Wish You Were Here: Being Together Through Composite Video and Digital Keepsakes

Gina Venolia, John C. Tang, Kori Inkpen, Baris Unver
Microsoft Research
Redmond, WA USA
ginav | johntang | kori@microsoft.com, verxx003@umn.edu

ABSTRACT
We developed a prototype which overlays local and remote participants in a video call and enables them to take group pictures together. These pictures serve as keepsakes of the event. The application uses real-time chroma key background removal to composite the remote person into the scene with the local group. We tested the prototype in a museum setting, and compared it to a more standard picture-in-picture (PiP) model. Users rated the composite mode as being significantly more fun, creating a greater sense of copresence and involvement than the PiP mode. Composite snapshots were also strongly preferred over picture-in-picture. Based on results from the study, we added a pinch-zoom and positioning interface to make it easier to frame remote people together into the snapshot, and conducted a second study. We conclude that combining composite video calls and picture-taking on a mobile device enables people to socially construct a shared activity with a remote person.

Author Keywords
Video calling; snapshots; shared experiences, video conferencing, mobile video.

ACM Classification Keywords
H.5.3. [Information Interfaces and Presentation] Group and Organization Interfaces – Synchronous interaction.

INTRODUCTION
Group photos are an important way to mark shared activities with family or friends. Yet, there are many occasions when a member of the group is unable attend the activity in person. While commercial mobile video apps enable people to remotely connect for such events, they do not go beyond a video link for seeing each other, nor do they afford capturing a group photo as a keepsake of being there together. We developed WishU, a system to virtually include a remote person into an activity and into group photos. The system combines aspects of a video call and a camera application.

WishU enables a person at an activity to use their smartphone to initiate a video call with a remote person. The remote person answers the call on whatever device is available to them – smartphone, laptop, desktop computer, or smart TV. The system uses real-time image processing to remove the background from the remote person's video, composite it with the video from the local person's smartphone, and show the composite video to both the local and the remote people alike (see Figure 1). Either party can take a photo of the current state of the composite video. Photos are made available to both parties, to be used as keepsakes of the event.

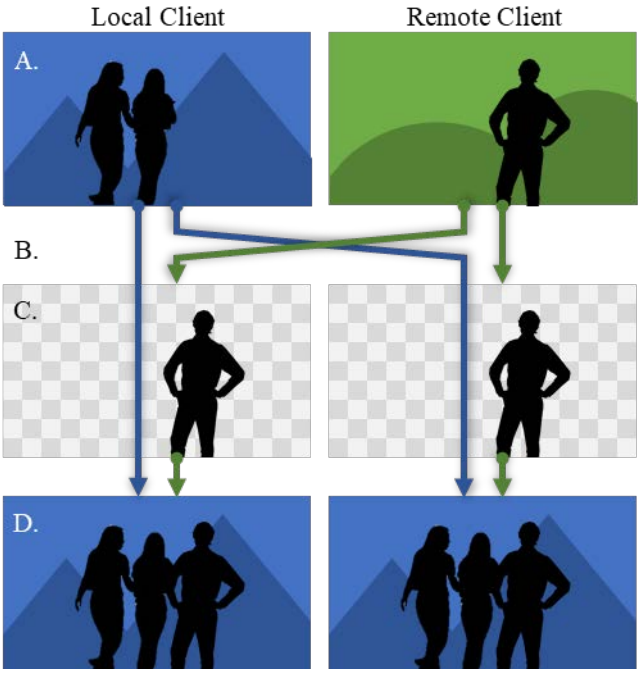

Figure 1: Functional schematic of the prototype. Each client (A) captures the camera stream, (B) exchanges it with each other using WebRTC, (C) performs chroma key background removal on the remote client's video, and then (D) overlays it on the local client's video.

We expected that a group of people at an activity would engage in their experience, find a moment that they

MobileHCI '18, September 3–6, 2018, Barcelona, Spain

ACM ISBN 978-1-4503-5898-9/18/09…$15.00
https://doi.org/10.1145/3229434.3229476

would want to share with friends or loved ones, or warranted a group photo, call up their remote loved one, talk with them and take a few photos, end the call, and then continue with their activity. This process might repeat whenever another group photo opportunity presented itself.

This experience comprises two features not typically found in today's video calling apps. First, instead of a traditional picture-in-picture view of the two video feeds, there is a single, composite video feed. Second, taking group snapshots is a main activity. While each feature in isolation is straightforward, we anticipated that the combination of would lead to a fun and engaging experience. We prototyped the system, evaluated it in a field trial, iterated on its design to allow visually positioning the remote's image in the composite, and conducted a second field trial.

RELATED WORK

As smartphones and networking connectivity that can support mobile video calling have become pervasive, there have been increased opportunities for sharing activities between remote people. O'Hara et al. [10] documented early experiences with mobile video, which focused more on keeping in touch over small talk than showing things to talk about. More recent research has focused on providing an embodiment that gives the remote person a physical presence in the local activity and enables people to more naturally interact together with a remote person [6]. Kim et al. [7] explored ways of increasing the involvement of the remote by sharing more contextual information of the mobile video partner, such as a map or a stream of high quality snapshots, to enable the remote person to take a more active role in the activity. These research experiences have demonstrated the potential for using video connections to involve remote people in mobile activities using the traditional picture-in-picture interface, with separate video windows for each participant. We focus our review of related work on video interfaces where the remote and local sides appear together and the role of taking pictures in sharing experiences.

Visually Appearing Together

Prior research has explored ways of visually incorporating remote people together into a common environment. The HyperMirror system [8] was an early prototype that evoked the experience of people in remote locations standing in front of a mirror together in a visually unified environment. The composite video was created by optically combining the two video feeds. Users could see each other and point at artifacts together in a shared frame of reference.

The *well* [11] took a different approach, uing video compositing software to create a variety of arrangements combining remote people together into a common visual view. They did not use background removal of each participant, instead relying on the visual juxtaposition of the participants in a display to evoke gazing down into a well to create an experience of sharing a visual space together.

The Being Here System [9] used Kinect depth cameras to separate the remote user's image from their background and superimpose their image into the local activity background. They found that their prototype significantly enhanced the perceived presence of the remote person, in comparison with conventional videoconferencing. The WaaZam system [5] also used Kinect to create a mirror mode where remote participants appeared together in the same environment as if looking into a mirror. They found that participants tended to focus on each other in this mode, whereas in other modes they offered, such as a digital or a customized environment, would draw more attention toward the space that they shared.

While these research prototypes demonstrated the potential of creating a visually shared user experience among remote video callers, they typically were restricted to specially equipped rooms due to the specialized video equipment needed to create the composite video effect. However, prior research has shown that many activities of interest occur out in the wild, which has recently become more accessible via mobile video [6, 10]. We believe there is an opportunity to visually include remote people into shared mobile activities.

Meanwhile mobile video applications such as Snapchat have popularized digital masks which can be inserted into a video stream and even track objects in the video to stay attached to them. While these masks show the increased capability of digitally compositing video streams, even on mobile devices, they have yet to be used to insert a dynamic video stream of a remote user into a video scene.

The Role of Pictures

In this age of widely accessible consumer photography, pictures play a major role in capturing and preserving memories of important social activities. The emergence of digital photography and social media as a platform to share pictures has increased the prominence of pictures in our social lives [1]. Despite the ubiquity of photos in our lives, little research has examined how taking photos affects our experience in an activity. In a range of field and lab studies, Diehl et al. [3] found that taking photos enhances the enjoyment of positive experiences and increases the displeasure of negative experiences. They concluded that these experiential effects focus greater visual attention on the activity, thereby increasing engagement with the experience.

The role of pictures seems more prominent in group activities. Wang et al. [13] examined the psychological effects of viewing selfies and groupies (including a group of people in selfies). Compared to viewing selfies, they found that viewing groupies correlated positively with perceptions of self-esteem and life satisfaction. Thus, taking pictures increases one's engagement with experiences and viewing pictures of being part of a group also has positive benefits.

Another indicator of the power of photos in social relationships is the use of life-sized cutouts for absent family members. Cutouts for U.S. military personnel serving abroad have been included in pictures of family events, such as weddings, memorial services, or just playtime with kids [15]. Capturing and sharing a picture which includes the missing family member, even when it is just a static picture cutout, provides strong emotional support for military families.

Commercial tools for capturing pictures that includes people on both sides of the smartphone have been explored in the marketplace. Samsung offered a phone with a Dual Camera feature that included pictures or video from both the front and back camera to include the person taking the picture into the scene being taken [12]. While this could be conceptually broadened to include a remote participant via video calling, the interface simply adds an inset picture of the remote person to the scene being captured, which does not effectively include the person into the scene.

A Design Gap

We believe the related work identifies a design gap which our prototype explores. Given the research documenting the engaging and psychologically beneficial impact of picture-taking on social events [1, 3,13] and the prevalence of apps focused on socially sharing pictures (e.g., Facebook, Snapchat), why do the current video calling applications (e.g., Skype, Facetime) not explicitly support taking a picture in the interface to capture sharing an experience with a remote person? And given the research showing the compelling nature of a video experience where local and remote participants appear together [5, 8, 9, 11], why have video calling applications stuck with a traditional layout of keeping participants in separate video windows? With the advent of real-time background removal technologies [2], we investigated what user experiences they could enable. We designed and developed WishU, a prototype to explore whether we can support close relationships over distance by combining picture-taking with a composite video interface.

PROTOTYPE

The WishU prototype was implemented as a web application, using Chrome and JavaScript. The system uses WebRTC and the PeerJS library to implement the video call capability, which runs at 1280×720 resolution. While we anticipate that digital video background removal will soon become a technology available in consumer devices, we used a green screen to prototype the user experience now to guide the design of using that technology in future apps.

To extract the remote person from their background, we used a greenscreen and a chroma key background removal algorithm. An initial calibration process calculated the mean and distribution of colors of the greenscreen. During the call, each client then examined each pixel of each frame of the remote person's video stream. A pixel was characterized as background if it was close to the mean color of the green background, foreground if it was distant from the mean color, or transitional if it fell into a middle range. Background pixels were set transparent; foreground pixels were set opaque; and transitional pixels were set translucent, proportional to their distance from the calibrated color. Additionally, to remove the green "fringe," transitional pixels colors were adjusted to limit the green component.

We created the composite video, shown in Figure 2, by simply overlaying the remote foreground video over the local video. After considering a variety of techniques for positioning the remote person's image relative to the local group's video, we chose to overlay the videos with a fixed alignment. Thus the remote person could change their position and scale within their video by physically moving relative to their camera.

For the purposes of our study, we also implemented a picture-in-picture (PiP) mode, shown in Figure 2. The local group's video was full-screen and the remote person's video appeared as a thumbnail, scaled to 25% in each dimension, in the lower-right corner of the screen. This approach is similar to a live streaming interface with a PiP inset to see the remote viewer, resulting in *identical* views for both the local and remote parties. This interface differs from typical video calling systems, where the other person's video appears full-screen with a thumbnail preview of the outgoing video, resulting in *different* views for each party.

The same web application ran on both the local party's smartphone (shown in Figure 2) and the remote person's device. One button started the call, a second button toggled between PiP and Composite mode, and a third button ended the call. During a call, both the local and remote applications streamed full-duplex audio and presented a full-screen view of the combined video, either Composite or PiP. Tapping anywhere in the video took a snapshot, playing a sound effect of a camera shutter click and displaying the snapshot on both clients for a few seconds. The smartphone client also had an additional button to change between front and back cameras. The accompanying video figure shows how the

participants switched between PiP and Composite mode, between front and back camera, and took snapshots.

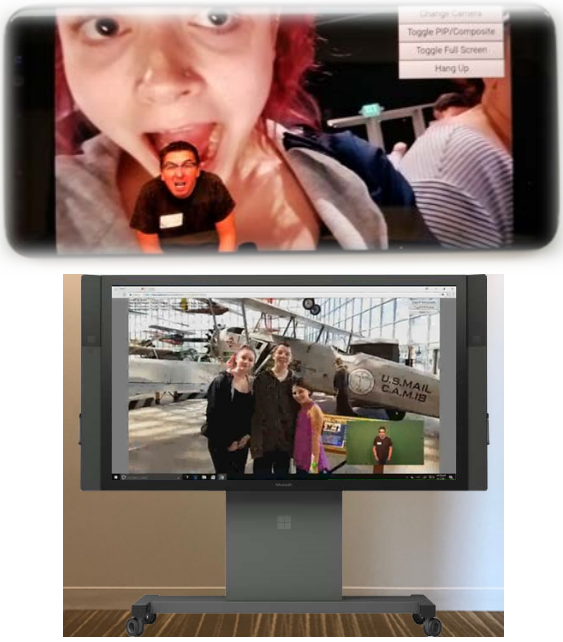

Figure 2: Comparison of the Composite mode on the local, mobile phone interface (above) and the PiP mode on the remote, touchscreen Surface Hub interface (below).

STUDY

We wanted to understand how groups of people would use these capabilities while engaged in a social activity. We performed a field trial of the system where a few friends or family members went to a museum, while one additional person from the group came to our lab to participate in the activity remotely. We chose the Museum of Flight in Seattle because it has displays that are interesting to a broad population and has high-quality Wi-Fi throughout the site.

We used our organization's usability recruiting department to recruit nine participants who were 18-60 years old, used video calling at least once a month, and would be interested in visiting the museum. We recruited a balanced number of males and females. These participants recruited 2-3 friends or family members to form their group. The resulting nine groups comprised 32 participants, 12 females and 20 males. Five were less than 11 years old, seven were 11-18 years old, five were 18-25, nine were 26-40, and eleven were over 40. There were five 3-person groups, three 4-person groups, and one 5-person group. Relationships among group members included parent/child, spouse, friend, coworker, and neighbor. Up to four participants per group were given a $100 gratuity each for their participation, which involved coordinating among a group for an hour-long time slot.

Each group selected one person to go to our lab as the remote person, and the rest met an experimenter at the museum. The lab participant, or *remote*, was situated in a room with a 55" Microsoft Surface Hub touchscreen (shown in Figure 2) mounted at a comfortable standing height, opposite a green background. The participants *local* to the museum were given a Samsung Galaxy S8 smartphone with the web application running full screen. They were instructed to click the button to call the remote person[1], and the museum experimenter introduced the interface for changing between PiP and Composite modes, taking photos, changing between front and back cameras, and starting and ending the call. After the initial introduction, groups were given 20 minutes to explore the museum and take snapshots using both modes as they wish, keeping the call open or starting/ending as desired. A video of the remote person in the lab as well as a screen recording of the prototype were recorded for analysis.

At the end of the 20-minute time period, the experimenters ended the call and gave each participant over 10 years old a survey. It drew upon the engagement & play and presence-in-absence subset of questions from the ABCCT survey [14], which included four paired questions where participants rated whether they had fun, felt involved, felt together, and felt comfortable with the level of control in both Composite and PiP modes. We collected 27 completed post-session surveys. Some participants did not answer some survey questions, so the number of responses varies by question.

After the survey, a semi-structured interview of about 5-13 minutes was conducted at each site (as a group at the museum), which was recorded for later analysis. We asked them to elaborate on which mode they preferred and discuss which features they liked or wanted to improve. It probed about their feelings of control, the effort in framing snapshots, and how they expected to use the snapshots.

Within an hour of the end of the session, each participant was sent a link to a webpage where they could download the snapshots that they took during their session. We offered participants 4"×6" printed photos of all their snapshots, which 11 people chose to receive. The photos arrived by postal mail about five days after the study. We gave participants the option of taking a follow-up survey after two weeks, for an additional $10 gratuity upon

[1]For this experiment, we preconfigured the smartphone client to call our lab, and the lab client to automatically answer the call, whereas a real system would provide a list of people to call and require explicit action to accept the call.

completion. 24 people opted in, and we received 13 completed surveys.

ANALYSIS

We analyzed the survey data according to responses for the PiP and Composite modes and whether the participant was at the museum or remote. We analyzed the follow-up survey according to whether they utilized the digital or print photos we made available to them. Wilcoxon signed ranks tests were used to compare between conditions.

The 410 snapshots taken during the sessions were coded as intentional (274), a test (49), or other (87). The "other" category consisted largely of duplicate photos involved in framing the snapshot just right, blurry photos that were replaced, or accidental touches that triggered a snapshot. Only intentional snapshots are considered for the remainder of the analysis. Each group took 17 - 50 snapshots (median= 27), or 0.93 - 2.50 snapshots per minute (median=1.37). One author coded each snapshot on a variety of factors:

- Mode: PiP or Composite
- Subjects: Whether the entire group, a partial group, just the remote or local members appeared in the snapshot
- Remote/Local Interaction: The type of interaction between the remote and local participants and the environment
- Playful: An obvious attempt do to something fun

Open and axial coding by one author was used to review the interview recordings and study observation notes to provide a qualitative context for our quantitative data. By leveraging the questions that participants rated in the surveys, we identified recurring themes during the interviews that helped explain the numerical ratings in the surveys.

RESULTS

Participants found the experience to be enjoyable; one said: *You have something cool – I mean to create this environment where people seem together – it's pretty cool* [6R[2]]. Even in the follow-up survey two weeks later, their enthusiasm was still strong: *It was an amazing experience, and I would love to do it again* [1M2]. The video figure shows examples of the participants interacting through the prototype.

Some people used PiP and Composite differently in support of their activity. Composite was generally preferred for taking snapshots. But some local participants switched to PiP when walking around the museum or showing something to the remote participant: *I liked the PiP mode like while walking to something, but actually wanting to take a picture and have them in it, the Composite worked better* [7M2].

When asked which condition they preferred, a Chi-squared analysis revealed that significantly more people stated a preference for Composite than PiP (93%, n=27, X^2=19.59, $p<.001$). The two participants who choose PiP stated that they felt the PiP mode better supported walking around the museum and showing things.

In general, the PiP mode was perceived as conventional, whereas taking snapshots in Composite mode made the experience unique and enjoyable. One participant said: *[I liked] the times when the background where they were at the [museum] allowed for some fun picture-taking. That was fun. Just getting from place to place and looking at stuff, you can do that with your iPhone* [8R].

Taking Composite Snapshots

Participants' preference for the Composite mode for taking snapshots over PiP was reflected in the relative proportion of snapshots taken in each mode. Of the 274 intentional snapshots, 252 (92%) were taken in Composite mode compared to 22 (8%) taken in PiP mode, which a Chi-squared analysis revealed was significantly different (X^2=193.07, $p<.001$).

This preference for Composite snapshots was also evident in their subsequent use. Twelve of the 13 respondents to the follow-up survey two weeks later reported downloading digital snapshots and/or receiving printed snapshots after the session (the one who did not access snapshots reported that she left it to her husband to utilize the snapshots). Of those 12, eight respondents shared or displayed Composite snapshots, two used both PiP and Composite snapshots, and none used PiP snapshots alone (the remaining two just looked at the downloaded snapshots).

We identified several reasons in the interviews for the preference of Composite mode over PiP: Composite was more enjoyable, created a greater sense of copresence, and afforded interaction of the remote person with the environment. However, composing Composite snapshots highlighted the asymmetry of control, and created framing issues. We consider each of these factors in turn.

Composite was More Fun

It was apparent from observing the sessions that participants enjoyed the experience – laughing, joking, and creatively working together to capture fun snapshots. Participants rated Composite significantly more **fun** than PiP (Z=-3.114, $p=.002$, Figure 3). All

[2]Participants are identified by group number (1-9), M for museum or R for remote, and, for the museum participants, a distinguishing digit. Their informed consent allows us to include their images in pictures and videos.

respondents rated the fun of Composite positively and 48% of participants gave Composite a more positive response than PiP, while the remainder were equal (none were less positive).

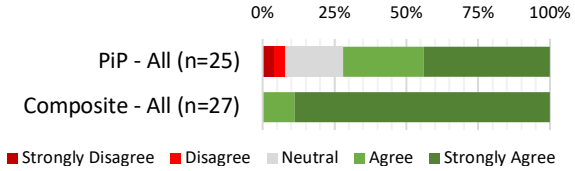

Figure 3: Likert responses to the survey questions, "I had *fun* with my remote partner(s) using the [PiP | Composite] mode."

We coded snapshots as *playful* when they exhibited some obvious attempt do to something fun (such as appearing to drive a vehicle on display or pretending to feed a snack to a remote participant or even eat a participant, as in Figure 2). 36 composite snapshots were coded as playful (14%, n=252), compared to only one playful PiP image (5%, n=22).

One participant said she preferred Composite: *because it was like he was there with us, and we could have more fun with it than when he was on the side [i.e. PiP]. Because then he could do silly things, or I could do silly things. It just seems like we're together* [5M1].

Composite Engendered Copresence

Participants rated Composite as creating a greater sense of **copresence** than PiP (Z=-3.868, p<.001, Figure 4). All respondents rated the copresence of Composite positively, except for two who were neutral. Relative to PiP, 73% of participants gave Composite a more positive answer and the remainder were equal (none were less positive). In the follow-up survey, one person reflected: *Overall, I think we were a lot more engaged with the person on the other end than I thought we would be.* [2M2]

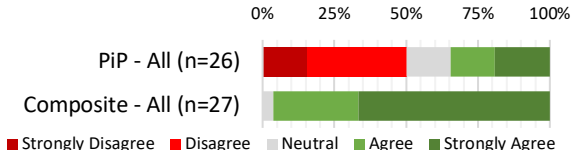

Figure 4: Likert responses to the survey questions, "I felt like I was *together* with my remote partner(s) at the activity while using the [PiP | Composite] mode."

For snapshots that showed both museum and remote participants, we coded the interaction among the people in the picture as one of:

- Pose: posing together for a picture (Figure 5A)
- Augmented: Going beyond posing in any way, such as pointing, appearing to hug the remote, mimicking figures in the museum (Figure 5B)
- Beyond Being There (BBT): Positioning the remote in a way that was not possible or permissible if physically there at the museum, such as sitting in a vehicle on display (Figure 5D), or appearing on the ceiling of the museum

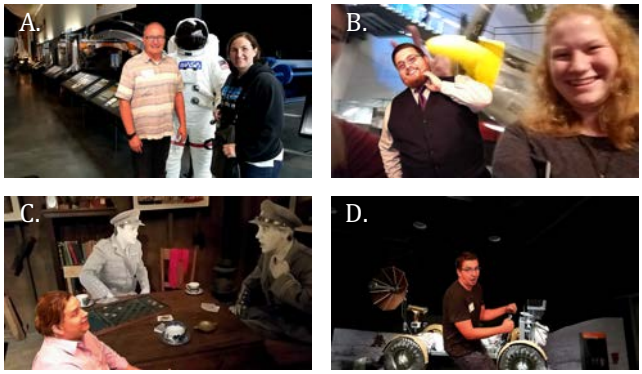

Figure 5: Examples of Composite snapshots: (A) Pose taken with the back-camera, (B) Augmented pose using front-camera, (C) Augmented pose showing only the remote person, and (D) Beyond Being There showing only the remote person.

There were more augmented and BBT interaction Composite snapshots (51 of 128, 40%) than PiP snapshots (4 of 20, 20%). However, a Chi-squared analyses did not find this difference to be statistically significant (X^2=2.917, p=.069).

One participant said that he preferred: *Composite mode, 100%. It actually felt like you were doing an activity together* [7M1]. Another likewise preferred: *Composite - I liked that because it was like I was there - does that make sense? It was very enjoyable. I came up with all kinds of neat ideas that I could use. I could use it as a teaching experience. I could use it with my mom and dad - they can't go anywhere, and they live over in [a place about 160 miles away]. And so, I could actually go through the [museum] with my dad and take pictures with my dad and mom. I just got goosebumps again. I know it was just a picture, but I still felt like I was there with her. That's what was really cool about it* [2R].

Composite Created a Sense of Involvement

Although both conditions helped participants feel involved with their partner(s), Composite created a greater sense of **involvement** than PiP (Z=-3.785, p<.001, Figure **Error! Reference source not found.**6). In the Composite condition, all respondents rated the involvement with their partner(s) positively. Relative to PiP, 72% of participants gave a more positive answer regarding Composite and the rest were equal (none were less positive). Figure 6 breaks out responses according to museum and remote participants, which indicates that this improvement was stronger for those in the museum.

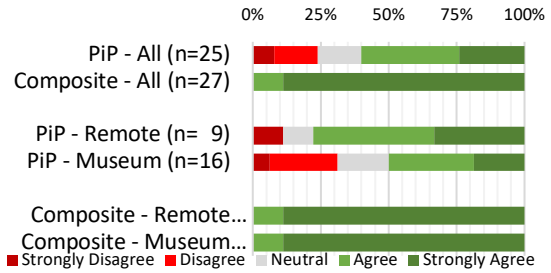

Figure 6: Likert responses to the survey questions, "I felt *involved* with my remote partner(s) in the activity while using the [PiP | Composite] mode."

For snapshots that only showed the remote without any museum participants, we used a similar coding scheme for the interaction of the remote person with the environment (illustrated in Figure 5 C-D). For Composite snapshots, 38% (45 of 119) went beyond just posing in the environment, to having the remote person interact with it. There were only two PiP snapshots taken with only the remote participant, and both were simple poses.

One participant described setting up a shot where the remote person became part of a museum diorama (shown in Figure 5C): *What we were trying to do is make it look like he was sitting at the table with the other people [in a museum display]. That actually worked out quite well. The soldier was sitting there, and he moved himself here, and now he's part of the scene.* [9M2].

4% of the Composite shots included interaction of the remote participant with the environment that would not have been possible if they were physically present (11 of 248 were BBT). A participant said: *I like the fact that it felt like I was there, that I was still able to participate in the antics. In fact, it added another layer of fun because physically when you're there you can't do these kinds of things. So, this added another dimension to the whole experience* [6R].

Control Issues

From observing the sessions, it was clear that the asymmetry of this activity could lead to an imbalance of control, where the local participants would have a much greater influence over the experience than the remote person. This was especially evident when composing the Composite snapshots. The local participants had control over their own movement and the timing and framing of snapshots. They would frequently instruct the remote about where to stand, how to pose, etc. (as shown in the video figure). There was also some verbal direction from the remote to the local participants, but much less. We wondered if this imbalance would have a negative effect on the experience, as speculated by one of the museum participants: *It felt like we had more control. I felt kind of bad for him because he just had to stand there the whole time. We could do whatever we wanted – we put him down the stairs one time* [2M2].

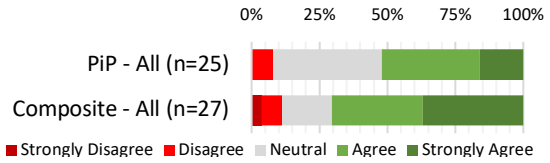

Figure 7: Likert responses to the survey questions, "I felt okay about my level of *control* of the experience in the [PiP | Composite] mode."

Although Figure 7 shows a higher percentage of participants rated themselves as more comfortable about their level of control in Composite (36%) than PiP (12%), it was not a statistically significant difference (Z=-1.588, p=.112). In the interviews, participants noted awareness of the control asymmetry, but did not express negative feelings about it. One local participant commented: *Looking back, I felt like we were just here together. I think you're right – we did have more control, but I think if he said, "hey, let's go over here and see this person" we would have* [2M1].

A remote participant commented: *I actually thought I had a pretty great deal of control. I could position myself in front of the camera. I could move in and out of the depth of field. It wasn't a huge amount of control, but it was a little bit. I was able to feel like I was participating with the activity* [1R].

Certainly, the open voice channel enabled control to be socially negotiated as needed. A remote participant said that the asymmetrical control: *is fine. They have the mobile device, so they're absolutely in control. If I say, "I want to look at this," or "turn the camera here," or "set this picture up," then it's still like I have control – not physical, manual control over the device but at least I can direct them* [6R].

Framing

Taking a group photo in-person can be a challenge – getting everyone in the shot and getting the background just right. It is even more challenging when there is a remote person in the shot, who must move to the correct position and size within the moving video image. Figure 8 shows the participants in the museum framing a snapshot.

Figure 8: Participants in the museum framing a Composite snapshot (left) and the resulting snapshot (right).

Many participants said in the interview that positioning the remote person by their physical movement in Composite mode was challenging, but that the challenge

was often part of the fun. One local participant said: *I won't use the word difficult, but I will use the word challenging. It was a learning experience. Because … she's here in this part of the picture, you have to move the camera in order to position her, and then sometimes you have to move the group to make it look right* [3M1].

A remote participant also expressed the challenge of framing: *Being able to position myself in the situation at a similar scale to them was challenging, because they were holding a camera a foot away from themselves and I was trying to position myself in front of a much more powerful camera* [1R], but went on to say: *The framing was one of the most fun parts, because it was like, oh, you know, I move, you move, I move my head closer and it just gets huge, it was kind of a fun experience* [1R].

Others found framing a Composite shot to be frustrating: *It was really hard for us to try to line things up. I was either too big or too small or I wasn't facing right or whatever* [4R].

Participants talked about alternatives to composing shots using the remote person's physical motion. Some mentioned manually positioning the remote person in the scene, for example by using a pinch-zoom interaction. Others suggested that positioning could be automated: *It would be neat if there was an autocorrect for sizing. If it could actually capture the size we are and then put him in there at the same size you wouldn't have to worry about it* [7M1].

We expected that WishU would be used primarily for group shots, including both local and remote participants. Of the 274 intentional snapshots, 48 (17.5%) included the entire group from both sites, 101 (36.9%) had the remote with some of the local participants, 121 (44.2%) showed only the remote, 3 (1.1%) showed only locals, and 1 (0.3%) did not include any people. Furthermore, 73% of the snapshots were taken using the back camera. This use of the prototype to take Composite snapshots of just the remote person in the museum environment suggests a design space outside of our initial focus. Perhaps this use grew out of a desire to create a keepsake to show that the remote person "was there" or the creative opportunities for placing the remote in the scene. This usage suggests an opportunity for elevating the focus and involvement of the remote person in the activity.

Keepsakes and Sharing

In addition to enhancing the experience in the moment by providing a shared activity, snapshots resulted in keepsakes for the participants. Furthermore, they shared photos outside the group to tell others about the experience.

The use of snapshots to reminisce within the group was anticipated by one participant in the post-session interview: *Probably [use the digital snapshots] as memories to look back on, like, "Hey, we were able to do this," in a different way than just using Snapchat, like, "Hey look at what I'm doing right now"* [1M1]. Another said, *I personally will upload them to my Google Drive and then years down the road [review them] as fun memories* [7M1]. A respondent to the follow-up survey reflected: *The best reactions were from those who I did the snapshots with … We were all having a good laugh and reminding each other what it was like doing the exercise.* [1M2]

Snapshots from the study were also shared beyond the participants. In the post-session interview, many participants anticipated how they would share them: *They will go on Facebook. They will go on Instagram. I will have to print them out and send them to my mom and dad* [2R]. Following the study, six (out of eight) participants reported sharing 31 downloaded snapshots (posting to social media, sending by email, etc.), while eight (out of ten) participants reported sharing 65 printed snapshots. One said: *Shared on WhatsApp with my different groups. Shared with friends by looking directly at my phone* [7R], while another commented: *Shared my son's pictures, most by regular mail, with grandkids who are now in college* [9M2].

DESIGN ITERATION

Since our study identified that the biggest concern about the user experience was the effort involved in positioning the remote person's image for framing snapshots, we made a design iteration that addressed that issue. Leveraging the pinch-zoom and dragging gestures used for sizing and positioning images on touch interfaces, we developed a manual positioning interface for the prototype. Either the remote or local participants could manipulate the remote's image to scale and place them in the composite view. This user experience gesture is demonstrated in the video figure, and some resulting snapshots are illustrated in Figure 9.

Figure 9: Snapshots where the remote was scaled and strategically positioned into a display at the museum (left) and scaled and placed into a group picture (right).

We conducted a similar study with eight different groups (2-4 people per group, 21 people total, 10-75 yrs. old) at the same museum, focused on participants' reactions to the positioning interface. In this second study, the prototype only showed a Composite image, but in two different modes. Fixed-Screen overlay mode relied on the remote user to position themselves in the scene by

physically moving around within the view, similar to how the Composite mode operated in the first study. Positioning mode enabled scaling the remote user's image through a pinch-zoom gesture and positioning it to any place within the scene by dragging. The user at either site could manipulate the image position. In this study, the remote user was seated in front of a touch-enabled laptop, rather than standing in front of a TV-sized screen, to make it easier to perform the positioning gesture.

Participants spent about seven minutes in each condition (counter-balanced on which condition they experienced first) and in the last seven minutes, were invited to freely choose which mode they preferred to use. Participants completed a survey of questions about their preferences and joined in a group interview for about 9-20 minutes to elaborate on their survey reponses. In this study, all participants came together to the museum, and one person was selected as the "virtual" remote by going to an isolated room in the museum. For the one hour time slot, each participant was given a $75 gratuity for their participation, since the logistics were somewhat simpler than the earlier study.

The participants continued to positively rate the prototype as a fun experience, with all participants rating it positively (9 Somewhat Agree and 12 Strongly Agree). They also expressed a preference for the Positioning mode. The survey question asked, "I liked being able to Position the remote person's image by pinch zooming and panning within the view". All participated rated the positioning mode positively (16 Strongly Agree and 5 Somewhat Agree).

We expected that the Positioning mode would affect the control dynamics between the sites, as digitally placing the image of the remote could be done from either site. Although the survey responses about the comfort with the level of control between the two conditions indicated that they felt like they had more control in the Positioning mode, this difference was not statistically significant (Z=-1.028, $p<.304$**Error! Reference source not found.**) (Figure 10).

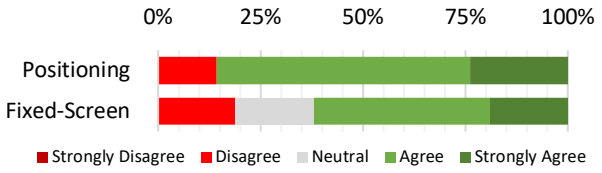

Figure 10: Likert responses to the survey questions, "I felt comfortable about my level of *control* of the experience in the [Positioning | Fixed-Screen] mode."

The interviews also indicated that there were some issues about negotiating who was in control of the positioning gesture at any given time. Comments from the participants suggested: *Maybe there should be some kind of button where you could turn off one person's control* [2R], and *It'd be interesting if you could toggle the control* [5R].

Another participant commented on the negotiation needed between the users: *We have to communicate, who is controlling. It's like OK, I'm not going to move you anymore, you move yourself wherever you want to be…* [5M1]

Since either side could control the positioning of the image, the interface needs to be more clear about when someone else is actively engaged in adjusting the positioning to support that coordination. While the prototype succeeded in giving both sides positioning control, the usability of the interface needs to be refined to help them coordinate using it.

Having addressed the most substantial issue from the first study of being able to digitally position the remote's video image, people raised other issues to address. Participants commented that differences between the sites in resolution, lighting, color balance, etc. detracted from the sense of looking like they were together in the snapshots. People also wanted to adjust the overlay layer so that they could control which site would be visually in front of the other.

DISCUSSION

This work combines two well-known techniques – composite video calls and photos – to produce a novel and compelling experience. Our prototype combines the benefits of both of these features to enhance remote engagement in an activity. Enabling these features on a mobile device, which has the flexibility of going wherever there is an activity of interest with network connectivity, builds on prior work that showed the potential of sharing an experience with remote people over video [6, 10]. Participants reported that they had fun, were engaged, and felt like they were together, all while connecting between remote sites. Just as prior work [3] found that taking photos increases user engagement with the experience in person, enabling taking snapshots in a video call offers strong engagement in the shared remote activity.

We believe it is this combination of visually appearing together and taking pictures that makes the experience compelling. Building on research distinctions between space and place [4], our prototype creates a shared *place* from remote *spaces* by providing a visual illusion of being in the same space and enabling the social practice of taking pictures together which is often a marker for meaningful shared interaction in a place. While our protoype uses very straightforward technology, the shared social place it generates is remarkable for enabling close ties to enact shared activities and create shared memories, even while separated by distance.

While our prototype was designed to enable remote people to take group pictures together, we were somewhat surprised that 44% of the snapshots only included the remote person. Upon further reflection, only a fraction of the pictures taken while visiting a place together in person include the entire group of people. Nonetheless, the practice of taking pictures imbues social meaning to the event, so our study helped us to see that we should not overly focus on including all the people in the picture. Indeed, the affordance of creatively positioning the remote's image within a scene enabled a new form of engaging interaction which elevates the remote person's involvement in the activity.

While control is unequally distributed between the two sites, since the people in the museum control what is in view of the mobile camera, we explored how people *felt* about that asymmetry of control. While the survey responses graphed in Figure 7 show that more responses were positive than negative in the first study, we expected that the design iteration of adding Positioning mode would change the control dynamics. Since users at either site could position the remote's image in the scene, we were curious how that affected their comfort with their level of control. The survey responses graphed in Figure 10 suggest that people felt better about the level of control in Positioning mode compared to Fixed-Screen mode. While the fundamental asymmetry of control persists, adding more agency within that constraint could improve users' perception of control.

Our research suggests several opportunities for future work. First, we used simple chroma keying to enable overlaying the remote into a composite video scene, but other technologies for background removal could provide the remote person with more flexibility about where and what device they could use to join the experience. Technologies are emerging that enable real-time digital background removal (e.g., [2]), which we would expect to be available in consumer devices in the near future. This capability would enable either side to be the host site where remote participants are integrated. This also suggests exploring other endpoint device pairs (e.g., phone-phone or laptop-laptop) to discover what interactional and technological issues might arise.

Second, although our study explored one social / leisure activity (visiting a museum), future work should examine other leisure or even work opportunities for this type of interaction, such as guiding people in repair or Do-It-Yourself activities. The experience may also vary according the user demographics (e.g., grandparents with grandkids, teens at social activities). Studies with more people in each age range are needed to explore the effects of age, and we are especially interested to see how youth, who already use a wide range of video apps, would react to the experience. Finally, it is important to explore how this functionality would be used if broadly deployed in current video calling and social media sharing tools with the opportunity to share with their social graphs of connections.

Exploring the combination of composite video and picture-taking helped us identify a social approach to creating a compelling sense of remote people being there together. Our prototype leverages commodity computing devices and existing use practices with them. Picture taking also taps into existing social practices around reminiscing with those who shared the experience with you and sharing photos with others to tell them about a social experience. In some ways, our prototype demonstrates a "discount" way of *being there* that relies more on social practices than sensory fidelity to create an experience of being together. Similar to how the Flat Daddy phenomenon leverages simple picture cut outs of absent military family members to include them in meaningful social events [15], the emotional response in WishU came not so much from the fidelity of the representation, but in the social meaning imbued by *including* the missing friends and the resulting keepsake. Perhaps this approach opens up new design spaces for creating social experiences of being together.

ACKNOWLEDGEMENTS

We thank the anonymous participants in the field deployments who helped us explore the potential of this concept. We also thank the Museum of Flight in Seattle, Washington, USA for providing a rich environment for conducting this study.

REFERENCES

1. Morgan G. Ames, Dean Eckles, Mor Naaman, Mirjana Spasojevic, and Nancy Van House. 2010. Requirements for mobile photoware. *Personal and Ubiquitous Computing*. 14. 95-109. http://dx.doi.org/10.1007/s00779-009-0237-4
2. Sebastian Brutzer, Benjamin Höferlin, and Gunther Heidemann. 2011. Evaluation of background subtraction techniques for video surveillance. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2011).* 2011, 1937-1944. doi: 10.1109/CVPR.2011.5995508
3. Kristin Diehl, Gal Zauberman, and Alixandra Barasch. 2016. How Taking Photos Increases Enjoyment of Experiences. *Journal of Personality and Social Psychology*, 111(2), 119-140. http://dx.doi.org/10.1037/pspa0000055
4. Steve Harrison and Paul Dourish. 1996. Re-Place-ing Space: The Roles of Place and Space in Collaborative Systems. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW '96).* 1996, 67-76. http://dx.doi.org/10.1145/240080.240193

5. Seth Hunter, Pattie Maes, Anthony Tang, Kori Inkpen, Sue Hessey. 2014. WaaZam! Supporting Creative Play at a Distance in Customized Video Environments. In *Proceedings of the Conference on Human Factors in Computing Systems* (*CHI 2014).* 2014, 1197-1206. http://dx.doi.org/10.1145/2556288.2557382
6. Kori Inkpen, Brett Taylor, Sasa Junuzovic, John Tang, and Gina Venolia. 2013. Experiences2Go: Sharing Kids' Activities Outside the Home with Remote Family Members. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW 2013).* 2013, 1329-1340. http://dx.doi.org/10.1145/2441776.2441926
7. Seungwon Kim, Sasa Junuzovic, and Kori Inkpen. 2014. The Nomad and the Couch Potato: Enriching Mobile Shared Experiences with Contextual Information. In *Proceedings of the 18th International Conference on Supporting Group Work (GROUP 2014)*, 167-177. http://dx.doi.org/10.1145/2660398.2660409
8. Osamu Morikawa and Takanori Maesako. 1998. HyperMirror: Toward Pleasant-to-use Video Mediated Communication System. In *Proc. CSCW '98*, ACM (1998), 149-158 http://dx.doi.org/10.1145/289444.289489
9. Mamoun Nawahdah and Tomoo Inoue. 2012. Being Here: Enhancing the Presence of a Remote Person through Real-Time Display Integration of the Remote Figure and the Local Background, *The Open Virtual Reality Journal,* June 2012, 17, 2, 10 pages.
10. Kenton O'Hara, Alison Black, and Matthew Lipson. 2006. Everyday Practices with Mobile Video Telephony. In *Proceedings of the Conference on Human Factors in Computing Systems* (*CHI 2006).* 2006, 871-880. http://dx.doi.org/10.1145/1124772.1124900
11. Nicolas Roussel. 2002. Experiences in the Design of the Well, a Group Communication Device for Teleconviviality. In *Proceedings of the ACM International Conference on Multimedia (Multimedia 2002),* 146-152. http://dx.doi.org/10.1145/641007.641036
12. Samsung. 2014. How do I use the Dual Camera feature on my Samsung Galaxy S5 to record video and capture photos using the front and rear cameras simultaneously? Retrieved September 12, 2017, from http://www.samsung.com/ie/support/skp/faq/1053544
13. Ruoxu Wang, Fan Yang, and Michel M. Haigh. 2017. Let me Take a Selfie: Exploring the Psychological Effects of Posting and Viewing Selfies and Groupies on Social Media. In *Telematics and Informatics*, 34, 4, July 2017, 274-283. https://doi.org/10.1016/j.tele.2016.07.004
14. Svetlana Yarosh, Panos Markopoulos, Gregory D. Abowd. 2014. Towards a questionnaire for measuring affective benefits and costs of communication technologies. In *Proceedings of the 2014 Conference on Computer Supported Cooperative Work (CSCW 2014).* 2014, 84-96. https://dx.doi.org/10.1145/2531602.2531634
15. Katie Zezima. 2006. When Soldiers Go to War, Flat Daddies Hold Their Place at Home. In *New York Times*, September 30, 2006. Retrieved September 12, 2017 from http://www.nytimes.com/2006/09/30/us/30daddy.html