Polly Wanna Show You: Examining Viewpoint-Conveyance Techniques for a Shoulder-Worn Telepresence System

Sven Kratz
Daniel Avrahami
Don Kimber
Jim Vaughan
FX Palo Alto Laboratory
3174 Porter Drive, Palo Alto, CA,
94304, USA
kratz@fxpal.com
daniel@fxpal.com
kimber@fxpal.com
jimv@fxpal.com

Patrick Proppe
University of Munich
Professor-Huber-Platz 2, 80539
Munich, Germany
patrick.proppe@gmail.com

Don Severns
RCManChild
San Francisco, CA, USA
rcmanchild@gmail.com

Abstract
In this paper we report findings from two user studies that explore the problem of establishing common viewpoint in the context of a wearable telepresence system. In our first study, we assessed the ability of a local person (the *guide*) to identify the view orientation of the remote person by looking at the physical pose of the telepresence device. In the follow-up study, we explored visual feedback methods for communicating the relative viewpoints of the remote user and the guide via a head-mounted display. Our results show that actively observing the pose of the device is useful for viewpoint estimation. However, in the case of telepresence devices without physical directional affordances, a live video feed may yield comparable results. Lastly, more abstract visualizations lead to significantly longer recognition times, but may be necessary in more complex environments.

Author Keywords
mobile; telepresence; view orientation conveyance; HMD; visualization

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

When working together in the same space, understanding what another person is looking at is relatively easy. However, when one person is remote and collaboration is mediated, sharing a common viewpoint is challenging. For example, in work by Fussell et al. [2], co-located pairs were significantly faster at completing a bicycle-repair task than pairs collaborating through video or through audio only, in part, due to the inability to know what is in the remote person's field of view.

In this paper we investigate the challenge of establishing a common viewpoint in a wearable telepresence system. Specifically, we consider the case of a telepresence system that allows a local person (*guide*) and a remote person to explore the same space. We use Polly [5], a wearable system consisting a smartphone on a stabilized gimbal that is carried by the guide on a wearable frame (see Figure 1). With Polly, the gimbal can be controlled remotely in two axes (pitch, yaw), allowing the remote user to choose their own viewpoint into the space. Polly's gimbal controller is run in *heading lock* mode, where the absolute orientation set by the remote user is kept constant, even if the guide's body rotates to a new orientation.

This ability of the remote user to control a personal viewpoint is one of Polly's key advantages, yet it creates a challenge for the guide to know what the remote person is looking at and vice versa. In our field studies, reported in [5], we observed this disconnect between the guide and the remote user, where the telepresence participants were often not aware of their counterparts' view directions. This diminished the sense of embodiment for the remote user and the sense of presence of the remote user for the guide.

As reported in [5], we were able to resolve much of the remote user's awareness of the guide's view orientation by visualizing for the remote user the relative orientations of the remote user and the guide. Using a mouse click, the remote user can realign her view with that of the guide. This is possible because the sensors on Polly's gimbal system provide the relative orientations of Polly's viewpoint and the guide's torso. However, conveying the remote user's view direction to the guide—the focus of this current work—is more difficult, as guides need to move their heads and torsos to change view direction, which, for obvious reasons, cannot be directly computer-controlled.

Figure 1: Guide with Polly device showing the remote user. The guide and the remote user are looking in different directions and may be unaware of each other's viewpoints.

In this paper we report findings from two user studies that we conducted in order to gain a better understanding of this challenge and investigate the use of visual feedback via a head-mounted display (HMD) to convey the view target of the remote user to the local guide.

In the first study, we tested how well subjects can estimate the view direction of a remote user by observing the movement of the Polly device itself. We compared this to the baseline co-located case, where the subject was asked to observe the view direction of another person standing in close proximity.

Since other telepresence devices (and future versions of Polly) may not rely on mechanical actuation and instead use a wide-angle or panoramic camera, it may not be possible to rely on the physical pose of the device for conveying a view point. Therefore, in a second study, we tested using visual-feedback mechanisms to convey the remote user's view orientation to the guide.

This paper makes the following key contributions: (1) We show that relying on the physical affordances of a telepresence device in sparse environments is a fast way for creating a shared viewpoint between a remote user and a local guide; (2) we show that providing a live view of the remote person's viewpoint provides sufficiently high performance in sparse environments if the device does not support observation of physical directionality; (3) we propose a set of visualizations that can be used, for example in more complex environments, that have slower task performance, but similar error rates.

Related Work

Providing remote telepresence through a wearable system has been the focus of several systems. For example, MH2 [11], Teroos [4], and the Polly system used here [5, 6] all involve a wearable system that can be operated by a remote user and be used for communications with the guide or 3rd parties. In our previous papers we describe the design and implementation of Polly and described its use for guided tours of museums and university campuses. These papers also report the issues of viewpoint alignment awareness between the remote operator and the local guide that we explore further in the present work. We note that while the MH2 and Teroos systems do not include a video-based representation of the remote user and lack attitude stabilization, the problem addressed in this paper and the solutions explored are is certainly also applicable to those systems.

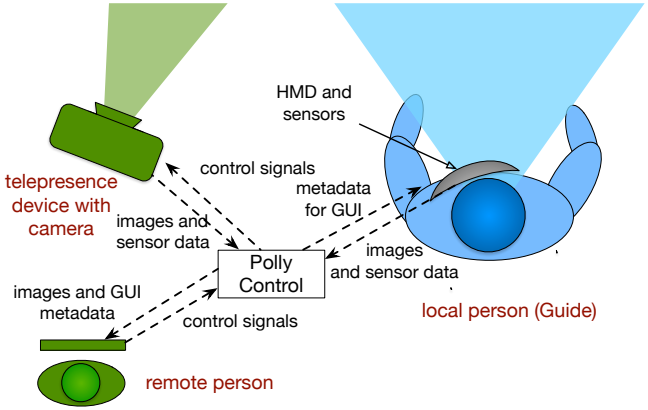


Figure 2: This architecture diagram of Polly shows the types of messages sent between the remote user, local user (guide) and the telepresence device.

There has been significant previous work addressing the issue of pointing to off-screen objects. Burigat and Chittaro conducted a user study on pointing to the location of external objects in 3D virtual environments [1]. They showed that inexperienced users performed best using a 3D arrow-

based visualization. Tonnis et al. developed a 3D arrow-based visualization for directing the users' attention in the direction of dangers in the environment [9]. For Polly, we have tried to avoid 3D-based visualizations and rely on 2D visualizations (see Figure 4). One reason to stay with 2D is that for Polly, we believe that conveying the relative horizontal view directions of the guide and Polly is more important than the vertical direction, as humans have a wider field of view in the horizontal than in the vertical axis [12]. The visualizations we use, *Cones* in particular, thus lay more emphasis on the horizontal plane.

In fields such as aviation, the problem of displaying multiple relative orientations has been solved with dial-style displays on instruments such as radio magnetic indicators [13]. Schinke et al. [8] proposed the "scene from top" visualization that, together with the Wedge visualization by Gustavson et al. [3], inspired the *Cones* visualization that we used in our second user study. The *Regions* visualization we propose is similar to minimap-based approaches [7, 8], although unlike in [7] and [8], Regions does not represent a geographic space but the space of pitch and yaw orientations—with the relative positions of Polly and the guide in that space indicated.

Viewpoint Conveyance Experiments

We conducted two user studies to explore the viewpoint-conveyance challenge between a local guide and a remote user. The first study serves as a baseline and compares the use of the Polly device, with its mechanical actuation serving to convey viewpoint direction, to the co-located case, where one person observes the gaze of another person. In a second study, we examine if visual feedback, presented through an HMD can be used to convey the view direction of the remote user to the guide, in case that physical cues are not available.

Study 1

In this study, we wanted to find out how well subjects can estimate the view direction of the remote user by looking at the orientation of a Polly device (Polly-to-Human, **P2H**) compared to the co-located case where standing next to another person and observing their gaze (Human-to-Human, **H2H**).

We hypothesize that, as humans, subjects would be quicker to infer the view target of the other human (**H2H**) and be slower and more error-prone when observing the Polly device (**P2H**). This study is motivated by our assumption that looking at another person to infer his or her view direction serves as a performance baseline for judging the speed subjects can infer the "view direction" of a mechanical device.

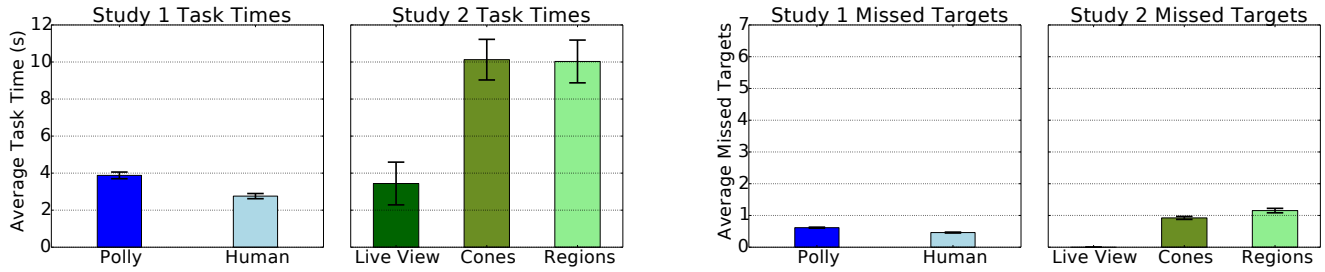
Participants

We recruited 13 participants from an industrial research lab, 3 were female. The age range was between 21 and 60 years old.

Apparatus, Task and Measures

Participants were asked to act as the Polly guide and try to identify at which artifact out of a set another person (either in the room, or represented by Polly) was looking at.

We used a medium-sized office lounge space (about 8×6 m in size) as our testing environment. All participants stood at a fixed location in one corner of the space. We distributed seven easily recognizable target objects (mainly colored plush toys) in the space. We took care to equally distribute the artifacts in height, distance and angle from the test subject. The horizontal angular range was 180° and the vertical angular range was 45°.



(a) Task Times for Studies 1 and 2

(b) Error Counts for Studies 1 and 2

Figure 3: The results of the two studies we conducted. (a) Shows the average task completion times for studies 1 and 2. (b) Shows the average error counts for studies 1 and 2. The error bars depict the standard error of the mean.

In the Polly-to-Human (**P2H**) condition, we used a Polly device mounted next to the test subject's left shoulder. The Polly device could be steered towards several predefined orientations via an application controlled by the experimenter.

In the Human-to-Human (**H2H**) condition, an experimenter acted as the human cue and stood 1 m to the left and in front of the test subject's left shoulder. The subject could use any cue (e.g., upper body posture or head rotation) provided to them by the experimenter to determine the target location. However, any type of explicit gesturing with any body part was disallowed. The experimenter was instructed to keep his hands at the sides of his body at all times.

The task for the test subjects was to observe the Polly device (**P2H**) or the other person (**H2H**), and identify what artifact was being viewed by the Polly device or or the other person. The experimenter manually measured the task completion time. The task completion time is determined as the time interval from setting the Polly or experimenter pose until the correct target is pointed out verbally by the test subject.

Design

We followed a within-subjects design, with each participant completing one trial for every target artifact in both conditions.

Results

Participants took an average of 3.88 seconds to identify targets in **P2H** and 2.76 seconds in **H2H** (see Figure 3(a)). A mixed model ANOVA on the task execution times (log-transformed) with Condition as a fixed effect and the User ID modeled as a random effect, shows that there is a significant difference between **P2H** and **H2H** ($\mathrm{F}_{1,156} = 40.69;\ p < 0.001$). Looking at error rates, a Wilcoxon Signed Rank test found no significant difference in error rates (Figure 3(b)) between the two conditions ($\mathrm{T}[11] = 7.5;\ p = 0.3$).

These results confirm our hypothesis and show that it takes Polly guides about one second longer to identify the remote user's view direction by observing the Polly device

compared to observing another human. However, we see that the identification time for **P2H** is still relatively fast. We use these results as an upper bound for gauging the performance of the visual feedback mechanisms evaluated in Study 2.

Study 2

Our first study showed that in our lab setup participants were able to infer the remote user's view target with relative ease just by looking at Polly's orientation. However, relying on a physical indication of a remote person's viewpoint may not always be possible. As mentioned earlier, future versions of Polly or other, similar telepresence devices may not use mechanical means of changing the view direction. Solutions using panoramic cameras or fisheye lenses with digital image stabilization—using a cropped image region for view panning—are likely. In this case, the guide would have no visual cues that could be utilized to infer Polly's view direction. Thus, in this study we examine alternative mechanisms to address the problem of conveying the relative view orientation of the remote viewer to the guide.

As a possible solution to this problem, we chose to examine the use of an HMD to provide visual feedback to the guide. In comparison to a phone or tablet, an HMD has the advantage of leaving the guide's hands free in order for her to interact with the Polly device or other objects in the environment, and does not require gaze shifts between a handheld device and the intended view direction.

Other technologies, such as tactile belts [10] could also be used to convey viewpoint orientation, but we feel HMDs are a more versatile device for the guide than belts as HMDs are easier to put on and to set up, and, more importantly, can provide functionality beyond viewpoint indication.

We use the HMD to show a visualization (Figure 4) of the relative alignment between Polly and the guide. The HMD we used was a Google Glass Explorer device, which has a display resolution of 640×360 pixels. To obtain the relative alignment data for Polly and the guide, we used two Bluetooth inertial measurement units (IMUs).[1] The reason for using external IMUs is that preliminary tests showed that the IMU on Polly's gimbal and the built-in IMU on Google Glass provide relatively noisy and drift-prone orientation data.

We implemented three visualizations:

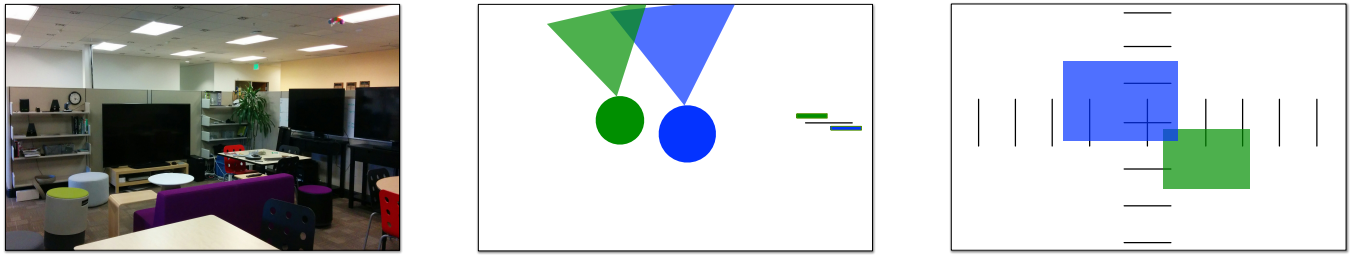
- *Live View* (**V1**, Figure 4(a)): Shows the remote user's video feed (i.e., what the remote person sees.)
- *Cones* (**V2**, Figure 4(b)): Shows the relative alignment of the guide and the remote user from above.
- *Regions* (**V3**, figure 4(c)): Shows the relative viewport alignment of the guide and the remote user on a plane.

Because the abstract visual representations provide a complete view of the orientations of the guide and remote user, we hypothesized that they (i.e., **V2** and **V3**) would have lower task completion times than Live View (**V1**).

Participants, Apparatus, Task and Measures

We recruited a new set of 13 participants from the same industrial research lab as in study 1. Two participants were female. The age range was between 29 and 58 years old.

[1] We used "YEI 3-Space Sensors" (http://ow.ly/BKJsu).



(a) **V1** (Live View) (b) **V2** (Cones) (c) **V3** (Regions)

Figure 4: The three visualizations studied in experiment 1. *Live View* (left) shows a live video feed from Polly's camera at the resolution of the HMD's display (in our case 640×360 pixels). *Cones* (Center) shows a top-down view of the scene, with the orientation and approximate view field and direction of the guide (blue) and Polly (green). The respective vertical view position is shown by small bars on the right of the displays that move on the vertical axis. *Regions* (right) shows a 2D representation of the field of view and direction of the guide (blue) and Polly (green).

We conducted this user study in the same setting as in Study 1, with identical positioning for participants, Polly device and target artifacts. However, as we wanted to evaluate the effect of the visualizations, the Polly device was hidden from the participants' view via a cardboard shield. Study participants were given the same task to perform—recognize which artifact the remote user is looking at—but using the HMD setup as described previously. We measured task execution times and the number of misrecognized view targets.

Design

Again, using a within-subjects design, participants were asked to perform one trial for every target in each of the three visual feedback conditions (**V1**–**V3**), for a total of 21 trials per participant. We note that in over 50% of trials in the Live View condition, more than one object was visible in the live camera feed.

Results

We used a mixed-model ANOVA with task time as the dependent measure, Visual Feedback condition (**V1**–**V3**) as a fixed effect, and user ID as random effect. Results show a significant effect of Visual Feedback condition on task time ($\mathrm{F}_{2,247} = 45.75$, $p < 0.001$). A Tukey HSD post-hoc comparison shows a significant difference between the Live View and Cones ($p < 0.001$) and between the Live View and Regions ($p < 0.001$). In both cases, Live View resulted in faster task execution time (see Figure 3a). This result was unexpected, as we hypothesized that the Live View would be less useful than the graphical visualizations.

Looking at errors, a Wilcoxon Signed Rank test found no significant difference in error rates (Figure 3(b)) for the three conditions (for the three matched pairs, T statistics were 7.5, 5.0 and 1.5).

Discussion

We conducted two studies to examine ways of conveying the view direction of the remote user to the Polly guide. The first study was a grounding experiment where we tested the capabilities of subjects to judge the remote viewer's view direction using either Polly or another human as a reference. In the second study, we asked subjects to use a visualization on an HMD to accomplish the same task.

The results of the first study were expected but promising: test subjects were able to acquire the view direction of the other human fastest, although the difference vs. the Polly device was only about one second. This shows that physical affordances on the telepresence device can be advantageous. However, as the results of the second study show, a live view may offer comparable performance, if distinct viewing targets are available.

However, in our field studies, which took place in relatively complex environments (both indoors and outdoors), the physical orientation of the smartphone was reportedly insufficient for establishing a joint viewpoint. Furthermore, as discussed previously, physical affordances may not always be available. For example, instead of a moveable smartphone, other systems may use a panoramic camera that continuously streams a 360° image from which the remote user controls only a virtual camera view. As the results of our second study show, a live view may offer reasonable task completion time, if distinct viewing targets are available. However, we believe that because we asked the users to look for distinct targets, and not only match up with Polly's view direction, the live video feed may have simplified the task. Again, we note that for the majority of trials, multiple targets were visible, so the live video feed cannot be regarded as a pure matching task.

We obtained the worst performance (highest task execution times) when participants used the abstract Cones and Regions visual feedback mechanisms. However, we believe that these might be useful in more complex spaces, or in cases where the remote user is not viewing a distinct target. In those cases, it may be harder for users to match the live view with an orientation in the space that they are in.

Finally, Our results indicate that while abstract visual feedback can be used as a fallback mechanism in ambiguous spaces with similar error rates, the target acquisition times will be significantly higher. The best option may be to combine the visualization mechanisms, i.e., overlaying a live video feed with a graphical visualization such as **V2** or **V3**.

Conclusions and Future Work

Establishing a joint viewpoint in collaboration through a remote telepresence system is important but often difficult. Through two controlled studies, we explored the use of physical indicators as well as approaches for visual feedback with a head-mounted display, useful when no physical cues in the device are available. In future work we plan to explore several important general questions raised by this work. For example, we plan research with the goal of better understanding what makes a space visually distinct, and whether this can be computationally estimated, e.g., through computer vision.

REFERENCES

1. Stefano Burigat and Luca Chittaro. 2007. Navigation in 3D virtual environments: Effects of user experience and location-pointing navigation aids. *International Journal of Human-Computer Studies* 65, 11 (2007), 945–958.

2. Susan R Fussell, Robert E Kraut, and Jane Siegel. 2000. Coordination of communication: Effects of

shared visual context on collaborative work. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 21–30.

3. Sean Gustafson, Patrick Baudisch, Carl Gutwin, and Pourang Irani. 2008. Wedge: clutter-free visualization of off-screen locations. In *Proc. CHI*. ACM, 787–796.
4. Tadakazu Kashiwabara, Hirotaka Osawa, Kazuhiko Shinozawa, and Michita Imai. 2012. TEROOS: a wearable avatar to enhance joint activities. In *Proc. CHI*. ACM, 2001–2004.
5. Don Kimber, Patrick Proppe, Jim Vaughan, Sven Kratz, and Don Severns. 2014. Telepresence from a Guide's Shoulder. In *Second Workshop on Assistive Computer Vision and Robotics at ECCV2014*.
6. Sven Kratz, Don Kimber, Weiqing Su, Gwen Gordon, and Don Severns. 2014. Polly: "Being There" through the Parrot and a Guide. In *Proc. MobileHCI, Industrial Case Studies*. ACM.
7. Virpi Roto, Andrei Popescu, Antti Koivisto, and Elina Vartiainen. 2006. Minimap: a web page visualization method for mobile phones. In *Proc. CHI*. ACM, 35–44.
8. Torben Schinke, Niels Henze, and Susanne Boll. 2010. Visualization of off-screen objects in mobile augmented reality. In *Proc. MobileHCI*. ACM, 313–316.
9. Marcus Tonnis and Gudrun Klinker. 2006. Effective control of a car driver's attention for visual and acoustic guidance towards the direction of imminent dangers. In *Proc. ISMAR*. IEEE Computer Society, 13–22.
10. Koji Tsukada and Michiaki Yasumura. 2004. Activebelt: Belt-type wearable tactile display for directional navigation. In *UbiComp 2004: Ubiquitous Computing*. Springer, 384–399.
11. Yuichi Tsumaki, Fumiaki Ono, and Taisuke Tsukuda. 2012. The 20-DOF miniature humanoid MH-2: A wearable communication system. (2012).
12. U.S. Department of Defense. 1998. Design Criteria For Military Systems, Equipment, And Facilities, MIL-STD-1472F. *Department of Defense Design Criteria Standard, Human Engineering* (1998).
13. Edward S White. 1969. RADIO MAGNETIC INDICATOR. (aug 1969). US Patent 3,460,146.