate space similar to the CRAIVE-Lab during the IBM Cognitive Colloquium in November 2015. Various research projects were presented using this system, in which the presenter walked around the room explaining the projects while the videos played and paused automatically along with automatic speech transcription. The response from the attending participants was very positive and successfully showcased the occupant-awareness of the room and its ability to support natural user interactions.

Conclusions and Future Work

In near-term future work, we plan to integrate various cross-surface devices such as tablets with the large screen. A user could easily use his/her mobile surface to share local content onto the display wall; similar to the "smart presentation" interface, the content's screen location could be synchronized with the user's spatial location. A reverse scenario of content sharing from the display wall to specific user devices could also be realized using the system we have in place.

The CRAIVE-Lab can be further improved to include various gestural or voice interactions apart from the spatially controlled interactions. For example, floor-mounted cameras and range sensors like the Microsoft Kinect can capture users' gestures, head poses, and facial expressions for finer control of presentation elements. In addition, multimedia elements in the environment could be controlled directly from speech understanding, preferably using ambient microphones instead of lapel-worn wireless microphones.

Our ultimate goal and current work in the CRAIVE-Lab focuses on making effective use of the large space with respect to service systems that aid both individual and group interactions, interfacing with advanced cognitive computing algorithms "behind the screen". Ultimately, we envision an intelligent environment that understands its occupants' locations, movement, speech, vocabulary, and intentions, and is able to actively facilitate meetings, aware of the day's agenda and action items, the participants and their roles, and what happened previously. The environment could keep meetings on track, assess participation and progress towards goals, maintain action items, summarize discussion, and mediate brainstorming.

Acknowledgements

The construction of the CRAIVE-Lab was supported by the National Science Foundation and by Rensselaer Polytechnic Institute [2]. Robert Carney and J. Parkman Carter worked with the CRAIVE-Lab project team to plan and direct the installation of the CRAIVE-Lab screen and loudspeakers and to calibrate the system. The metal frame for the CRAIVE-lab was built by Adirondack Studios. Jacob Braun installed the cameras and implemented the visual tracking algorithm based on an original algorithm by Ziyan Wu. The research in this paper was partially supported by NSF's PFI:BIC program under award IIP-1631674, and by IBM Research via Rensselaer's Cognitive and Immersive Systems Laboratory.

References

[1] Ardito, C., Buono, P., Francesca, C. and Desolada, G., Interaction with large displays: A survey. *ACM Computing Survey*, 47, 2015.

[2] Braasch, J. (PI), Chang, B. (Co-PI), Cutler, B. (Co-PI), Goebel, J. (Co-PI) and Radke, R. (Co-PI), MRI/Dev.: Collaborative-Research Augmented Immersive Virtual Environment Laboratory (CRAIVE-Lab), NSF/Major Research Instrumentation, 10/12–09/15, CNS-1229391.

[3] Bradel, L., Endert, A., Koch, K., Andrews, C. and North, M., Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality. *International Journal of Human Computer Studies*, 71:1078-1088, 2013.

[4] Jeon, S., Hwang, J., Kim, G.J. and Billinghurst, M., Interaction techniques in large display environments using handheld devices. *Proceedings – ACM Symposium on Virtual Reality Software and Technology – VRST '06*, 2006.

[5] Malik, S., Ranjan, A., and Balakrishnan, R., Interacting with large displays from a distance with vision-tracked multi-finger gestural input. *Proceedings – ACM symposium on User Interfave Software and Technology – UIST '05*, 2005.

[6] Nancel, M., Chapuis, O., Pietriga, E., Yand, X.D., Pourand, I. and Beaudouin-Lafon, M., High-precision pointing on large wall displays using small handheld devices. *Proceedings – SIGCHI Conference on Human Factors in Computing Systems – CHI '13*, 831-840, 2013.

[7] Ni, T., Schmidt, G.S., Staadt, O.G., Livingston, M.A., Ball, R. and May, R., A survey of large high-resolution display technologies, techniques, and applications. *Proceedings - IEEE Virtual Reality*, page 31, 2006.

[8] North, M.M. and North, S., A comparative study of sense of presence of traditional virtual reality and immersive environments. *Australasian Journal of Information Systems*, 20:1-15, 2016.

[9] Pixelwarp warp and blend software. http://pixelwix.com/6-warp-and-blend-software

[10] Vogel, D., and Balakrishnan, R., Distant freehand pointing and clicking on very large, high resolution displays. *Proceedings of the ACM symposium on User Interface Software and Technology – UIST '05*, 2005.

[11] Zabulis, X., Grammenos, D., Sarmis, T., Tzevanidis, K., Padeleris, P., Koutlemanis, P. and Argyros, A.A., Multicamera human detection and tracking supporting natural interaction with large-scale displays. *Machine Vision and Applications*, 24:319-336, 2013.