The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Symbiotic Cognitive Systems: Technical Report WS-16-14

Ceding Control: Empowering Remote Participants in Meetings involving Smart Conference Rooms

V. Venkataraman[1,2], J. Lenchner[1], S. Trewin[1], M. Ashoori[1], S. Guo[1], M. Dholakia[1], P. Turaga[2]
[1]IBM Thomas J Watson Research Center, Yorktown Heights, New York, USA
[2]Arizona State University, Tempe, Arizona, USA

Abstract

We present a system that provides an immersive experience to a remote participant collaborating with other participants using a technologically advanced "smart" meeting room. Traditional solutions for virtual collaboration, such as video conferencing or chat rooms, do not allow remote participants to access or control the technological capabilities of such rooms. In this work, we demonstrate a working system for immersive virtual telepresence in a smart conference room that does allow such control.

Introduction

Technologically advanced "smart" conference rooms are becoming more common place (Chen, Finin, and Joshi 2004). In order to support more fluid interaction with users, these conference rooms employ high bandwidth audio-visual equipment that also afford a high level of interactivity, especially using voice and/or gestural input. In such environments, the idea of virtual telepresence is an interesting problem. If the user is not present in the room it has, until now, been impossible to experience anything similar to what is experienced in the room itself. In addition, the remote participant is passive and has virtually no ability to interact with any of the components of the "smart" room. In our work, we present a working system which provides an immersive experience for remote participants while also enabling them to control most aspects of the smart conference room.

Following failed attempts in building virtual reality (VR) systems over the past couple of decades, the VR community has become re-energized as a result of recent advances in virtual reality head-mounted display devices such as the Oculus Rift (Oculus 2015). The Rift offers compelling immersive experiences in virtual worlds suitable for gaming and the more direct experiencing of stories. Technology giants such as Facebook and Microsoft are working on a consumer version of virtual reality headsets with the gaming audience as their main target. It is noteworthy that the environments considered in a vast majority of the applications have been virtual, lacking any correspondence with the real world. Such an approach, which makes use of a static 3D model, does not offer a complete solution for a remote participant in a smart conference room. Dynamic environments such as a conference room, with people moving around and content on screens changing, require frequent update lest the remote person risk losing important information about what is transpiring in the room. In this work we utilize the live video feeds from IP cameras to overlay the dynamics of the environment on a static 3D model of the room. The IP cameras capture the content displayed on screens as well as people in the room, and this content is projected on the walls of the 3D model. Figure 1 shows the architecture of the system. The individual blocks, along with their functionality, are discussed in greater detail in future sections.

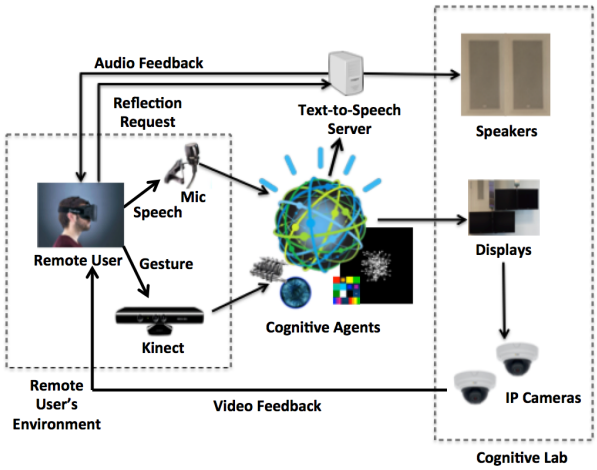


Figure 1: Architecture of our system that enables a remote user to interact with the capabilities of an agent-based smart conference room, using technologies such as the Oculus Rift and the Microsoft Kinect.

Contributions and Outline:

The contributions of this work are two-fold:

1. We enable an immersive experience for remote participants in a smart conference room using a multi camera network and a head-mounted display such as the Oculus Rift, and
2. We enable gestural interaction from a remote location to directly access/modify content in the local meeting room.

Related Work

Rheingold (Rheingold 1991) defines virtual reality as an experience where a person is "surrounded by a three dimensional computer–generated representation, and is able to move around in the virtual world and see it from different angles, to reach into it, grab it, and reshape it." Most sources trace the first virtual reality head-mounted display system to Sutherland (Sutherland 1968), who created a system based on 3D modeling. Sutherland's head mounted display was so heavy it had to be suspended from the ceiling, and the display was limited to a wireframe model of a room. Numerous virtual reality systems have been proposed since then. (Mazuryk and Gervautz 1996) provides a thorough historical survey through 1996.

Of particular note among the older virtual reality systems is the CAVE virtual reality theater system of Cruz-Neira et al. (Cruz-Neira, Sandin, and DeFanti 1993). The CAVE consisted of rear projection screens on three sides (no display in the back of the room) together with a floor projection system. In addition, users wore 3D glasses, a motion capture system tracked user's motion, and with the help of the glasses the system rendered a view that was consistent with the user's position within the room. More recent versions of the CAVE system have appeared with enhanced video projection and faithful rendering of audio from appropriate directions within the CAVE. Many universities today have their own CAVE systems and a variety of software libraries dedicated to CAVE development are in use. The CAVE's projection system is very similar to our own scheme for projecting imagery onto the walls in the 3D model of our cognitive lab and then rendering the view in a perspectively correct way using the inherent capability of the Oculus Rift.

Although virtual reality is most commonly associated with gaming, a recent article in the online journal TechRepublic (Carson 2015) described nine industries actively utilizing virtual reality, including healthcare, automotive, education, tourism, the military, and law enforcement. In particular they mention that the Ford Motor Company is using the Oculus Rift to envision new automobiles under design, and that the British military is using the Rift to train trauma medics for work under actual battle conditions. NASA has been using virtual reality for many years, both for training and to achieve some degree of control of remote vehicles (e.g., the Lunar Lander).

We believe that our use of the Oculus Rift for remote control of a smart conference room to be the first time a heads up display type virtual reality system has been used for real-time remote control of an actual physical environment.

The Cognitive Environments Lab

The Cognitive Environments Lab (CEL) is a conference room equipped with a variety of "smart" devices and sensors. In the front of the room a 4×4 array of high definition (1920×1080 pixel) monitors acts like a single large display surface. On the right and left of the room, respectively, are two pairs of high definition monitors on tracks. The tracks allow the monitors to be moved anywhere along the periphery of the room. In the back of the room there is an 84 inch, 4K (3840×2160 pixel) display.

Content can be moved from monitor to monitor with the aid of a special cross-monitor pointing device called a "wand." Such content can also be moved programmatically. The underlying 3D-spatial computing framework underlying the multi-display system is known as g-speak, a product of Oblong Industries (Oblong 2015). In addition, the room is outfitted with a large number of microphones and speakers. With the aid of a speech-to-text transcription system, the room can understand elements of what is being said, and with the aid of a complimentary text-to-speech system, the room can synthesize appropriate responses.

Applications for the room are built as distributed multi-agent systems. We give some of the more advanced agents the special designation of being "cognitive agents," or "cogs" for short. For a formal definition of the notion of a cognitive agent, consistent with our own informal notion see (Babu 2015).

The aim of this room is to offer the power of advanced computing to assist users in complex decision making tasks. It is designed as a space for humans and systems to collaborate in what we call a "symbiotic cognitive computing experience," where humans are given the opportunity to do what they do best, e.g. identifying important problems and applying value judgements, while the computing systems seamlessly do what they do best, e.g., performing analytics on big data and running simulations (Kephart and Lenchner 2015).

System Architecture

In our system the remote user interacts with the agents of the cognitive lab using speech and gesture. The control flow of the system is depicted in Figure 1. The speech signal is transcribed into word-by-word transcripts using the IBM Attila speech recognition system (Soltau, Saon, and Kingsbury 2010). The system recognizes pointing gestures using the skeletal tracking of body joints from the Microsoft Kinect in conjunction with the Kinect for Windows SDK. Using the 3D coordinates of the body joints, we estimate the pointing direction as the intersection of the line joining the head center (eye position) to the most outstretched position of the hand, with the virtual screen plane (assumed to be parallel to the user's body plane). The output of the interaction is displayed on the room display screens as a cursor. We use live feeds going to the Oculus Rift from strategically situated IP video cameras in the room to provide video feedback to the remote user regarding where the cursor is on any display as well as the possibly changed state of the virtual content shown on the displays.

The Oculus Rift DK2 head-mounted display system used in this work offers a resolution of 960 x 1080 in each eye, and a refresh rate of 75 Hz. The device incorporates complete rotational and positional tracking of the user's head. The positional tracking is performed by an external IR-based tracking unit, which is included with each Rift and normally sits on the user's desk. This system allows for use of the Rift while standing and walking around in a room.

To provide a representative immersive experience to remote participants, we created a 3D model of the room using Unity 3D as shown in Figure 2. Common objects in the

room, such as tables and chairs, were inserted into the model using Google SketchUp. We place the virtual Oculus camera at the outset in the center of the 3D model and thereafter the camera is controlled by the user's head movements as tracked by the Rift. The 3D model allows the system to render in real time a close facsimile to what a user in the real room would see as they take steps toward/away from displays, or turn their head.

Although it is not possible to recreate the entire 3D model of the room as the dynamics of the room change, we make the decision to just change the parts of the scene that are likely of most consequence to the remote user. Thus, we capture the content information on the display screens and anything near the periphery of the room through strategically placed IP cameras, and project these video feeds onto the vertical planes in the 3D model that are associated with the walls of the room. This manner of capturing the dynamics allows the user to see and access the content displayed on the screens and also to see most meeting participants, given the U-shaped seating arrangement that is typical for occupants of the room.

In addition to controlling content via gesture, as sensed by the Kinect, the remote user also receives direct speech output from the system. Before beginning the interaction, the remote user's system sends a reflection requests to the Text-to-Speech server which then echoes the audio signal sent to the room speakers, to the machine supporting the remote user. On the whole, the proposed system offers an immersive experience for remote users that is not too dissimilar to that of local users of the smart conference room.

The speech transcript and Kinect skeletal coordinates are synchronously sent to the speech and gesture comprehension module (SaGComp Module) using the ZeroMQ pub-sub messaging library, which facilitates distributed messaging among our various agents. The SaGComp Module then has the logic to transform speech and gestural input into a request to modify the content displayed on any of the screens in the room.

The system is symbiotic by virtue of the fact that it empowers the remote user via giving them control of the room, while simultaneously empowering the room, by enabling it to understand the remote user, just like it does users in the room.

User Evaluation Study

We decided to study the effectiveness of the proposed system in offering an immersive experience to a remote user collaborating with a local (in-room) user to complete a task. In this study, participants were instructed to collaboratively solve a picture puzzle game displayed on a 4×4 display screen as shown in Figure 3. The remote user can point and speak to give directions, while the local user actually moves the puzzle pieces in response. Participants were instructed to perform the task (a) collaboratively in the remote location using our Oculus VR + Kinect based system, and (b) locally in the room, without collaboration, just using the wand. Puzzle pieces were chosen specifically so they would not be easy to describe, thus necessitating the use of gesture on the part of the remote participant. The task was considered complete

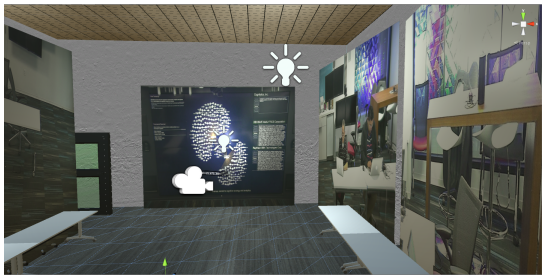


Figure 2: 3D model of the smart conference room created using Unity3D. The virtual light bulbs indicate the location of the virtual light sources. The camera icon on the front screen indicates the initial pointing direction of the camera.

when all tiles were in the required finishing positions. The solved puzzle was displayed on a side screen to aid participants in solving the puzzle.

We recruited a total of 10 participants, and the order in which they performed the tasks was balanced. Thus, five of the participants first played the role of the remote participant in the collaborative exercise and the other five first took on the non-collaborative local task. Nine of the ten participants completed the task both as local and remote participants. One participant did not complete the remote task stating discomfort associated with feelings of nausea/disequilibrium. The details of the user study are tabulated in Table 1. Figure 4 plots the average time taken in seconds to solve the puzzle by local and remote participants as a function of their age (in years). Participants were asked to provide their age in 10-year intervals, hence the necessity of presenting the data in this manner.

We see that when remote, older participants tend to take more time than younger ones. While the same effect is found for the local task, the effect is substantially greater for the remote task. The remote users take on average about twice as long to complete a given task compared with users doing the same task locally.

After completion of the tasks, the participants were given a Questionnaire (**Q1**–**Q4**) to evaluate their experience with the system:

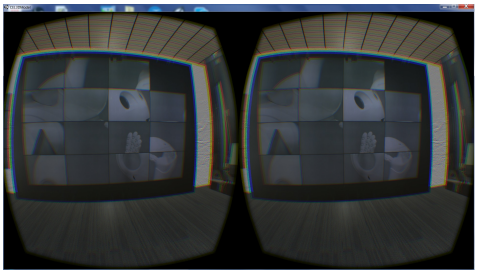
Q1 How did you feel when performing the task remotely?

Q2 How did you feel when performing the task locally?

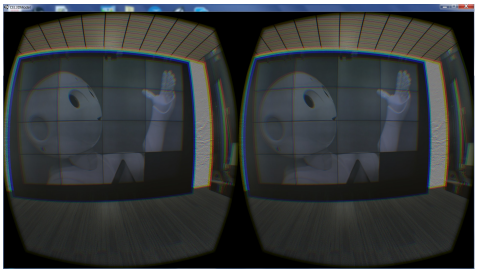
Q3 What did you like the most about remote experience?

Q4 What did you like the most about local experience?

User responses, in the form of a statements, were documented. 7 of the 10 participants felt some amount of dizziness and disequilibrium while performing the task remotely, which is at present one of the known limitations of the Oculus VR headset. Future iterations of the device may ameliorate this sensation. 8 of the 10 participants acknowledged that they liked the 3D model of the room and that it offered a realistic immersive experience while performing the task, while the remaining 2 were not impressed and did not find the experience to be a satisfying one. All 10 participants re-



(a) Picture Puzzle



(b) Solved Picture

Figure 3: Illustration of the picture puzzle game (a variant of the well-known 15-game) displayed on the screens, as seen through the Oculus Rift display.

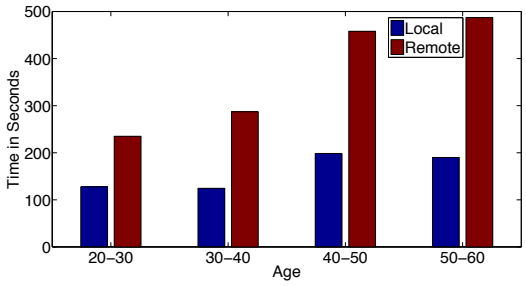


Figure 4: Comparison of time taken in seconds to solve the puzzle by participants in the room (blue) and remote (brown) based on their age groups.

garded their local experience as "natural" with no significant system lag reported.

Conclusion and Future Work

In this paper, we have presented a working system for immersive virtual telepresence in a smart conference room using real-time video feeds from IP cameras along with an Oculus Rift. Using a Kinect, we provide remote users with the ability to control aspects of the room using a combination of gesture and speech. A preliminary user study indicates that this technology provides a reasonable vehicle for effective collaboration among local and remote participants.

Additional work remains to establish whether reported feelings of disequilibrium by our study subjects was due to this known issue with the Oculus Rift or whether the sensation was aggravated by either (i) the mixed 3D and 2D environment users experience when we project part of the 3D scene onto the virtual walls in the 3D model, or (ii) the slight lag incurred by the network and processing associated with capturing and then projecting content from the IP cameras.

Although a long way off, this technology offers the possibility of some day enabling such applications as remotely taking over for an airline pilot in distress, or for a surgeon who finds something unexpected in the midst of a surgery.

Table 1: Summary of our user experience study, showing, for each participant, their age and time taken to complete the task in seconds. Subject 6 did not complete the task remotely due to feelings of nausea and dissequilibrium.

Subject No.	**Age Range** (in years)	**Time Taken** (secs)	
		Local	**Remote**
1	20 - 30	96	230
2	20 - 30	159	240
3	30 - 40	138	194
4	40 - 50	276	303
5	40 - 50	121	613
6	40 - 50	161	DNC
7	30 - 40	101	470
8	50 - 60	190	487
9	30 - 40	178	340
10	30 - 40	81	144
Average		150	335

References

Babu, B. S. 2015. Cognitive agents. http://pet.ece.iisc.ernet.in/sathish/cognitive.pdf. Accessed: 2015-10-18.

Carson, E. 2015. Nine Industries Using Virtual Reality. http://www.techrepublic.com/article/9-industries-using-virtual-reality/. Accessed: 2015-10-11.

Chen, H.; Finin, T.; and Joshi, A. 2004. An ontology for context-aware pervasive computing environments. *The Knowledge Engineering Review* 18(3):197–207.

Cruz-Neira, C.; Sandin, D. J.; and DeFanti, T. A. 1993. Surround-screen projection-based virtual reality: The design and implementation of the cave. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, 135–142.

Kephart, J. O., and Lenchner, J. 2015. A symbiotic cognitive computing perspective on autonomic computing. In *Proceedings of the 2015 International Conference on Autonomic Computing*, 109–114.

Mazuryk, T., and Gervautz, M. 1996. Virtual reality history, applications, technology and future. *Technical Report, TR-186-2-96-06, Institute of Computer Graphics and Algorithms, Vienna University of Technology*.

Oblong. 2015. g-speak. http://www.oblong.com/g-speak/. Accessed: 2015-10-25.

Oculus, V. 2015. Oculus Rift. http://www.oculusvr.com/rift. Accessed: 2015-10-11.

Rheingold, H. 1991. *Virtual Reality*. New York: Summit.

Soltau, H.; Saon, G.; and Kingsbury, B. 2010. The IBM Attila speech recognition toolkit. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, 97–102.

Sutherland, I. 1968. A head-mounted three dimensional display. *Fall Joint Computer Conference, AFIPS Conference Proceedings* (33):757–764.