Mobileportation: Nomadic Telepresence for Mobile Devices

JACOB YOUNG, TOBIAS LANGLOTZ, STEVEN MILLS, and HOLGER REGENBRECHT,
University of Otago, New Zealand

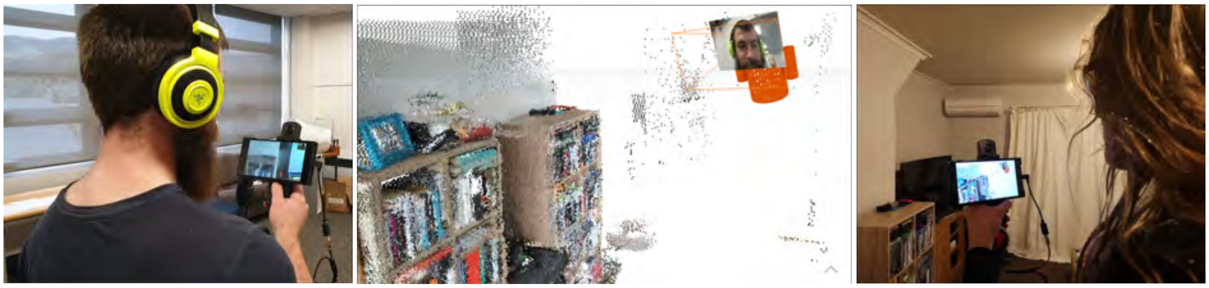

Fig. 1. Two users at different locations communicate within a shared environment constructed using consumer-grade mobile phones with an integrated RGBD sensor and external 360° camera. Here, the local user (right) shares his environment with the remote user (left). Both "meet" in an incrementally constructed shared virtual space (centre) where they can freely move between ego- and exocentric viewing positions. This reconstruction can be created and transmitted in real time over a standard cellular network, and each user's position within it is shown as an avatar to allow for nomadic face-to-face communication.

The desire to stay connected to one another over large distances has guided decades of telepresence research. Most focuses on stationary solutions that can deliver high-fidelity telepresence experiences, but these are usually impractical for the wider population who cannot afford the necessary proprietary equipment or are unwilling to regress to non-mobile communication. In this paper we present Mobileportation, a nomadic telepresence prototype that takes advantage of recent developments in mobile technology to provide immersive experiences wherever the user desires by allowing for seamless transitions between ego- and exocentric views within a mutually shared three-dimensional environment. The results of a user study are also discussed that show Mobileportation's ability to induce a sense of presence within this environment and with the remote communication partner, as well as the potential of this platform for future telepresence research.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *User studies*; *Collaborative and social computing systems and tools.*

Additional Key Words and Phrases: co-presence, 3D reconstruction, video communication, augmented reality

Authors' address: Jacob Young, jacobyoung.research@gmail.com; Tobias Langlotz, tobias.langlotz@otago.ac.nz; Steven Mills, steven@cs.otago.ac.nz; Holger Regenbrecht, holger.regenbrecht@otago.ac.nz,
University of Otago, P.O. Box 56, Dunedin, New Zealand, 9054.

2474-9567/2020/6-ART65 $15.00
https://doi.org/10.1145/3397331

ACM Reference Format:
Jacob Young, Tobias Langlotz, Steven Mills, and Holger Regenbrecht. 2020. Mobileportation: Nomadic Telepresence for Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 65 (June 2020), 16 pages. https://doi.org/10.1145/3397331

1 INTRODUCTION

Today's computing power, the availability of fast data networks, and the spread of wireless connectivity to even remote areas allows us to virtually connect people wherever they may be through technologies such as Zoom, FaceTime, and Skype. Further technological advancements could make this experience even more realistic by giving users the impression that they really are *spatially present* (the sense of "being there" [4]) in another place, or *socially present* (feeling that their communication is unmediated [18]) and *co-present* (mutually attentive and entrained [5]) with their remote communication partner.

Many solutions have been proposed that attempt to give users this feeling of presence by either surrounding them with sensors or tethering them to stationary systems, but with how pervasive mobile phones have become it seems unlikely that users would revert to a desktop computer as their primary telecommunication device. This has led to mobile devices being integrated into several advanced telepresence prototypes as simple portable displays powered by stationary hardware [23, 26], though this nullifies any advantages in mobility and spontaneity that mobile phones can provide [2].

We present *Mobileportation*, a prototypical system that provides an immersive nomadic telepresence experience on purely mobile devices. We use a mobile phone with an in-built RGBD sensor to incrementally create a 3D reconstruction of the *local user*'s environment as they move around it, which the *remote user* can freely explore in 6DoF (*Degrees of Freedom*) in an *exocentric* view (from *outside of* the local user's position). An attached 360° camera also captures a live feed of the space encompassing the local user, allowing for a higher resolution *egocentric* view of the environment (from *within* the local user's position) revisionby simply walking to their current position. Each user's face is also captured and spatially displayed in the shared environment on a virtual avatar, allowing Mobileportation to provide the face-to-face communication users expect from such systems while also giving them complete freedom to move around the shared environment independently.

We show that despite the technical limitations of mobile devices this experience can be achieved in real time and with minimal latency over a cellular network, allowing for ad-hoc immersive communication and exploration of remote environments without requiring any additional hardware, and provide technical details on how this was accomplished. We also present the results of a user study showing that despite the reduced visual fidelity of its incrementally constructed environments, Mobileportation is similar to video-only 360° mobile videoconferencing systems in its ability to invoke a sense of presence between its users and within the virtual space while providing a significantly more social experience with more freedom for its users. Finally, we explore the possible innovations that this hardware platform could provide in the future, showing how this application can be extended to support advanced features such as view-dependent rendering, full-body capture and display, and 3D hand gestures once mobile phones become sufficiently powerful.

2 RELATED WORK

Many approaches have been proposed that attempt to bring distant people into a user's local environment. Tang *et al.* [28] explored how a 360° camera could be used to enhance standard videoconferencing by allowing a remote user to obtain independent views within a shared environment. Such views have been shown to significantly increase users' spatial awareness of the environment [14] and spatial presence within it [21] as well as allow them to guide the topic of conversation rather than have it dictated to them. While rotational independence can thus be achieved, this only shifts the asymmetry of interaction as the remote user still has no control over their position within the space.

The solution to this asymmetry is often to create a three-dimensional reconstruction of the users or their environment and combine the two in an exocentrically viewed space. Holoportation [22] attempted to virtually bring the remote user into the shared space by volumetrically scanning and displaying them within the local user's Hololens display through several custom RGB-D camera arrays placed about the space. This volumetric reconstruction also required expensive and powerful desktop computers, severely limiting the system's user base and potential use cases by raising the cost of entry.

Many systems use this outside-in method of depth sensor placement to create these reconstructions [1, 10, 19] as it allows the full space to be visible once communication starts, however this environment and the users within it are constrained to the limited area covered by the cameras. To remove this constraint, technologies such as KinectFusion [13] or photogrammetry [24] can be used, which combine multiple captures from a moving Kinect sensor or RGB camera into a coherent 3D mesh to create larger interaction spaces. This requires a lengthy pre-construction stage before conversation can begin, and any updates to the environment will not be reflected in the reconstruction, making this method unsuitable for dynamic spaces where the task object or topic of conversation is likely to be moving. Some recent work explored how light fields could be used in place of a mesh for object-focused interaction [20], and while this offers advantages in visual quality even on mobile devices, it is unlikely that this can be extended to room-scale environments.

The alternative is to stream the RGB-D data to the remote user as it is captured in real time, dynamically reconstructing the environment as the local user walks around it. Stotko *et al.* use this method in SLAMCast [27], which combines real-time data captured from a handheld Kinect sensor to allow the remote user freeform exploration of unconstrained environments. The size of this environment is still limited given that the Kinect must be tethered to power and a stationary computer; the authors considered using a mobile phone as their capture device, though this was not further explored. Creating the environment in this way also means that the remote user must either wait until all areas have been captured or be satisfied with an incomplete model.

The ideal system would combine the benefits of these various depth capture methods, allowing free and immediate movement within an arbitrarily sized environment while also allowing it to be shown and updated in real time. Komiyama *et al.* proposed such a system with JackIn Space [15], which reconstructed a small area to be shared by placing several RGBD cameras around it. Users could explore this limited space, or could transition to an egocentric view of it captured by a fisheye camera attached to the local user's head, allowing them to view the space's wider context. This transition was appreciated by users, though the computer-mediated mechanism to do so was not as well regarded.

Teo *et al.* [29] proposed hand gestures as a more natural way to transition between these ego- and exocentric views. Users found an object search task faster and significantly easier when in an egocentric view, with social presence and the ability to gauge their partner's focus rated significantly higher in that mode. However, the exocentric view was overall preferred by users, while the ability to freely transition between the two was significantly preferred over either view alone.

To the best of our knowledge such an experience has yet to be achieved using purely mobile hardware, which would make it immediately available to most of the population and as convenient and effortless as mobile video calling is today. The inherent portability of these devices would mean that true exocentric views within arbitrarily sized environments could be attained without being constrained by tracking spaces or cables, but despite this the few existing purely mobile approaches [21, 28, 31] only allow egocentric viewing. In this work we close this gap by providing a prototypical solution that provides this real-time exocentric and egocentric exploration with spatially-rendered avatars on consumer-grade hardware.

3 MOBILEPORTATION SYSTEM OVERVIEW

We present *Mobileportation*, a novel approach for nomadic mobile telepresence that combines the strengths of 2D and 3D environmental reconstructions to enable a new and novel communication experience not yet seen on mobile devices. Combining different environment representations in this way has several advantages: panoramic representations are easy to build but are usually only valid from the point of view of the panoramic camera [31], meaning the remote user is not free to stray from the local user's position without creating visible distortions. 3D reconstructions enable this straying through 6DoF exploration, though usually at the cost of either visual fidelity or the size of the explorable area.

To combine these two reconstruction methods, Mobileportation provides two ways of viewing the environment. No intentional mode-switching is required; rather, the way the environment is viewed depends on the distance between users, allowing them to focus on exploration rather than operating the application. These two viewing modes and the transition between them are illustrated in Figure 2 and are as follows:

- *Exocentric View:* A 3D reconstruction of the local user's surroundings is incrementally captured using an RGBD sensor embedded in their mobile phone. Both users may then move about this shared virtual space by simply walking around their real one, with 6DoF tracking provided by the same RGBD sensor. Each user's current position is shown as a 3D avatar with their face overlaid on top of it to allow for face-to-face communication.
- *Egocentric View:* Live 360° video is shown from the local user's position to provide an immediate, high-detail view of the environment that users can independently explore in 3DoF by rotating their phone. Any sufficiently close 3D data will still be visible, and each user's face is now shown in the top-right corner of their partner's display.

To transition from an exocentric to an egocentric view, a user simply has to walk toward their partner's avatar. As the distance between them decreases, the 360° video will slowly fade in and a "snap in" mechanism will smoothly transition the two users together. Similarly, when in this egocentric view either user can simply walk away from their partner to gradually transition back into the exocentric view. In either mode, users may speak to each other via integrated voice chat no matter the distance between their virtual positions.

All that is required to experience this application is a mobile phone with an integrated RGBD camera and a 360° camera. While this may seem a rare combination, recent trends toward integrating depth cameras and wide-angle lenses in mobile phones such as the Galaxy S10+[1], or including modular 360° cameras such as with the Essential Phone[2], suggest that these features will soon be available in one self-contained device.

4 IMPLEMENTATION

A Lenovo Phab 2 Pro[3] is used for all capturing, rendering, tracking, computation, and display with no external computer required. Depth and tracking data are acquired from the phone's inbuilt ToF (*Time-of-Flight*) sensor, accessed via Google's Tango API. These are combined with images from the inbuilt RGB camera to create coloured, oriented point clouds of the user's surroundings, which are then sent to the remote user and integrated into a cumulative point cloud stored on each device's GPU.

360° video is captured using a Ricoh Theta S[4] connected to the phone via USB using the UVCCamera library[5] and attached using the purpose-built 3D-printed mounted shown in Figure 3. Captures from this camera are sent

[1] https://www.samsung.com/us/mobile/galaxy-s10/
[2] https://www.essential.com/
[3] https://www.lenovo.com/us/en/smart-devices/-lenovo-smartphones/phab-series/Lenovo-Phab-2-Pro/p/WMD00000220
[4] https://theta360.com/en/about/theta/s.html
[5] https://github.com/saki4510t/UVCCamera

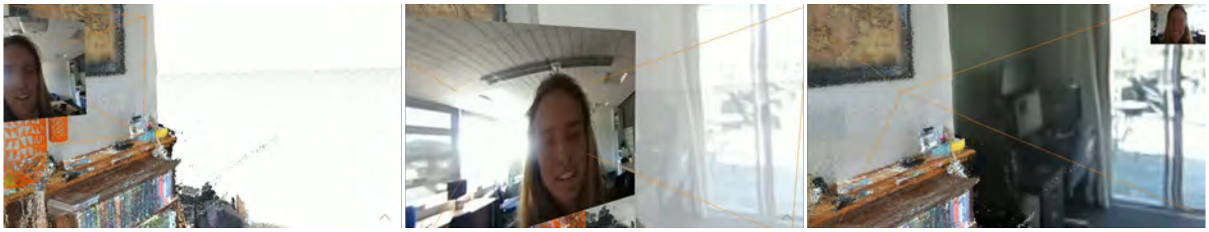

Fig. 2. The user's perspective of the environment. (Left): Exocentric view. Users can explore the space independently with only the three-dimensional reconstruction visible. Users see each other as avatars with a face camera and gaze indicator. (Centre): As the users approach each other, the video captured by the 360° camera slowly becomes visible behind the 3D data. (Right): Egocentric view. When users are co-located, the 360° video capture becomes fully opaque to provide a higher-resolution representation of unmapped areas. Their partner's face is now displayed in the top-right corner.

to the remote user unprocessed; once received, images are passed through a face detection algorithm, the result of which is used alongside each user's tracking data to spatially render an avatar of them within the environment.

In the following sections we detail the algorithms used to perform these various processes as well as key optimisations required for this processing to occur in real time on the mobile phone's modest hardware. An overview of the system architecture and how these various processes interact can be seen in Figure 4.

4.1 Depth Acquisition

The Tango API uses the mobile phone's inbuilt ToF sensor to track the user's position and construct a point cloud of their surroundings. However, this cloud lacks colour and its position is not defined in world space so additional work is required to create a recognisable reconstruction that can expand over time.

At the beginning of each frame we request the latest point cloud from the Tango API, which the Lenovo is only capable of producing five times per second. This cloud is then transformed into world space by multiplying each point's coordinate vector with the local user's current pose matrix within an OpenGL compute shader. This shader also projects each point into the latest image captured by the RGB camera using both cameras' calibration matrices, producing a coloured point cloud in the depth camera's current field of view.

Once captured, the coloured point cloud is then passed to the render thread so that it can be stored in the GPU. New clouds are simply appended to the end of a vertex buffer object; it was found that to fulfill our real-time requirement more complex storage options such as an octree were not possible, resulting in many duplicate points being captured, stored and rendered. To combat this, points are also stored in a separate buffer on the CPU which is managed by the Point Cloud Library[6]. Each time the GPU buffer becomes full it is swapped with a back buffer which will receive all new points and continue being rendered. Each point within the CPU buffer is then filtered through a voxel grid in a separate thread, which ensures only one point will exist within a given area and imposes a uniform structure upon the cloud. The resulting cloud is then uploaded to the back buffer, ensuring that any points captured during the filtering process are not lost.

4.2 Two-Dimensional Environment Capture

To display the panoramic images retrieved from the Ricoh we construct a sphere mesh around the local user. This sphere has a two metre radius, so any depth information within this distance will still be visible. Images are captured in a dual-fisheye format, and so to correctly display them on the sphere we first find each vertex's

[6] www.pointclouds.org

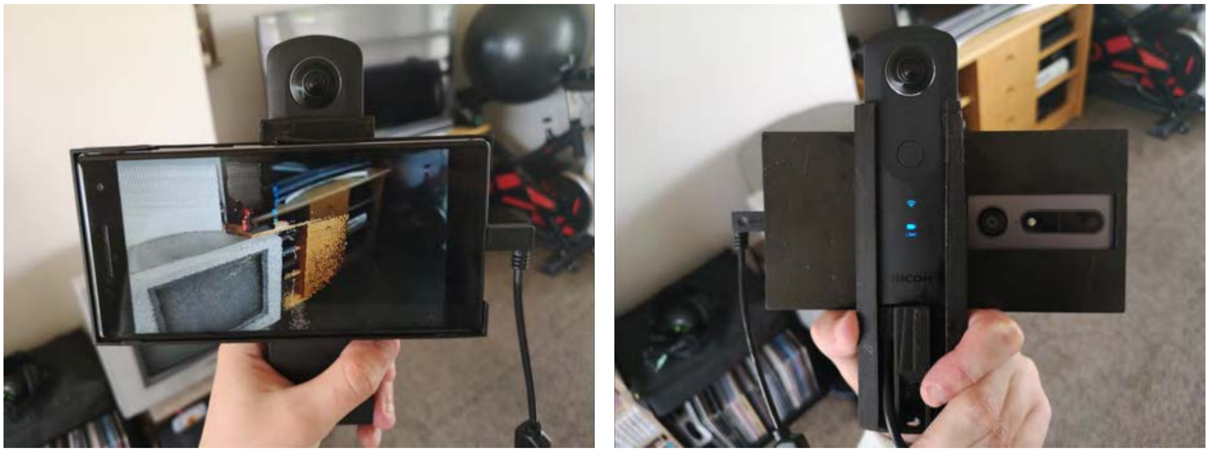

Fig. 3. The hardware used in our research. The Lenovo Phab 2 Pro is used for reconstructing and viewing the environment, while the Ricoh Theta S is used for capturing users' faces and live 360° video for egocentric viewing. The two devices are connected via USB and combined using a custom-built 3D-printed mount.

texture coordinate using the method described by Young *et al.* [31], using a unit vector to each vertex coordinate for the calculation rather than the user's current view direction.

As objects within the point cloud and panorama are unlikely to align within the remote user's view when they are sufficiently distant from the local user, the panorama only becomes visible once the two are within a short distance of one another. To aid in this transition the panorama's opacity is increased exponentially as this distance lessens, losing full transparency at one metre and gaining full opacity at 10cm. Any movements made by the remote user toward the local user's position will also be exaggerated using this same exponential curve, with the two users "snapping" together once their distance falls within a small threshold. This allows for the transition from 3D to 2D to be as seamless and natural for the users as possible so that they can focus on the experience rather than operating the application.

4.3 Facial Capture

In traditional telepresence systems users must to choose whether to show their face or, in the case of remote collaboration, the task space. This is not an easy choice: views of the task space provide greater opportunities for conversational grounding [9], while views of the user's face can provide helpful emotional and conversational cues [8]. Our use of a 360° camera removes this issue as it will always be able to capture a view of each user's face, no matter the angle they are viewing their mobile phone from. This capture is then overlaid on a plane positioned in front of their avatar, but if at any point users transition into the egocentric view it will instead be displayed in the corner of their partner's display so it remains visible while users are co-located.

As users will be rotating their device to look around the environment it is likely their face will move around this capture area. To resolve this, the user's face is tracked using OpenCV's[7] implementation of Haar-cascade detection; this is performed on the unprocessed fisheye image to maintain performance as users are unlikely to be viewing their device from extreme angles and so their face won't see any extreme distortion. Once a bounding box containing a face is found, the latitude and longitude of its centre point within the 360° capture sphere is found and used as the centre of projection, again using the algorithm by Young *et al.* [31].

[7] https://opencv.org/

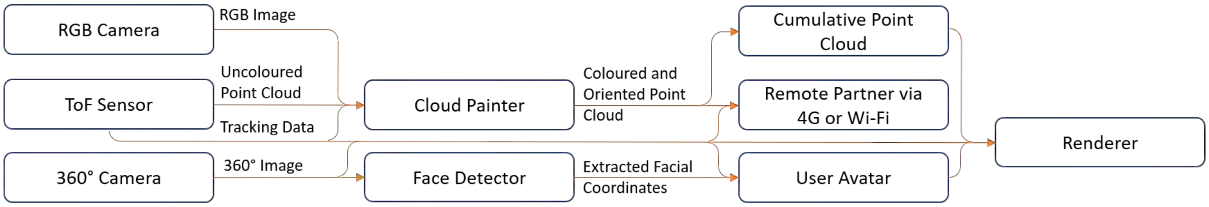

Fig. 4. An overview of the interactions between the various modules and data streams within the system. Data here moves from left to right, so all is captured by either the RGB, ToF or 360° camera and processed in several modules before being rendered or sent to the remote peer.

4.4 Rendering

While some systems opt to create a mesh from unprocessed point data [24], we instead keep the scene in its cloud form and render each point as its own primitive. This allows for the lowest latency possible as no processing is required from the time the point is coloured to when it is rendered, though has the unfortunate side effect of creating holes whenever insufficient data is captured.

This approach also produces many more primitives that need rendered. To mitigate this a random noise texture is generated, with each texel assigned a random number in the range $[0.1, 1]$. This texture is sampled in the vertex shader for each point based on its coordinate, and the resulting number is multiplied by the distance to the far plane to calculate a maximum distance at which that point can be rendered. To prevent the resulting lower cloud density from creating holes in distant objects, each point's size is also increased exponentially based on its distance to the user, effectively lowering the resolution of distant objects in a way not visible to the user.

4.5 Networking

Networking is achieved through Google's open-source implementation of WebRTC[8]. A central server is used for matchmaking, but once a connection between devices is established all further communication is entirely peer-to-peer. Support for STUN and TURN servers is also implemented for NAT traversal which has allowed for Mobileportation to be successfully tested over international connections. Each peer has its own dedicated video and audio channel as well as a data channel for sending each user's positions and 3D point clouds. These clouds are painted before transmission so no synchronisation between these channels is required.

4.6 Performance

One of our goals when developing Mobileportation was to ensure real-time frame rates so that users would have the smooth, seamless experience they are accustomed to from existing videoconferencing solutions. As shown in Figure 5 this was achieved for reasonably sized environments, with the application beginning at 60fps and decreasing linearly as more points are captured, dropping to around 15fps at five million points. This is approximately what was required for the two-storied building shown in Figure 6 which we expect to be uncommon; typical use cases will see an average frame rate of 20-30fps. Despite all processing being performed on a mobile phone, this is better performance than achieved by many similar desktop-based systems [1, 10, 11, 23].

The Lenovo Phab 2 Pro's ToF camera captures at 5fps which enforces a limit on the application's overall capture rate. Each frame has an average of 5,000 points, and once captured this cloud requires 31 milliseconds to transform to world space, colour, and upload to the GPU. Mobileportation thus provides this full 5fps update rate,

[8] https://webrtc.org/

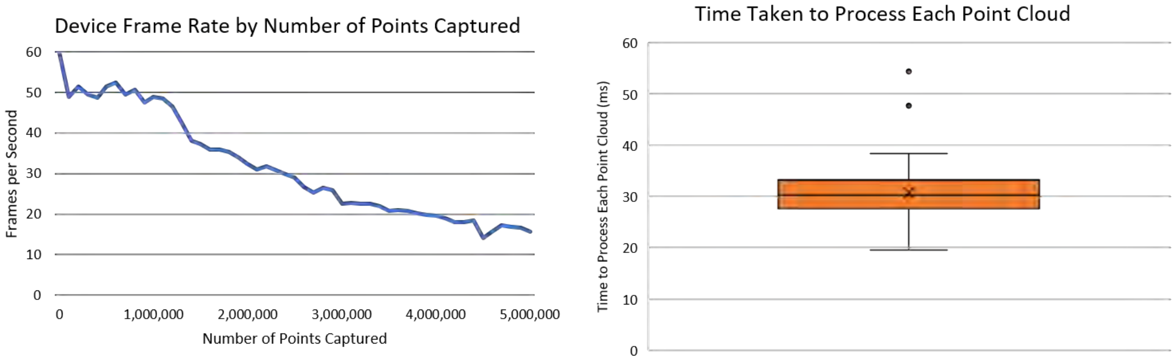

Fig. 5. (Left): The mobile phone's frame rate as a function of the number of points captured. More points result in a linear decrease in frame rate, though it remains interactive for all but the largest captures. (Right): The milliseconds required to process each point cloud after it has been captured. This averages 30ms in almost all cases, suggesting that 30fps capture is possible with a faster camera.

and in theory could capture depth maps at 30fps with a faster camera without any other hardware or algorithmic changes.

5 USER EVALUATION

In developing Mobileportation, we hypothesised that it would be preferred by users to conventional 360° videoconferencing. To determine which factors could contribute to this preference, we identified the following sub-hypotheses:

(1) Despite the reduction in visual fidelity and immediacy of information introduced by incremental 3D capture, allowing transitions between 6DoF exploration of a 3D environment and 3DoF exploration of a 2D one would induce the same heightened sense of spatial presence as conventional 360° videoconferencing.
(2) Showing each user's position and face via a 3D avatar would induce more co-presence than their face alone.
(3) The above features would also contribute to a higher sense of social presence than if they were absent.

To test these hypotheses an experiment was conducted on novice users, who were asked to use the application within an unprepared environment with a live, remotely situated study mediator. Mobileportation was developed with the average consumer in mind and so a large focus was placed on their personal experience; this is contrary to related systems that aim to support professional users in a remote expert scenarios and so the usual industry-focused measures of task performance or cognitive workload were omitted.

Our system was compared to video-only 360° videoconferencing, which to keep all other factors consistent was Mobileportation locked to an egocentric view with depth capture disabled. Users' faces and gaze indicators were still captured and displayed, and participants could freely look around the live 360° video capture, however they could not walk around the space as their position was locked to the mediator's. As comparisons between traditional and 360° videoconferencing are already well documented [21, 28, 31] this gives a good indication to how Mobileportation performs relative to other systems and allows it to be used as a benchmark for future telepresence research.

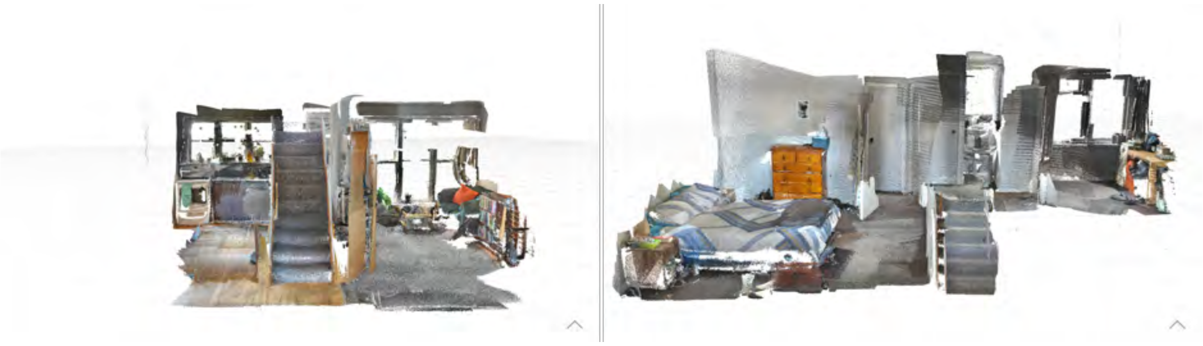

Fig. 6. A two-storied residential property reconstructed in real time using Mobileportation. Each story is constructed here separately for the sake of clarity, but the portability of our system allows for such a large reconstruction to be made in one pass.

5.1 Study Design

14 participants between the ages of 18 and 65 were recruited, of which eight were female; given this limited sample size, the results of this study should be considered as indicative rather than conclusive. Each was gifted a $10 supermarket voucher upon completion of the experiment. We used a within-subjects design where the independent variable was the application used; either Mobileportation or standard 360° videoconferencing. The order of conditions was randomly chosen for each participant to reduce potential learning effects. Each consisted of an informal guided tour through one of two floors of a rental property, which was randomly assigned per condition for each participant to minimise the effects the floor's contents may have on the participant's engagement with the space. A reconstruction of the explored property is shown in Figure 6.

Spatial presence was measured using the IPQ questionnaire by Schubert *et al.* [25], while social and co-presence were measured using questionnaires by Biocca *et al.* [4], Bailenson *et al.* [3], and Hauber *et al.* [12]. These consisted of statements about the user's experience while using the system, with participants noting the degree to which they agree with each on a 7-point Likert scale. A blank space was provided at the end of the questionnaire for participants to write free-form comments about their experience.

A post-experiment questionnaire was given to participants after both conditions had been completed with the following questions:

(1) "Which of the two systems did you find easiest to use?"
(2) "Which of the two systems made you feel more 'present' in the virtual environment?"
(3) "Which of the two systems made it feel more like the remote partner was present with you?"
(4) "Which of the two systems did you prefer overall?"

Space was provided after each to allow participants to justify their decision.

5.2 Procedure

For each condition, participants were connected to a study mediator who was physically located within the rental property several kilometres away and connected via either WiFi or 4G. The participant was given a brief overview of how to operate the application for that condition and allowed five minutes to familiarise themselves with its use with their camera and microphone disabled to allow them to experiment without fear of being observed. One room within the rental property was set aside for this purpose.

Once comfortable with the system, a brief and informal tour of the assigned floor was given. The mediator would pretend the participant was interested in renting the property and show them through various rooms while describing their contents, though participants were encouraged to explore on their own and could (and often did) ignore the mediator entirely. All video, audio, and tracking data was captured and transmitted between the two parties in real time so that the mediator could react to participant comments and requests. Participants were encouraged to ask the mediator to revisit areas they wished to see more of in both conditions. No set task was to be completed other than experience the system in a realistic social scenario. Once the relevant floor had been extensively shown, which took approximately five to ten minutes, participants were asked to complete the presence questionnaires. This was repeated for the remaining condition on the other floor, after which participants completed the post-experiment preference questionnaire.

5.3 Results

According to the post-experiment questionnaire, twelve of the fourteen participants (85%) found Mobileportation more difficult to operate than conventional 360° videoconferencing. This did not dissuade nine of them (64%) claiming it made them feel more present in the remote environment, ten (71%) claiming it made it feel more like their remote partner was with them, and nine (64%) choosing it as their preferred system overall.

The presence questionnaires indicated that the amount of spatial presence induced within the environment was rated similarly for both Mobileportation ($\mu = 4.30$, $\sigma = 0.86$) and 360° videoconferencing ($\mu = 4.27$, $\sigma = 0.79$), with a Wilcox signed rank test ($N = 14$, $\alpha = 0.05$) showing no significant difference between the two conditions ($p = 0.59$). The amount of co-presence induced between communication partners was rated slightly higher in Mobileportation ($\mu = 4.93$, $\sigma = 0.81$) than in 360° videoconferencing ($\mu = 4.20$, $\sigma = 1.10$), though another Wilcox signed rank test revealed this difference to be insignificant ($p = 0.12$). A significant difference was however found in the social presence induced between users ($p = 0.03$), with Mobileportation ($\mu = 5.99$, $\sigma = 0.90$) rated significantly higher than 360° videoconferencing ($\mu = 5.13$, $\sigma = 1.57$).

6 DISCUSSION AND FUTURE DESIGN SPACE

Our overall hypothesis was that participants would prefer Mobileportation over conventional 360° videoconferencing. Nine participants (64%) indicated that this was the case; all that chose otherwise said visual quality was the main deciding factor, suggesting future hardware improvements could improve this ratio. Those that preferred Mobileportation often cited how much fun they had with the system and the sense of presence it invoked within the shared environment; while promising, this could be due to the novelty of the experience and so further research is required to determine if this remains the case after extended use.

The first sub-hypothesis we proposed that could contribute to this preference was that Mobileportation would provide a similar level of spatial presence within the shared environment as conventional 360° videoconferencing, which was confirmed by our experimental results. While it may seem counter-intuitive that a combination of exo- and ego-centric views would provide similar spatial presence as purely egocentric ones, this makes sense in the context of the visual fidelity of the captured environments. With 360° cameras the entire environment is captured and available as soon as the application starts in a high resolution; this is not the case when incrementally constructing the 3D space, and our point cloud representation of it is lower resolution than what an image can provide. All participants that chose 360° videoconferencing as providing a higher sense of spatial presence specified visual fidelity as the main reason for this, with some becoming confused at seeing the room as "fragmented" and "not fully rendered", and another stating that "3D would have been awesome if I could perceive my surroundings without the need to wait". We also speculate that visual inconsistency between the 2D and 3D data could inhibit spatial presence, however this requires more research.

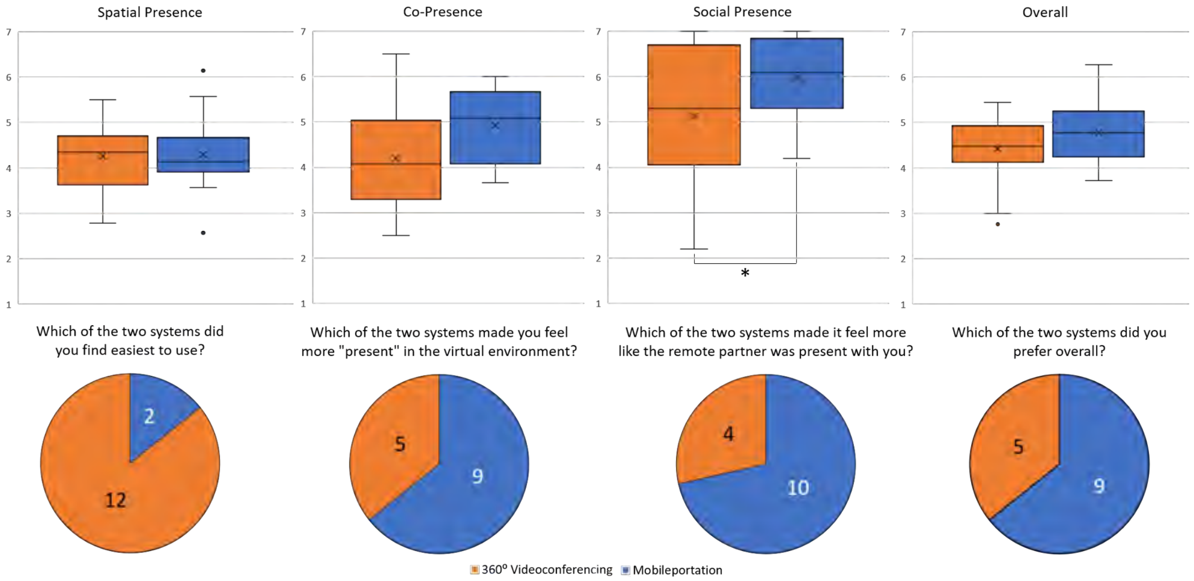

Fig. 7. The results of our preliminary user study comparing Mobileportation to video-only 360° videoconferencing. The top row shows aggregate results of various presence questionnaires for each condition, while the bottom row shows the number of participants that preferred each system over several categories as indicated by a post-experiment questionnaire.

Despite the decreased visual fidelity of the 3D data, a comparatively immersive experience was still achieved. A likely reason for this is the ability it provides users to freely explore the shared environment, which led to an "overall more enjoyable", "more seamless", and "more fun and immersive" experience. Participants that experienced Mobileportation as their first condition often missed this during their second, with one stating that "without being able to physically move around I felt it's more restrictive. I prefer [Mobileportation] where the room is rendered out despite video being clearer", and another saying of 360° videoconferencing: "The thing that I missed in this experience is that I do not have the freedom to move and interact with the virtual world, it felt more like a virtual tour". With this freedom of exploration making up for the lack in visual quality, we believe that hardware improvements could lead to Mobileportation becoming a significantly more presence-inducing experience than what is currently available on mobile phones with no algorithmic changes required.

Our second sub-hypothesis was that spatially rendering representations of each user within the shared environment would increase the co-presence felt between them. Unfortunately the presence questionnaires indicate that this is not the case, even though 71% of participants claimed otherwise in the post-experiment questionnaire. The most likely reason for this is that 360° videoconferencing has constant views of the mediator's face, while Mobileportation only shows this when the participant actively seeks them out. This led to participants having trouble locating the tour guide at times, with one having "difficulty as sometimes I was not able to catch him [the mediator] or not able to spot him" and another finding it "hard when [their partner] moved without me knowing, finding him again was confusing at times" and that it was "easier to track where the [gaze indicator] is". This is a unique problem only made possible with the arbitrarily-sized environments afforded by Mobileportation; in existing systems with limited tracking areas finding a peer would be trivial as they must remain close, but in our experiment it was common for participants to completely abandon the mediator in favour of self-guided exploration and get lost which could have limited the sense of co-presence they felt.

Participants still felt that the spatially rendered avatars were helpful in providing a sense of co-presence with the mediator, stating that it was "like the remote partner was in the same room compared to just seeing the face all time" and that it "allowed for a feeling of being there with the person rather than being on call with them". This could speak for a hybrid approach where the user's face is displayed in the corner whenever their avatar is not visible, or some other indicator as in other systems [16, 31] could show their relative positioning at all times to ensure users have constant representation of their partner and thus do not lose each other. Spatial audio could additionally be added such as by Langlotz et al. [17], allowing auditory cues to give a sense of the remote partner's location within the space and guide users together when they become visually separated.

Our third sub-hypothesis was that Mobileportation would provide more social presence between its users than 360° videoconferencing, which was confirmed by our experimental results. Participants complained that 360° videoconferencing felt too much like a "360 visual tool" or a "virtual tour", or that it was "a bit like a 'presentation"' or "like Google street view"; these complaints seem unfounded as they are accurate descriptions of the task, but in the context of our experiment imply that they saw Mobileportation as a completely different experience than what they're accustomed to due to being able to explore and interact with their partner. One felt that without this exploration aspect "it was easy to ignore and just listen, which would be the same as a pre-recorded video", and another that it was more like "being there while they show you something" than actively participating.

6.1 Future Research

This experiment focused heavily on socialisation and presence, however many other aspects remain to be explored. For instance, while we have evaluated the sense of spatial presence Mobileportation induces within shared environments, it is not known how it affects a user's understanding of that space. In initial pilot tests of the system we asked participants to sketch the layout of the rental property after it had been explored and rate the mental workload required to do so, though this was soon abandoned as it was found to distract too much from the social aspect of the experiment as participants focused on memorising the space rather than interacting with the mediator. A future experiment could reintroduce this task, allowing us to determine whether free exploration of an incrementally constructed space could lead to a better understanding of it.

The mechanism used to transition between ego- and exo-centric views also has yet to be fully evaluated. In our experiment participants only used it sparingly to view areas yet to be constructed, though never to view the space with higher fidelity as we initially expected. Consequently, participants spent almost all their time in an exocentric view, sometimes briefly transitioning to egocentric mode whenever they entered a new room but would favour exocentric viewing of it once it had been reconstructed. Despite the long training period, some even forgot that this transition was possible and were subsequently surprised after accidentally triggering it during the experiment. Future studies could focus on how to make view transitioning more useful and appealing to participants, record how long they spend in each mode, and discover scenarios in which they would use it.

Our focus on social scenarios also meant that traditional industry-focused collaborative tasks were ignored. Despite this, it is easy to see what benefits such a system could have for remote collaboration; free view exploration reduces the time taken to complete collaborative tasks [23], and our mobile implementation means such collaboration could occur wherever and whenever is convenient to the collaborators. It could be that independent ego- and exo-centric views are beneficial for different types of tasks, or that some combination of the two is required to achieve the greatest effect; this could be determined in future experiments focused on more collaborative scenarios.

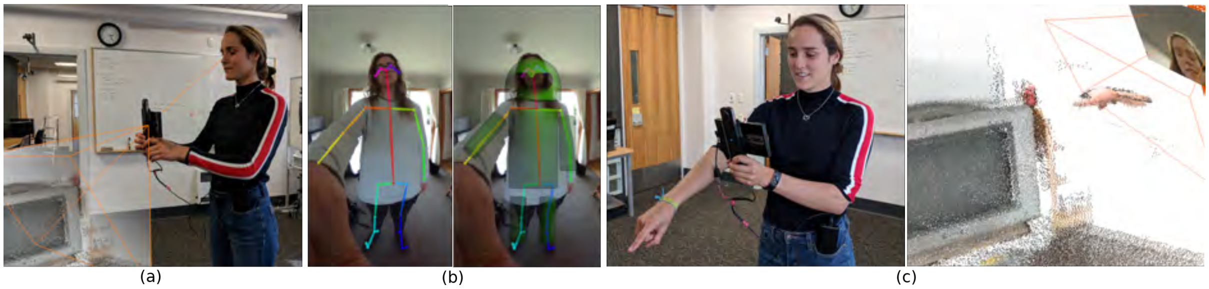

Fig. 8. Prototypical features that are also supported through Mobileportation's hardware configuration. (a): The user's view direction and gaze indicator are determine by the angle between their face and the display, providing a "smart window" into the virtual world. (b): The position of the user's body is detected and used to manipulate the limbs of their 3D avatar. (c): The user's hands are detected and rendered in the environment to allow for gestural communication.

6.2 Future Design Space

Although the Tango API is officially deprecated, we have recently seen a resurgence of RGBD sensors being included in standard devices such as the Huawei P30 Pro[9], and even phones without this sensor are capable of stereoscopic depth capture due to the increasing trend of including multiple lenses on the back of the device. Some manufacturers have also started to experiment with how panoramic cameras can be incorporated, with Essential[10] and Motorola[11] supporting modular 360° cameras and the Samsung Galaxy S10[12] integrating a 123° FoV lens in the phone itself. In our research we use a custom setup combining different off-the-shelf hardware, however with these developments it is likely that all of the features Mobileportation provides could soon be achieved using only a standard mobile phone, and we hope that our research provides incentive for manufacturers to include the necessary hardware. Future improvements in mobile processing power could also allow for more advanced features not currently feasible for real-time use; we explore these here, with prototypical demonstrations of each shown in Figure 8.

One common feature of telepresence applications is visualisations of the user's current gaze direction [16, 24]. This is currently approximated by visualising the camera's field of view, however given that the position of the user's face is already tracked, future iterations could provide a more accurate indicator by computing the angle between the user and the display. This would also allow for view-dependent rendering where the display becomes a "window" into the shared space. Future iterations utilising higher resolution cameras could even extract eye data from the panoramic image which could be used for even more precise gaze estimates.

Another possibility is to use the 360° camera to capture the user's entire body rather than just their face. This has previously been attempted using fisheye cameras integrated into a head-worn cap [30], though it would be preferable to use a device already owned by the user. To show how this could be achieved we extracted the user's body from the Ricoh footage and tracked their skeleton using OpenPose [6, 7]. This skeleton model could then be used to control the movements of the user's avatar as shown in Figure 8, or the body capture could be shown directly within the shared space, though both proved too computationally expensive for real-time use.

This hardware could also allow gestures to be used in conversation. As clouds are captured they could be searched for hands, which could be spatially rendered in the environment with the correct size, proportions and colour. Alternatively, an artificial mesh such as used by Sodhi *et al.* [26] could be shown to give a complete

[9] https://consumer.huawei.com/en/phones/p30-pro/
[10] https://www.essential.com/
[11] https://www.motorola.com/us/products/moto-z-gen-4-unlocked
[12] https://www.samsung.com/us/mobile/galaxy-s10/

model of the hand at the expense of realism. To explore how these gestures could be integrated we applied a color threshold to all captured points, marking those that are skin-coloured and within a set distance to the camera as belonging to a hand. These are not integrated into the cumulative mesh and are instead only displayed temporarily and can be even be captured by the remote user who is not actively scanning their environment.

7 CONCLUSION

In this paper we present Mobileportation, an immersive nomadic telepresence system based on a novel combination of panoramic and depth cameras in a handheld mobile form factor. Mobileportation utilises the strengths of these sensors to allows distant users to explore shared live environments using a combination of egocentric views of 2D panoramas and exocentric views of incrementally constructed 3D data. Users may freely explore this space with their position shown via a 3D avatar with live capture of their face overlaid, allowing conventional face-to-face communication as well as independent views of the environment.

Mobileportation is preferred by users to 360° videoconferencing as it provides a more social, immersive, and fun experience, inducing similar levels of spatial and co-presence while providing more social presence between communication partners. This is despite the lower visual fidelity unavoidable in current mobile systems, which future hardware improvements could remedy with no changes required to the application.

While it currently requires use of an external 360° camera, the recent prevalence of integrated depth cameras and wide-angle lenses could allow for the entire system to be available on a single device. Such revisions would also allow for new features to be integrated such as gaze-based rendering, full-body rendering, and gesture tracking, allowing for a fully immersive experience that so far has been restricted to high-power desktop systems.

ACKNOWLEDGMENTS

The authors would like to thank Jonathan Sutton and Rosa Lutz for their help in conducting the user study, and the rest of the HCI lab for their valuable input throughout the duration of this research.

REFERENCES

[1] Matt Adcock, Stuart Anderson, and Bruce Thomas. 2013. RemoteFusion: Real Time Depth Camera Fusion for Remote Collaboration on Physical Tasks. *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry - VRCAI '13* (2013), 235–242. https://doi.org/10.1145/2534329.2534331

[2] Ilkka Arminen and Alexandra Weilenmann. 2009. Mobile Presence and Intimacy - Reshaping Social Actions in Mobile Contextual Configuration. *Journal of Pragmatics* 41, 10 (oct 2009), 1905–1923. https://doi.org/10.1016/j.pragma.2008.09.016

[3] Jeremy N. Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. 2005. The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments. *Presence: Teleoperators and Virtual Environments* 14, 4 (2005), 379–393. https://doi.org/10.1162/105474605774785235 arXiv:https://doi.org/10.1162/105474605774785235

[4] Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 456–480. https://doi.org/10.1162/105474603322761270

[5] Celeste Campos-Castillo and Steven Hitlin. 2013. Copresence: Revisiting a Building Block for Social Interaction Theories. *Sociological Theory* 31, 2 (2013), 168–192. https://doi.org/10.1177/0735275113489811

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[8] Nick V. Flor. 1998. Side-by-side collaboration: a case study. *International Journal of Human-Computer Studies* 49, 3 (1998), 201 – 222. https://doi.org/10.1006/ijhc.1998.0203

[9] 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*. 21–30. https://doi.org/10.1145/358916.358947

[10] Lei Gao, Huidong Bai, Rob Lindeman, and Mark Billinghurst. 2017. Static Local Environment Capturing and Sharing for MR Remote Collaboration. *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications on - SA '17* (2017), 1–6. https://doi.org/10.1145/3132787.

3139204

[11] Lei Gao, Huidong Bai, Thammathip Piumsomboon, Gun A Lee, Robert W Lindeman, and Mark Billinghurst. 2017. Real-time Visual Representations for Mixed Reality Remote Collaboration. (2017), 87–95. https://doi.org/10.2312/egve.20171344

[12] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in Videoconferencing: Trade-offs Between Efficiency and Social Presence. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. ACM, New York, NY, USA, 413–422. https://doi.org/10.1145/1180875.1180937

[13] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and et al. 2011. KinectFusion: Real-Time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. Association for Computing Machinery, New York, NY, USA, 559–568. https://doi.org/10.1145/2047196.2047270

[14] Shunichi Kasahara and Jun Rekimoto. 2014. JackIn: Integrating First-Person View with Out-of-Body Vision Generation for Human-Human Augmentation. In *Proceedings of the 5th Augmented Human International Conference*. ACM, Kobe, Japan, 46:1–46:8. https://doi.org/10.1145/2582051.2582097

[15] Ryohei Komiyama, Takashi Miyaki, and Jun Rekimoto. 2017. JackIn Space: Designing a Seamless Transition Between First and Third Person View for Effective Telepresence Collaborations. In *Proceedings of the 8th Augmented Human International Conference (AH '17)*. ACM, New York, NY, USA, Article 14, 9 pages. https://doi.org/10.1145/3041164.3041183

[16] Sven Kratz, Daniel Avrahami, Don Kimber, Jim Vaughan, Patrick Proppe, and Don Severns. 2015. Polly Wanna Show You: Examining Viewpoint-Conveyance Techniques for a Shoulder-Worn Telepresence System. In *MobileHCI 2015 - Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Toronto, Canada, 567–575. https://doi.org/10.1145/2786567.2787134

[17] Tobias Langlotz, Holger Regenbrecht, Stefanie Zollmann, and Dieter Schmalstieg. 2013. Audio Stickies: Visually-guided Spatial Audio Annotations on a Mobile Augmented Reality Platform. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration (OzCHI '13)*. ACM, New York, NY, USA, 545–554. https://doi.org/10.1145/2541016.2541022

[18] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication* 3, 2 (09 1997). https://doi.org/10.1111/j.1083-6101.1997.tb00072.x JCMC321.

[19] Andrew Maimone and Henry Fuchs. 2011. Encumbrance-Free Telepresence System with Real-Time 3D Capture and Display using Commodity Depth Cameras. *2011 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011* (2011), 137–146. https://doi.org/10.1109/ISMAR.2011.6092379

[20] Peter Mohr, Shohei Mori, Tobias Langlotz, Bruce Thomas, Dieter Schmalstieg, and Denis Kalkofen. 2020. Mixed Reality Light Fields for Interactive Remote Assistance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA.

[21] Jörg Müller, Tobias Langlotz, and Holger Regenbrecht. 2016. PanoVC: Pervasive Telepresence using Mobile Phones. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–10. https://doi.org/10.1109/PERCOM.2016.7456508

[22] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[23] Fabrizio Pece, William Steptoe, Fabian Wanner, Simon Julier, Tim Weyrich, Jan Kautz, and Anthony Steed. 2013. Panoinserts: Mobile Spatial Teleconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems SE - CHI '13*. 1319–1328. https://doi.org/10.1145/2470654.2466173

[24] Thammathip Piumsomboon, Arindam Day, Barrett Ens, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. Exploring Enhancements for Remote Mixed Reality Collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications (SA '17)*. ACM, New York, NY, USA, Article 16, 5 pages. https://doi.org/10.1145/3132787.3139200

[25] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments* 10, 3 (2001), 266–281. https://doi.org/10.1162/105474601300343603

[26] Rajinder S Sodhi, Brett R Jones, David Forsyth, Brian P Bailey, and Giuliano Maciocci. 2013. BeThere: 3D Mobile Collaboration with Spatial Input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (2013), 179–188. https://doi.org/10.1145/2470654.2470679

[27] Patrick Stotko, Stefan Krumpen, Matthias B. Hullin, Michael Weinmann, and Reinhard Klein. 2019. SLAMCast: Large-Scale, Real-Time 3D Reconstruction and Streaming for Immersive Multi-Client Live Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (May 2019), 2102–2112. https://doi.org/10.1109/TVCG.2019.2899231

[28] Anthony Tang, Omid Fakourfar, Carman Neustaedter, and Scott Bateman. 2017. Collaboration in 360° Videochat: Challenges and Opportunities. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, Edinburgh, United Kingdom, 1327–1339.

https://doi.org/10.1145/3064663.3064707

[29] Theophilus Teo, Louise Lawrence, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 201, 14 pages. https://doi.org/10.1145/3290605.3300431

[30] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, P. Fua, H. Seidel, and C. Theobalt. 2019. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (May 2019), 2093–2101. https://doi.org/10.1109/TVCG.2019.2898650

[31] Jacob Young, Tobias Langlotz, Matthew Cook, Steven Mills, and Holger Regenbrecht. 2019. Immersive Telepresence and Remote Collaboration using Mobile and Wearable Devices. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (May 2019), 1908–1918. https://doi.org/10.1109/TVCG.2019.2898737