ShowMe: A Remote Collaboration System that Supports Immersive Gestural Communication

Judith Amores *
MIT Media Lab
75 Amherst Street, Cambridge,
02139, USA.
amores@media.mit.edu

Xavier Benavides *
MIT Media Lab
75 Amhers t Street, Cambridge,
02139, USA.
xavib@media.mit.edu

Pattie Maes
MIT Media Lab
75 Amherst Street, Cambridge,
02139, USA.
pattie@media.mit.edu * Equal contribution to this work

CHI'15 Extended Abstracts, Apr 18-23, 2015, Seoul, Republic of Korea
ACM 978-1-4503-3146-3/15/04.
http://dx.doi.org/10.1145/2702613.2732927

Abstract
ShowMe is an immersive mobile collaboration system that allows a remote user to communicate with a peer using video, audio and hand gestures. We explore the use of a Head Mounted Display and depth camera to create a system that (1) enables a remote user to be immersed in another user's first-person's point of view and (2) offers a new way for the remote expert to provide guidance through three dimensional hand gestures and voice. Using ShowMe both users are present in the same physical environment and can perceive real-time communication from one another in the form of 2-handed gestures and voice.

Author Keywords
Remote collaboration; Shared experiences; Telepresence; 3D interaction; Augmented reality; Hands free Interaction

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces - Training, help, and documentation; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces - Computer-supported cooperative work

Figure 1. System Overview

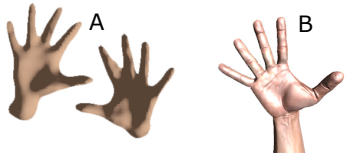

Figure 2. (A) Blob hand. (B) 3D hand mesh

Introduction

Today more than ever, people are able to collaborate at a distance. Commercial teleconferencing technologies are cheap and portable and are more immersive than traditional voice-only phone calls. Nevertheless, when it comes to performing a physical task collaboratively, existing technologies offer limited ways to employ gestures to interact in the remote user's view, as they tend to focus on face-to-face experiences. One problem derives from the user's need to have both hands to interact with their environment, while also using their hands to hold their devices in order to share the workspace with the remote participant. When using teleconference systems with smartphones or tablets, people tend to switch between the front and back camera or they may put the device in a fixed position so as to have freedom of movement [11]. In most cases, the user has to move the camera around in order for the remote person to perceive the entire scene.

In this paper, we propose a solution to this problem in the form of ShowMe, an immersive, mobile system that enables easy communication between remote users about a physical task.

The prototype is designed to be used by two users (Figure 1). For convenience we will refer to the user who would like to share his point of view and receive guidance as the "novice" user, and the one who is remotely assisting as the "expert" user, even though the roles may well be reversed. Both users wear a Head Mounted Display (HMD) and perceive the same view, namely the first-person point of view of the novice. The novice user has two cameras attached to their HMD and

Figure 3. Some examples of different hand gestures. A variety of movements can be performed using the blob hands.

the expert has a depth sensor camera attached to theirs. This set up allows the remote expert to perform hand gestures that are superimposed in real-time onto the novice user's immediate environment (Figure 2, 3). The novice sees his/her own hands as well as the hands of the expert in real-time in their immediate physical surroundings.

ShowMe is a proof of concept built to investigate how we can provide a more useful system for remote assistance with manually oriented tasks. The system shares the point of view of the novice user with the expert who in turn can make their hands inhabit that space so as to offer assistance. The system also shares real time audio for the two users to communicate.

This work contributes (1) the implementation of the hardware prototype of ShowMe, achieved by modifying existing devices adding a depth sensor and cameras, (2) and the software implementation of virtual and augmented reality space integration with 3D hand gestures.

ShowMe can be used in a variety of applications where remote collaboration is useful, for example in remote maintenance of complex machinery, training how to operate devices and finally shared collaboration on the design of physical artifacts. Our proposed system looks to aid in reducing travel expenses by allowing manual problems to be solved collaboratively at a distance.

Related Work

Commercial videoconferencing systems are abundant (Skype, Google Plus Hangouts, Cisco WebEx Conferencing, etc.). Most of these systems enable face-to-face communication from disparate locations but they do not allow remote users to share and reference a common physical workspace. Researchers tried to overcome this limitation using a variety of approaches [7], ranging from projected interfaces [10] to HMD technology [2]. Pioneering work aimed to create an interactive shared drawing surface that both users could work on [5]. Nevertheless, researchers still try to understand and build tools to support collaborative work so as to create a more heightened sense of physical co-presence.

Over the past several years, researchers have studied video communication systems that support collaborative work by remote users in a shared virtual space [4]. These systems integrate depth sensors that analyze body interactions and create a shared depth mirror that allows users to work together in the same space. Unlike ShowMe, these systems are not focused on sharing a first person view of the content and they are not mobile.

Some researchers have focused on proving the efficiency of gestures in shared workspaces. Kirk et al. [9] determined that gestures and visual information improve the speed and accuracy of remote collaboration activities, and Fussell et al. [16] demonstrated that collaborating users rely more on visual actions than on speech. Tang et al. [13] confirmed that 35% of the gestures performed in a collaborative task are performed to engage the other users and express ideas. This research motivated our work on ShowMe to support gestural communication in a remote collaboration system.

HMDs have lately attracted considerable attention as a human interface technology, even though they have been researched since the 60's [18, 19]. One of the most closely related projects to ShowMe is the JackIn project, which explores integrating a first person view with out of body vision for human-human communication [2]. One of the users wears a transparent HMD and shares his view to a remote user who sees an out of the body view displayed on a static desktop monitor. The difference between JackIn and ShowMe is that we create a mobile system where both users are wearing HMDs and are immersed in the same view, instead of having an out of body view. In contrast with JackIn and other related research [12], in ShowMe both hands can be tracked and displayed using 3D hand models (instead of using a flat graphical user interface that supports tele-pointing).

Another interesting project related to our work is "3D Helping Hands". In this system both users are fused in the same 3D rendering space and the remote expert uses an HMD to perform hands gestures in a shared virtual space [14]. However, in contrast with ShowMe, the novice user is not using an HMD and has to look through a screen; therefore s/he is not immersed in the same view as the expert.

BeThere [1] is another closely related project. It explores the use of 3D gestures and remote spatial input without any type of HMD. BeThere allows users to leverage a basic pointing gesture and orientation of the finger in order to control a virtual 3D hand. One limitation of BeThere, besides the fact that it only

detects one gesture and one hand, is that the whole device is too heavy to hold, forcing the user to use a monopod. The difference between ShowMe and these systems is that in ShowMe the user is able to perform full hand gestures with both hands and that novice and expert are completely immersed in the same experience. Moreover, we designed a portable setup with the novelty of incorporating a wide field of view that tracks full movements of both users' hands and shares this data over the Internet. A similar setup is used in recent work by Oda et al. [8] that present a system to guide the user to the perfect viewpoint of an object. To achieve their goal, they employ a 3D model and several pre-recorded angles. In contrast, ShowMe proposes a mobile system that can be used in new environments and relies on hands gestures to perform remote collaborative work.

Implementation

The system is currently implemented using two Oculus Rifts connected to an Apple MacBook Pro 13", an external battery (to power the Oculus Rift), a pair of headphones and the internal camera microphones. The HMD of the novice user also has two Logitech Pro 9000 cameras attached and the expert user's HMD has an Intel Creative Senz3D Depth Camera attached to capture and track her/his hands. We designed the system to be mobile, comfortable and small enough to fit in a daily handbag or backpack.

Unity3D is used to render, process and display the incoming data from the webcams and from the depth camera. Once the data is collected, we stream it and display it to both Oculus Rifts (superimposing the 3D hands meshes of the expert user). The whole system runs on Windows 7 installed on two laptops.

Both computers are sharing information remotely via Wi-Fi, which allows the user to wear the system in mobile situations. ShowMe uses the Oculus Rift Development Kit which has a horizontal field of view higher than 90º and a diagonal field of view higher than 110º to create an immersive experience.

Camera Selection

Looking at the characteristics of the human eye, and with the aim of providing something closest to natural vision, we chose to implement a stereoscopic view using a converged setup, where the cameras are angled towards each other. We designed two 3D printed mounts that can slide over the front part of the HMD so as to adjust the convergence and the inter-pupillary distance (Figure 1).

We chose the cameras based on the given Oculus Rift specifications. Oculus Rift has a diagonal field of view of 90º, 60 FPS and a resolution of 640x800 per eye, which means an aspect ratio of 1,25:1. We decided to use a high-resolution webcam (Logitech Pro 9000) because it best matched the specifications of Oculus Rift and human eyes.

The Logitech Pro 9000 has an aspect ratio of 1,33:1 (approximately the resolution per eye of the Oculus Rift). The Logitech Pro 9000 is a USB 2.0 high-resolution webcam that can achieve 30 FPS, which is enough for the human eye. Moreover, we selected a pair of lenses with a diagonal field of view of 120º that mounted on a camera with an aspect ratio of 1,33:1 provides a 90º field of view.

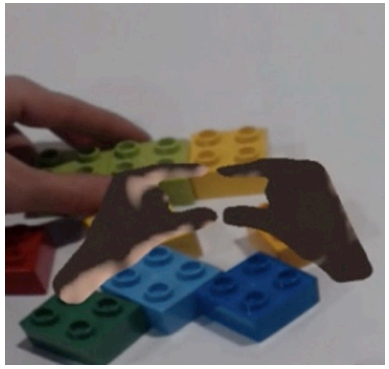

Figure 4. The novice user is building a Lego model while the expert is indicating how to join the pieces.

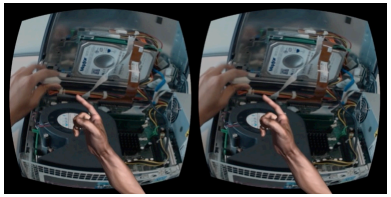

Figure 5. It is a screen capture of the Oculus Rift user's view.

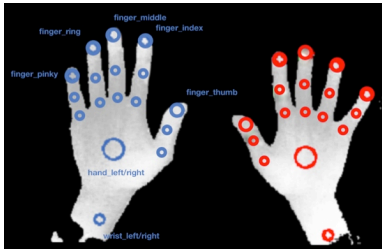

Figure 6. Depth data extracted from the Senz3D camera.

Hand Types and Tracking

In order to track the expert's hands, we fixed the Senz3D Depth Sensor Camera on top of the Oculus Rift with a 3D printed mount (Figure 1). We selected this sensor due to the fact that it is optimized for close-range interactivity and is a time of flight camera (TOF). TOF cameras are faster than other systems such as structured light cameras, and offer a simple and compact solution.

We developed different ways of displaying the expert's hand data and tested two different styles: "blob style" and "3D model style" (Figure 2). The first visualization, "blob style", is a blob stream where both hands have been cut out from the background image; they have natural skin color and no texture (Figure 4). The second hand representation is a 3D rigged model (Figure 5) of a hand where the data of the human hands are used to puppeteer the 3D mesh. This visualization is a rendering of the 3D hand model, with its own texture and applied lights.

We used the SoftKinetic SDK and Intel Perceptual SDK [6] to extract the data from the depth sensor and establish four types of outputs (blob information, geometric node, tracking result and pose, and gesture of the hands). Once the depth data is collected, we extract 17 points for each of the hands (Figure 6) and send it over a TCP/IP socket to be processed on the novice user's computer. Finally, we use the points to determine the pose of the hand model in Unity 3D.

Evaluation

As an initial evaluation, eighteen subjects (11 males, 7 females) were recruited in pairs to evaluate the system. We designed three exercises inspired by real-life situations where a person needs help from a remote expert, such as looking for an object in a room, repairing a device or assembling furniture. After performing all tasks, users filled a qualitative questionnaire summarizing their experiences and offering feedback for possible improvements. Overall, we received positive feedback, they expressed that ShowMe is a useful and helpful system for remote collaboration tasks. Some of the users acting as experts remarked that a first person point of view offers the best angle to teach another person, as the expert can see where the novice is looking and at the same time offers the same perspective for their own hands. Enabling the viewing of first-person hand gestures helped users to successfully perform the taks, which suggest that ShowMe can be effective for remote collaboration around manual tasks.

Future work

In future research we would like to add a depth sensor to the novice's HMD to detect the captured 3D model of the local environment and properly depth-merge the expert's 3D hands into it. We also intend to experiment with transparency of the remote expert's hand visualizations so as to prevent any occlusion from happening. Finally, we will also investigate upcoming transparent HMD technology as well as new generations of HMD with higher screen resolutions to push our approach further. Last but not least we are planning a larger user study to evaluate this approach.

Acknowledgments

We would like to thank the Fluid Interfaces group for their feedback and support and Roy Shilkrot for his proofreading of this document.

References

[1]. Rajinder S.Sodhi, Brett R.Jones, David Forsyth, Brian P.Bailey, and Giuliano Maciocci. BeThere: 3D Mobile Collaboration with Spatial Input. In Proc. CHI 2013, ACM Press (2013), 179-188.

[2]. Sunichi Kasahara, Jun Rekimoto. JackIn: Integrating First-Person View with Out-of-Body Vision Generation for Human-Human Augmentation. In Proc. AH 2014, ACM Press (2014).

[3]. A. Woods, T. Docherty, and R. Koch. Image Distortions in Stereoscopic Video Systems. In Proc. SPIE 1915 (1993), 36–49.

[4]. Seth Hunter, Pattie Maes, Anthony Tang, Kori Inkpen and Susan M Hessey. WaaZam! Supporting Creative Play at a Distance in Customized Video Environments. In Proc. CHI 2014, ACM Press (2014).

[5]. John C.Tang and Scott L.Minneman. VideoDraw: A Video Interface for Collaborative Drawing. In Proc. CHI 1990, ACM Press (1990), 313-320.

[6]. Stan Melax, Leonid Keselman and Sterling Orsten. Dynamics based 3D skeletal hand tracking. In Proc. GI 2013, ACM Press (2013), 63-70.

[7]. Keisuke Tajimi, Nobuchika Sakata, Keiji Uemura and Shogo Nishida: Remote Collaboration Using Real-world Projection Interface. In SMC 2010, IEEE (2010), 3008-3013.

[8]. Ohan Oda, Mengu Sukan, Steven Feiner, Barbara Tversky. Poster: 3D referencing for remote task assistance in augmented reality. 3D User Interfaces (3DUI), 2013.

[9]. David S. Kirk, Danaë Stanton Fraser. The Effects of Remote Gesturing on Distance Instruction. In Proc. CSCL 2005. ACM Press (2005), 301-310.

[10]. Pavel Gurevich, Joel Lanir, Benjamin Cohen, Ran Stone. TeleAdvisor: A Versatile Augmented Reality Tool for Remote Assitance. In Proc. CHI 2012, ACM Press (2012), 05-10.

[11]. Leonard Giusti, Kotval Xerxes, Amelia Schladow, Nicholas Wallen, Francis Zane, Federico Casalegno. Workspace Configurations: Setting The Stage For Remote Collaboration On Physical Tasks. In Proc. NordiCHI 2012, ACM Press (2012), 351-360.

[12]. Steffen Gauglitz, Cha Lee, Matthew Turk, Tobias Höllerer. Integrating the Physical Environment into Mobile Remote Collaboration. In Proc. MobileHCI 2012 ACM Press (2012), 241-250.

[13]. John C. Tang. Findings from observational studies of collaborative work. In Proc. Man-Machine Studies 1991, 143-160.

[14]. Franco Tecchia, Leila Alem, Weidong Huang. 3D Helping Hands: a Gesture Based MR System for Remote Collaboration. In Proc. VRCAI 2012, ACM Press (2012), 323-328.

[15]. Carl Gutwin and Reagan Penner. Improving Interpretation of Remote Gestures with Telepointer Traces. In Proc. CSCW 2002, ACM Press (2002), 49-57.

[16]. Darren Gergle, Robert E. Kraut, Susan R. Fussell. Action as Language in a Shared Visual Space. In Proc. CSCW 2004. ACM Press (2004), 487-496.

[17]. Laura Valenzeno, Martha Alibali, Roberta Klatzky. Teachers' gestures facilitate students' learning: A lesson in symmetry. Contemporary Educational Psychology (2003); Vol.28, 187–204.

[18]. Ivan E.Sutherland. A head-mounted three-dimensional display. In Proc. AFIPS 1968, ACM Press (1968), 757-764.

[19]. C.Comeau and J.Bryan: Headsight Television System Provides Remote Surveillance. In Electronics 1961, Vol.11 86-90.