Tourgether360: Collaborative Exploration of 360° Tour Videos using Pseudo-Spatial Navigation

This is Short Title of the paper, used in page headers

First Author's Name, Initials, and Last name*

First author's affiliation, an Institution with a very long name, xxxx@gmail.com

Second Author's Name, Initials, and Last Name

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Third Author's Name, Initials, and Last Name

Third author's affiliation, possibly the same institution, xxxx@gmail.com

Figure 1: Tourgether360 allows collaborators, represented by avatars, to tour together "inside" a 360-degree video.

Collaborative exploration of 360° videos with contemporary interfaces is challenging because collaborators do not have awareness of one another's viewing activities. Tourgether360 addresses this problem for 360° tour videos using a pseudo-spatial navigation technique that provides both an overhead "context" view of the environment as a minimap, as well as a shared pseudo-3D environment for exploring the video. Collaborators appear as avatars along a track depending on their position in the video timeline and can point and synchronize their playback. We evaluated Tourgether360 with a collaborative experience study and found that participants adopted the navigation approach with ease and enjoyed the shared social aspects of the experience. Participants reported finding the experience similar to an interactive social video game.

* Place the footnote text for the author (if applicable) here.

CCS CONCEPTS • Human-centered computing • Collaborative and social computing • Collaborative and social computing systems and tools

Additional Keywords and Phrases: Insert comma delimited author-supplied keyword list, Keyword number 2, Keyword number 3, Keyword number 4

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

360° tour videos are an increasingly popular way of exploring remote destinations and environments. Such videos are typically shot using an omnidirectional camera mounted atop a tripod as the cameraperson moves through an environment (e.g., by walking or driving). The videos provide viewers with the ability to freely look around, independent of the direction that the cameraperson was moving. Because of this freedom, they provide users with a rich sense of immersion—particularly when coupled with head mounted displays (e.g. [6,31]). Such videos are also increasing in popularity: beyond simply being used to view and compare vacation destinations, they are now increasingly being used by families to visit prospective college campus, or by realtors showing off homes for rent or sale. Thus, *collaborative* viewing is also becoming increasingly important. Collocated or remote friends may want to watch 360° videos to experience immersive and social entertainment together, or such videos may be used in the educational context, with a class of students going on a virtual museum tour or virtual trip to cultural locations.

The problem is that current 360° interfaces do not provide effective support for collaborative navigation and exploration of 360° videos (e.g. [30]). With only a handful of exceptions (e.g.[20,27]) 360° video players are intended for single-person use; in addition to normal video playback controls, such video players need to provide a special, separate means for controlling the view orientation. On a desktop, orientation is controlled by grabbing the scene and moving it; on tablets, this is augmented through gyroscopic sensors, and on a head-mounted display, orientation is controlled by turning or tilting one's head. Yet, there is little to no support for collaborative viewing of these immersive videos. Prior work has demonstrated that when collaborators are watching 360 videos together, collaborators may not want to be looking at the same thing at the same time [29,30]; in spite of this, they still want to maintain an awareness of what their collaborators are watching. To this point, we have yet to see collaborative 360 video viewing experiences that support these needs.

We propose a pseudo-spatial navigation metaphor for collaborative exploration of 360 videos inspired by a focus+context approach [28]. We realize this approach in a prototype system called Tourgether360, which allows several collaborators to explore a 360 video together. With Tourgether360, the video tour context is visualized as a path on a 2D map of the environment. Building on this approach first illustrated by Noronha and colleagues [21], users can scrub along the path to navigate both time and position in the 360 video (both in focus and in context views). This eases the coordination between finding points in the video with spatial locations visited in the video. When multiple users view the same video, their viewing position (time, space and orientation) is embodied by a viewing capsule, and they can position and mark points of interest for one another. Figure 1 illustrates how two collaborators are embodied in the

shared 360° video scene together. This allows collaborators to experience the video as if they were embodied together in the video. While the mapping for Tourgether360 relies on manual parameter tuning, commercial applications are now able to reconstruct these scenes using 360° photographs[1], and computer vision approaches have shown the ability to reconstruct rich environments with only a limited amount of video data [35].

We evaluated our navigation approach by conducting a collaborative user experience study involving 16 participants. Participants collaboratively explored 360° videos using the Tourgether360 interface, completing tasks that required navigating through the video and identifying points of interest. We found that participants had very little difficulty adopting and using the pseudo-spatial navigation technique, and used this to navigate the videos as opposed to using the timeline scrubber. Furthermore, we observed that participants were able to effectively use the cues to communicate and coordinate their interaction with the video.

We make three contributions in this work: first, we contribute extensions to a navigation technique for 360 tour videos that employs a pseudo-spatial approach; second, we contribute a system that realizes this technique and facilitates multi-user interaction; finally, based on our study, we uncover a number of new interaction challenges and opportunities based on pseudo-spatial navigation approaches that future designers ought to consider.

2 RELATED WORK

To set the stage for our work, we outline three related areas of research: first, we describe recent work that has explored new metaphors for navigating 360° videos, which consider spatial navigation approaches; second, we describe efforts to support collaborators exploring and using pre-recorded video streams, and then finally we outline some foundational work on collaborative virtual environments that inspired our approach.

2.1 Interaction with 360° Videos

Research exploring interaction with 360° videos has either focused on supporting orientation navigation (directing one's view in the video), or temporal navigation (controlling playback or directing one to interesting moments in the video). In terms of orientation navigation, considerable prior work has explored how to ensure the viewer does not miss important points of interest. Several approaches automate this through computational measures (e.g. [14]), while other researchers have designed mechanisms to signal to viewers where the view should be oriented. For creating pre-defined 360° video stories, Pavel et al. [22] propose two techniques to orient a viewer's perspective when the playback comes to a pre-defined point of interest. On the other hand, Lin et al. [15] propose visualizations for points of interest that are out of the FOV of the viewer, allowing the viewer to see the point of interest in an inset video. Mäkelä et al. [16] show that such indicators of others' viewing interest (social indicators) can improve the experience, even if they are subtly distracting.

Several researchers have also proposed new techniques for temporal navigation of videos. For instance, Petry & Huber [23] explore multimodal gestures for playback controls for 360° videos within a head-mounted display viewing context. Similarly, Ruiz et al. [26] apply this approach within a multi-person viewing context. Neng and Champbell [18] present a 360° video player that augments the traditional timeline with cues representing points of interest, and regular thumbnails for some frames in the 360 video. For scrubbing through videos, VRmiere provides a "Little Planet" navigation technique for these videos, which can provide some spatial awareness [19].

[1] http://www.reconstructinc.com

Like our work, Route Tapestries is a departure from this prior work, focusing on blurring the boundary between navigating time and space. In their work, they produce a navigation timeline that presents a slit-scan visualization immediately to the left and right of the forward-facing vector of the video. This timeline becomes a spatial "tapestry" of the scene, and presents the user with identifiable landmarks for navigating the video. This sort of "context" view is similar to the approach in Site Surfers [21], which provides an overhead 2D map with trajectories that can be used for navigation. Our work builds on these approaches to enable a similar kind of spatial navigation that simultaneously manages the temporal navigation, motivated from the perspective of collaboration.

2.2 Collaboration Over Pre-Recorded Video

Studies of people navigating 360° videos together have revealed several communication problems that can be resolved through technology [16,25,30]. Particularly for experiences where viewers can watch simultaneously, there is a strong need for systems to provide an awareness of where others are viewing, and potentially gesture support to support communication and coordination. When collaborators can be at different points in the video at the same time, there is also a need to support this sort of temporal awareness. CollaVR provides this awareness, enhancing the timeline view with extra scrubbers, as well as through a colour-coded rectangle in the viewport that represents the collaborator's perspective [20]. Systems that support playback of VR recordings also provide this type of support, where the scene can be played back and viewed from different perspectives [32].

We were also inspired by prior work that has studied how people communicate about video recordings. Yarmand et al. present a study of YouTube comments, noting that while some comments make use of time-based references to the video, the majority of comments make reference to intervals in the videos—with specific reference to visual entities [36]. Aligned with this, Dodson et al. present a study of how lecture videos are explored within the context of recorded lectures, noting that content-based features (i.e., transcripts) were more useful for gross navigation compared to timeline navigation [3]. Thus, while time is an easy computation index into videos, people rely more on content-based features to communicate and navigate through videos.

Our approach centers on the insight that navigation through video may be better supported through a semantic, contextual understanding of the content (i.e., what is in the video) rather than time. In the context of 360° videos, we explored a visual map-based approach that provides this context view, combined with the normal view of the scene, which is the focus view.

2.3 Collaboration in Virtual Environments

Many of the problems outlined by researchers studying collaborative viewing of 360° videos [25,29,30] are reminiscent of early CSCW research focused on Collaborative Virtual Environments (e.g. [1,2,7,9]). These virtual environments were designed to support multiple collaborators, and designers were forced to contend with basic awareness issues: (1) Where are my collaborators? (2) What can my collaborators see? (3) What are they looking at? (4) How can I draw someone's attention to what I am talking about? Early work by Benford identified many of these issues [1,2]. Follow-up work explored how virtual embodiments for collaborators could provide insight into presence and view orientation (e.g. [1]). Subsequent work by Fraser et al. [7] identified ways in which viewports, and gestures could be supported, where some of these could exaggerate or create representations that might not mimic real life.

And, while some work has explored how to provide additional cues for awareness in 3D workspaces [28,29], most embodiments we see in CVEs or video games have not pushed the boundaries of the embodiments first envisioned in the mid-1990s. Recent work has pointed to how some of these challenges with deictic references (and dereferencing) still

persisting to this day [33,34], with very few solutions addressing many of the challenges identified decades ago. Recent work has showed how these avatars and their presentation can be made to support effective mixed reality collaboration if some aspects of reality are ignored—e.g. size of collaborator [24].

We take inspiration from this prior work in laying down the user experience goals of our approach. While the domain of interest is slightly different (i.e., 360° tour videos vs CVEs), because we realize the 360° tour video in a CVE-like environment, many of the same embodiment approaches may still be applicable.

3 USER EXPERIENCE DESIGN

Our goal in designing Tourgether360 was to create an effective way to collaboratively watch 360° tour videos with others. These videos are characterized by a non-fixed position camera progressing along a path, and frequently focus on tours of tourist areas or cultural heritage locations. Building on prior work, we identified four major design goals:

- *DG1: Support semantic navigation of the video space.* Prior work on video navigation has demonstrated that temporal navigation can be meaningfully augmented with semantic navigation (i.e. with some understanding of the video content itself, e.g. [36]).
- *DG2: Support awareness of collaborators' perspectives and temporal position.* Watching 360° videos with others on independent displays means that collaborators may engage with the immersive experience independently, meaning that collaborators will separate—both in terms of their viewing perspective, as well as in their temporal navigation of the video [29]. Knowing what others are looking at, and where they are is an important part of feeling co-present [10,11].
- *DG3: Enable smooth engagement and disengagement with collaborators' perspectives.* We know collaborators sometimes like to explore the video independently, in a loosely-coupled mode of interaction. At the same time, having a shared perspective supports smooth, detailed coordination and conversation in a tightly-coupled mode of interaction. A tool should allow collaborators to smoothly move between these modes of interaction [29,30].
- *DG4: Support deictic reference with a semantic understanding of the environment.* Finally, the tool should allow collaborators to point and refer to things in the video and environment; further, these should be somewhat persistent given that collaborators may not be looking at the same thing all the time [33].

We executed on our approach by reconceptualizing 360° video as a shared virtual 3-dimensional space and implementing a number of unique features that increase users spatial understanding of the environment and enable allocentric navigation. These features include pseudo-spatial navigation mechanisms, user embodiment and pseudo-spatial annotation, and allocentric coordination affordances. We realized our design vision in the experimental video player called Tourgether360. The interface of the software is presented in Figure 2.

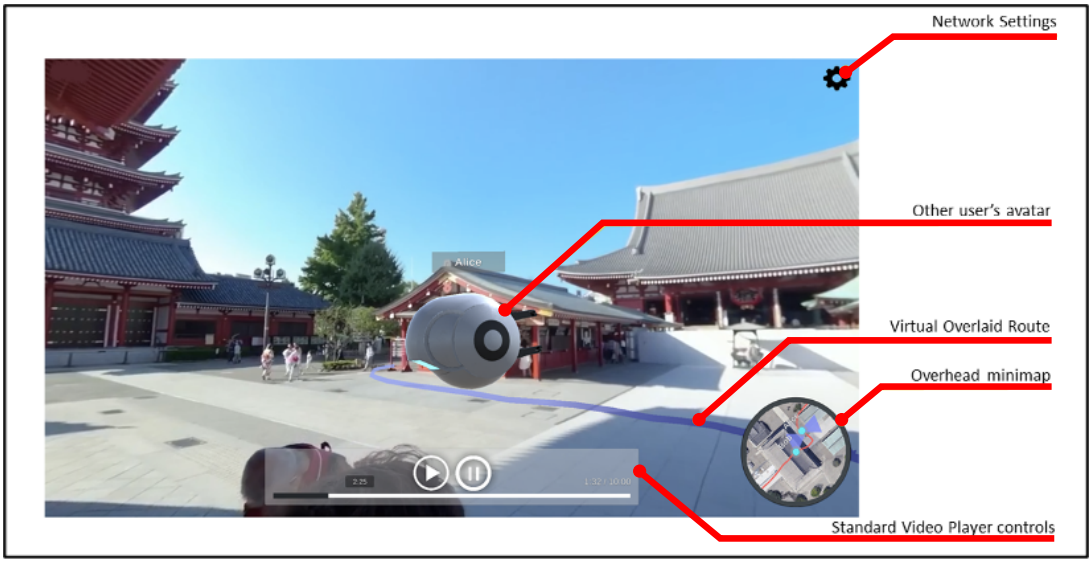

Figure 2: Full Interface of Tourgether360 from the perspective of a user (Bob).

In the following paragraphs we outline how each major function of Tourgether360 addresses the design goals above. **Pseudo-spatial navigation.** Inspired by prior work [21] and the UI of the 3D video games, navigation affordances in our application are supported by an overhead schematic interactive minimap of the 360° video tour environment Figure 3). The minimap provides a virtual path that represents the route on which the tour takes place, where the user's position in the video is represented by a blue dot in space. Depending on where users are currently positioned, their avatars are mapped on the corresponding place on top of the virtual path. The users' gaze direction is also represented on the minimap by the conical light beam from the avatars. The minimap allows users to navigate through the video using spatial landmarks visible from the minimap (DG1).

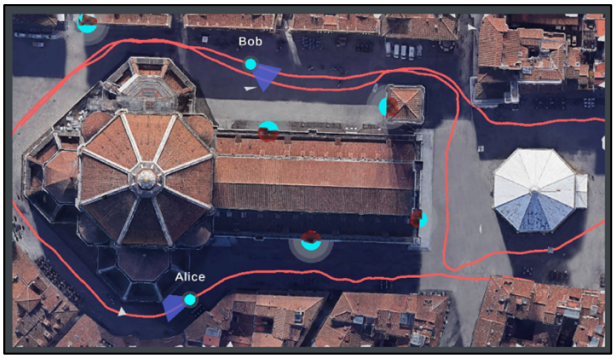

Figure 3: Minimap of the Florence Cathedral environment shown in one of the 360 videos we used in the study. Here, the path taken in the 360 video is represented by the red line. Alice and Bob are at different parts of the video, where their viewing orientations are represented by a cone. Finally, spheres represent marked "points of interest" that were placed by the collaborators.

The minimap serves the users as a navigation control mechanism. By clicking and dragging the mouse along the virtual path on the minimap, the users effectively scrub the video timeline back and forth, rewinding and forwarding the video. In addition, the users can scrub the video by scrolling the mouse wheel.

As illustrated in Figure 4, the metaphor of the 3D space is also realized in the main video view. Here, the path of the video is represented by a continuous line stuck to the ground, showing the forward and backwards route of the tour from the first-person perspective of the user. This virtual route served as a visual representation of the actual route in the environment through which the person with the camera moved while the video was originally recorded. Thus, virtual track provided users with an understanding of what physical places the users pass when playing the video. To enhance the illusion of being present in an environment with geometry, parts of the path are cropped if they could not be seen around the geometry of the space.

Figure 4: Representation of a virtual route overlayed on top of the video. Taken from Alice's point of view (from Figure 3), the path (highlighted blue) illustrates how the video tour will take them around the cathedral. Because Tourgether360 understands the architecture of the space represented in the video, the line path is cropped at the edge of the cathedral.

Collaborator Embodiment. As illustrated in Figure 5, each collaborator is embodied by an avatar in the 3D video tour scene. The avatar is a flying spherical robot with four antennas indicating the "face" part of the robot. This "face" is synchronized with the user's camera's forward vector to indicate the gaze orientation. This embodiment approach in the 3D scene provides awareness of others' temporal position and view orientation (DG2). The embodied avatar is shown in Figure 5.

Figure 5: Representation of the user avatar overlayed on the top of the played video on the virtual route line

When collaborators are watching the video together, the apparent distance between two avatars in the 360° scene is equivalent to the temporal distance between two collaborators. Thus, as illustrated in Figure 6, if one user pauses the video while the other continues to watch, the former will see the latter's avatar gradually going away and shrinking in size as it moves farther down the route in the video tour. As illustrated in Figure 7, to maintain the illusion that collaborators are navigating an environment rather than a video (DG1), collaborators' avatars are presented using a silhouette representation if they would normally be occluded by the architecture of the space.

Collaborators can use the embodiments to engaging and disengaging smoothly with each other through view synchronization. A user can also assume a spectator role by double clicking on another collaborator's avatar, which synchronizes both collaborators so that both playback and view orientation are synchronized to the guide. When the leader now navigates through the video or changes their orientation, the other user's view also changes. Users can regain control by simply moving their orientation or explicitly navigating once again. This allows them to move in and out of engagement with one another smoothly (DG3).

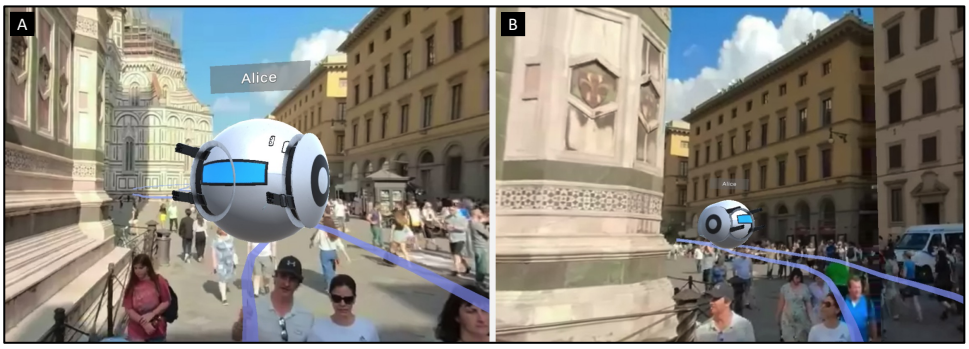

Figure 6: One user (Bob) sees his partner's (Alice) avatar while playing the video. (a) Both users are close to each other in the video (b) Alice pauses the video, while Bob continues to play it, thus seeing Alice's avatar gradually becoming farther and smaller as she stays in place.

Figure 7: A collaborator's avatar is rendered as a silhouette if they would normally be occluded by the environment (here, by the wall of the building).

Pseudo-spatial annotation and allocentric navigation. As illustrated in Figure 8, users can annotate the environment and navigate around it using external artificially created landmarks – pseudo-spatial markers that are placed by the users directly in the environment of 360° video tour during watching. The markers are visible to everyone and, from the user perspective, diminish or increase in size depending on the closeness of the user to them. This allows users to communicate about the environment via deictic reference (DG4) and reinforces the notion that the annotations are about the semantic space (DG1).

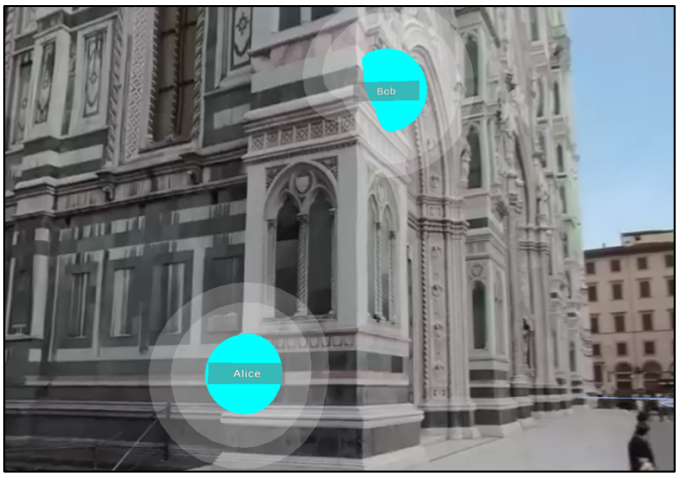

Figure 8: Representation of pseudo-spatial markers placed on the walls of the Florence Cathedral by two users (Alice and Bob).

Users can instantiate markers by double clicking at the point of interest in the environment where they want to place them. Clicking on a marker will teleport the users to the point of interest and time in the video when this marker was instantiated. Users also can delete the existing markers by double clicking on them.

Other features. In addition to the spatial interaction affordances, we implemented traditional controls universally used in video playback applications. The main component that supports video playing capabilities is the video control panel including the timeline slider and playback control buttons. The users are also able to pause and resume the video playback by pressing the spacebar on their keyboard.

4 SYSTEM ARCHITECTURE

We created Tourgether360 using the Unity game engine environment. The multi-user functionality was supported by Unity's Multiplayer Networking library (MLAPI). In the following paragraphs, we describe the implementation of system's main technical features.

Route Extraction from 360° Tour Video. The virtual route was extracted using Simultaneous Localization and Mapping technique (SLAM), specifically through its open-source implementation in the package ORB-SLAM [17] run on monocular 360° videos using the omnidirectional camera model on a per-frame basis. The algorithm generated the camera's spatial coordinates (x, y & z) and rotational orientation (quaternions) relative to a central coordinate system, and the virtual route was constructed by sampling from these spatial coordinates at an interval of 0.5 seconds and joining the resulting points.

Invisible 3D layer Overlaid on Top of The Video. In addition to the virtual route, to support all our pseudo-spatial design affordances, we overlaid the video environments played in Tourgether360 with the invisible 3D models of the environments depicted in the video. We extracted the models of the locations from Google Maps geo-information system and aligned the models with the 360° video of the location through a manual calibration process, in which the models were scaled, positioned and rotated to reflect their size, position and angle in the actual video. We used RenderDoc[2] to extract Google Map's 3D buffer cache of the location, removing the unnecessary parts of the 3D models that were not visible in the video using Blender[3]. The 360° video is rendered on a sphere around each user, which engulfs the 3D models, the virtual route, and the avatars of other users. On each client, only the sphere corresponding to the local player is rendered. Each sphere along with the parented user's avatar at their center, move independently along the fixed route in the unity space (Figure 9).

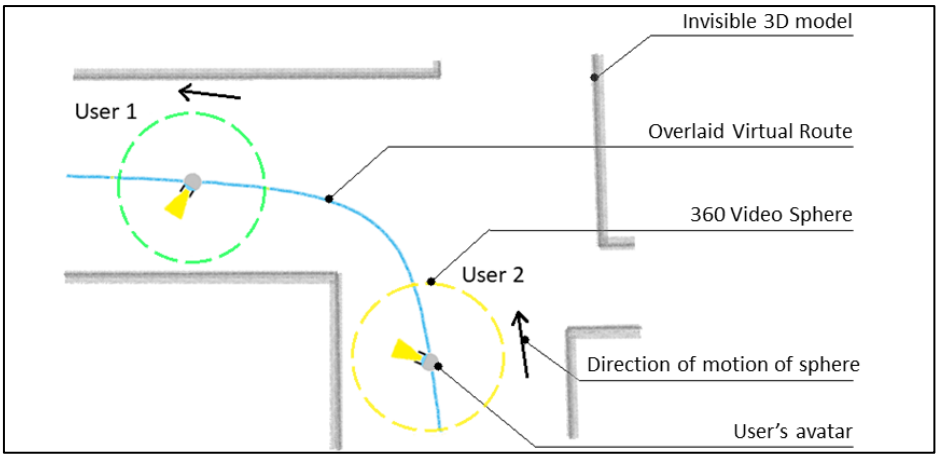

Figure 9: Overhead diagram of two users (and their video spheres) moving along the virtual route in the environment. Here for user 1 only the green sphere and the avatar of user 2 is rendered and vice versa. Note: The spheres are made small for visualization purposes, in the actual system they engulf all of the 3D model and the route.

[2] https://renderdoc.org/
[3] https://www.blender.org/

For aligning the video with the stationary invisible 3D model while the video is playing, a rotation correction is subsequently applied to the video sphere. This was done to account for the movements and rotation of the camera in the 360° video and keep the 3D model of the environment synchronized to the camera changes at all times.

Without applying the correction, the desynchronization between the video and the 3D overlay will occur if the camera in the video rotates. For example, as depicted in Figure 10a, when the person recording the video takes a left turn, the whole environment in the video reorients itself 90 degrees clockwise over the sphere (Figure 10a, slide A1). However, the viewer's avatar still faces the same direction because the video sphere in which the viewer's avatar is situated did not rotate accordingly to the rotation of the camera in the video. Consequently, while the user does not perceive any anomalies, the video is, in fact, no longer in sync with the stationary 3D model in the environment (Figure 10a, slide A2). To correct this, a rotation of the video sphere is applied in the moment of the change in the direction of the camera, where it rotates by the same number of degrees as does the camera. For example, as shown in Figure 10b, if the camera rotates 90 degrees anti-clockwise (left turn), the video sphere rotates 90 degrees in the same direction to account for the spatial inconsistency (Figure 10b, slide B1). This helps to synchronize the video with its 3D overlay when the camera rotates on any of the 3 axes (Figure 10b).

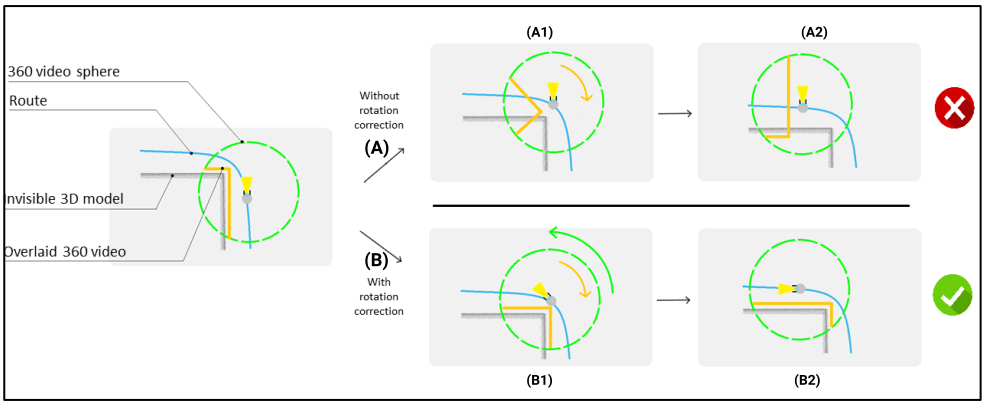

Figure 10: (a) Without Rotation Correction of the sphere, the video layer(orange) desynchonises with the invisible 3D layer(grey) when the camera takes a left turn and therefore the video moves right(90° clockwise) on the sphere.(b) To correct this, the whole sphere rotates back in the same direction as the camera to cancel the video rotation.

As in the case of computation of virtual route, the exact values of camera rotation used above for the rotation correction of the video sphere are computed through the SLAM algorithm, which reports these values in terms of quaternions for each frame in the video.

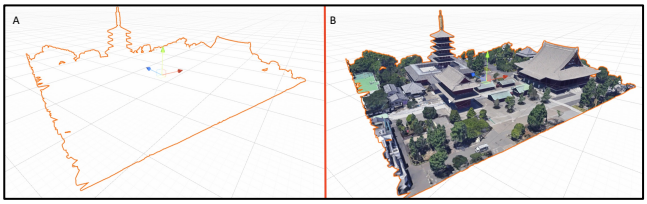

Figure 11: 3D model of Asakusa Shrine Complex that we used as a virtual overlay for the corresponding video used in the study. (a) The model with a custom transparent shader that is used as a direct video overlay (traced for clarity), (b) The textured model used as an overlay for the minimap.

This combination of the video and 3D models provides dynamic occlusion of users' avatars, and the virtual route. For instance, if the other user's avatar moves behind a wall, the avatar changes to a silhouette-like appearance. The implementation of this function was via a custom shader, which although is transparent (to allow for the unobstructed view of the video), tints the shaders of specific objects like avatars to a red fresnel (silhouette) shader when occluded (e.g., Figure 7). Similarly, the path is occluded when it is obstructed by any solid spatial entity (e.g., a wall or a building in the video, as in Figure 4).

Pseudo-Spatial Markers. We implemented the ability to mark spots in the environment by using pseudo-spatial markers, represented by the flashing sphere model. The pseudo-spatial positioning of the markers in space was implemented via a ray-casting technique, where a ray from the mouse cursor points on the screen determined the position of the marker in the location where this ray hit the 3D model of the environment. When markers are created, the markers are instantiated and positioned directly on the 3D model. Users perceive the elements as if they are synchronized with the actual video, appearing to stick to the place where they were instantiated. This is illustrated in Figure 7 where markers appear to be stuck to the actual walls of the building and get large/small depending on the user's temporal and spatial distance to them. Markers which should not be in direct view from the camera (example, being obstructed by a certain face of a building) are occluded by the hidden 3D model.

5 COLLABORATION EXPERIENCE EVALUATION

To evaluate the user experience of collaborative watching of 360° videos using Tourgether360, we conducted a user study where pairs of remote participants watched a series of videos together and performed a series of collaborative tasks. While connected via a video conferencing tool so they had full audio, participants each viewed the 360° videos from their own desktop or laptop computers, connected using the Tourgether360 tool. We asked participants to explore the videos while completing collaborative tasks together. Our goals were to assess the overall experience of the interaction, uncover behavioral and collaborative patterns, understand what functions participants find particularly useful during the study.

5.1 Video Selection

For the user study, we selected four publicly available 360° tour videos. All videos involved a person holding an overhead 360°-camera capturing the environment, moving along a path. In each video, the person holding the camera was not visible, creating an illusion of a first-person view of the environment for users. We chose the videos for novelty (we did not want participants to be familiar with the locations), as well as architectural distinctiveness (we wanted the experiences to be interesting to our participants). Each video represented a virtual tour of a specific popular tourist location: Florence Cathedral (duration: 15:10), Rome's Colosseum (duration: 4:07), and Asakusa Shrine Complex (duration: 10:00) were the locations in the videos used for the study sessions. The Asakusa Market area (duration: 6:00) was used as the learning/practice training video, while the remaining three were used for the main study.

5.2 Task Design

The study tasks were designed to encourage the participants to work in both tightly coupled and loosely coupled modes of interaction. In each of the tasks, the participants were free to use any functionality available to them in the software in whatever ways they deemed appropriate, communicate freely, and work in whatever style they wanted. The tasks were limited in time: the duration for each task was 60% of the duration of the respective video used for the task. This

was done to encourage some loosely coupled work, where participants could work independently of each other to cover more area of the location quicker, in addition to the tight coupling collaborative mode.

In the first task involving the Asakusa Shrine Complex, the participants were asked to locate the biggest shrine and find several points where they thought that the view on the shrine was the best. In the second task involving the video of the Colosseum, the participants were asked to identify several points where they would like to take a picture together. In the third task with the Florence Cathedral, the participants were asked to find the place with the best perspective on the Cathedral and its entrance.

5.3 Procedure

The study was conducted in three stages: introductory, training, and the study. At the introductory stage, we connected with both participants remotely using Skype video conferencing software. The participants were shared links to the written description of the study and the pre-study demographic questionnaire. Once we had verbally reviewed all aspects of the study, including our data collection and handling procedures, we obtained verbal consent to participate. We then invited participants to run the previously downloaded application and connect to the dedicated server.

Once participants had joined the shared server through Tourgether360, we asked them to start with the training video. The researchers then explained each aspect of the interface and all possible actions and operations possible while running the video, including operating the minimap, creating and deleting the markers, and synchronizing each other's views. Participants were then invited to engage in the practice session and play with the app together until they felt completely comfortable with using the software. Once the participants had mastered all the required interactions, they were invited to run the first experimental video and have explained the first study task, subsequently followed by the rest of the tasks. Upon completing all experimental tasks, the researchers had shared the link to the post-study questionnaires. When the participants finished filling the questionnaires, the researchers conducted a semi-structured interview with the participants, focusing on the various aspects of communication, coordination, and overall user experience of the interaction. This study protocol was approved by our Institutional Research Ethics Board.

5.4 Participants

We recruited 16 participants in pairs for the user study. Participants' age was between 18 and 25 years old, four identified as women. Eight of the participants were frequent video game players, with majority of the users playing at least weekly, and having experience with video games that involve navigating in 3D environments, such as Call of Duty, and Counter Strike titles. In addition, all participants indicated familiarity with 360° videos, but stated that they rarely engaged in watching such videos.

5.5 Data Collection and Analysis

We collected the field notes and logs of the participants' interactions with the software. Prior to the study, we also asked one of the participants to share their screen with us via Skype, which we recorded. We collected questionnaire data, as well as post-study interview data. Our analysis was grounded in our observations of participant behavior during the study and guided by their interview responses. We took a thematic analysis approach, grouping this data based on thematic relatedness, drawing common stories about participants' experiences from the data.

6 FINDINGS

Participants reported having very positive experiences with Tourgether360 during the study. Without exception, participants were able to learn and use all the functions of Tourgether360. All the users quickly and confidently navigated the video tours, creating collaborative markers, and using the video timeline slider and the minimap for coordination and communication. In the post-study user-experience survey, the participants rated their overall experience as quite easy (5.9 on the scale of 7), interesting (6.5 on the scale of 7), exciting (6 on the scale of 7), and inventive (6.1 on the scale of 7). The most salient aspects of participants' behaviors emerged around participants' perceptions of the social aspects of the experience, their use of spatial navigation using Tourgether360, and their use of the pseudo-spatial markers for communication and coordination.

6.1 Perception of the Experience

In discussing their experiences with Tourgether360 in the study, participants' overall descriptions centered on the highly social experience provided by Tourgether360, and how it resonated and reflected their experience with video games and virtual world.

Tourgether360 as a Social Experience. Participants described the experience as a fundamentally social one, and engaged in the study tasks as if they were in a shared virtual space. For instance [P2] reported, "*We could see what other people are doing, so it was sort of a social experience and not just a video*". The embodiment of participants' partners within Tourgether360 played a significant role in participants' perception of the social character of the experience: because the participants saw each other as the avatars the scene, they referred to objects in the scene *in relation* to each other's avatars. For example, in one instance, [P3] suggested to his partner: "*[P4], the building to your* left *could also be the main temple*". Groups relied on the embodiments to tell them whether their partners were "with" or "nearby", and adjusted speech and conversation accordingly, where they would beacon one another in different ways (e.g., "Come over here" [P4] if they were far apart, versus, "Look at this" if they were nearby one another). One pair [P9 and P10] took this to the extreme, where they ensured that they stayed in each other's view throughout the entire study. This would ensure they could see each other, and what each was looking at as they "walked" through the space in the video—a sort of visual confirmation that they were "with" each other as they went through the video.

Participants also enjoyed the ability to "synchronize" views with one another like a guide-follower pattern. When a participant would click onto their partner's view (i.e., to become a follower), the partner would then temporarily become a tour guide, showing specific points of interest for discussion. We observed every group do this at least once; however, of note, because participants could do this with relative ease, we observed many groups where partners would fluidly flip-flop these roles (as well as disengage) during the study session depending on the situation. For example, one participant [P10] requested his partner [P9] to sync his view to show him a number of potentially relevant locations. In the process, [P9] mentioned another angle which [P10] has missed. Immediately, [P10] synced to [P9]'s view to see this location. Similarly, P3 had his partner synchronize his view before showing each point of interesting to assesses its relevance [P3] before synchronizing his view to his partner's one to watch see the locations that his partner had found.

Tourgether360 as a Video Game. Participants reported that Tourgether360 felt more like an immersive video game than a 360 video player. When asked to compare their experience of Tourgether360 with other computer-related activities, [P6] reported, "*This reminds me mostly of video games. Although we watch video, the interactivity of the process makes it not like any other video viewing app, and I feel that we are* playing *rather than* watching *it*" (emphasis added). This suggests that navigation mechanism, albeit constrained by the nature of the video (i.e., along a track), allowed sufficient latitude and flexibility to give participants the feeling that they were moving through an active environment.

Similarly, [P1] reported, "*It was sort of an interactive video and not just a regular video*", which is consistent with [P9] who described it as, "*It was a mixture of both, a video watching experience as well as playing games, because although I was watching a normal video, simultaneously I was playing around with the app as well.*" Other participants reported that the experience was akin to online multiplayer games with virtual environments (e.g., first person shooter games)—even if they had relied primarily on the timeline slider to navigate time. Thus, the very presence of others (as they experienced through the embodiments) certainly helped to create this impression and perception. For example, [P11] explained: "*[The experience] felt similar to playing Role Playing Games (RPGs), since I see other users directly, and we can move close to each other or move away".*

The participants' understanding of the experience as an interactive video game was reflected in how they navigated the videos. All groups except one looked at the videos non-linearly, jumping back and forth to different places in the space (and video) to explore the environment. For example, one pair used the minimap at the start of the study session together and saw that the most interesting parts of the environment were located toward the end of the video [P1, P2]. One of the participants jumped straight to this final piece of video and then spent most of the time of the study there [P1]. The participants also engaged into playful interactions using the interactive functionality available to them. In particular, several participants at some point started to jokingly delete each other's markers and laughing about it (e.g., [P3, P4]).

Another common framing of the experience was participants' comparison of the interaction to being present in the real world. [P8] reported, "*I don't even need to go outside. I can go and browse the world with my friends however I want without even coming out of my desk*". Participants were excited by the potentially being able to use the app when planning to go to the real locations in the future. For example, [P10] suggested that a potential use case would be to go to the location in Tourgether360 and figure out and mark all of the interesting parts so that he would know where to look when he was to actually travels there.

6.2 Preference for Spatial Navigation and Wayfinding Behaviors

Participants were able to quickly appropriate the spatial navigation metaphors presented by Tourgether360, and we noted that participants generally used spatial navigation strategies rather than a temporal navigation approach. Tourgether360 gives users two primary ways to navigate and wayfind through the video: the minimap, and the timeline slider. The minimap gave participants the ability to navigate through the video relying on a spatial mental model (e.g., "near the main building in the Shrine Complex", or "at the front of the building" and so on); on the other hand, the timeline slider represents a traditional, time-indexed navigational model of the video (e.g. "at the start", "two minutes into the video", and so on).

While we observed participants using both tools to navigate through the video, participants overwhelmingly relied on spatial navigation strategies—both in terms of which UI elements they used to navigate through the video, and how they communicated with one another. All but three used minimap as the central hub of their navigation and wayfinding activities, rarely ever using the timeline scrubber. They even referred to architectural elements in the minimap and relying on spatial relationships when communicating with one another (e.g., [P15] "*Go to toward the front of this temple, there is the good spot*"; [P10]: "*We'll get a better view from the front side of the church.*"; [P11] *"At least one location is fixed for the photograph, the one from the front side of the shrine."*). Similarly, when [P1] was commenting on his partner's work, *"The markers that you have placed from the side of the building are also very good. Photographs from this angle would be great"* [P1]. Yet another user explained the particular time within the video by using spatial reference instead of time: "*The place where it goes in [the alley], right?*" [P12].

The minimap was also a way that they understood each other through the avatars, the route of the tour represented by the path, and the markers that they put into the environment. In many cases, the participants even said that they mainly looked at the minimap to make sense of their ongoing activity, rather than directly into the video. For example, P2 reported, "*[I] mostly looked at [my partner] on the minimap and not directly in the video because it was easier to understand precisely* where *he is now*" (emphasis added). The minimap allowed a type of precision in how they navigated to locations in space: [P7] reported enjoying the scrolling ability as it allowed him to reach precise locations of interest, since he could understand in detail where the location was exactly and jump to the corresponding route point in the minimap.

Only three of sixteen participants predominantly used a temporal navigation strategy [P3, P8, and P9]. They referred to their own locations and locations of the markers that they put into the environment as the time stamps in the video (e.g., *"I am at one minute and twelve seconds"* [P3], or "*Go to the marker that is on the second minute of the video*" [P3]). In addition, [P8] used temporal references to inform his partner where to go: "*For the front view [on the structure] come to 2 minutes and 20 seconds, and then go right to zero minutes*". At a different stage of the activity, P8 told his partner to go to a building, "*At the third minute*" of the video. The adoption of temporal navigation strategy seemed to result in participants having a harder time maintaining awareness of one another in the study video tours.

At the same time, participants could switch between strategies depending on their needs in the moment. For instance, P2 used both a spatial and a temporal reference (at different times) when trying to direct his partner: "*I am at the marker you placed at the back portion of the building*", adding "*at around 4 minutes*" later as an afterthought [P2]. Another participant commented on the markers placed by his partner as: "*The markers that you have placed from the side of the building are also very good [...] and I guess from the right [of the building] where it is around eight minutes until eight-thirty."*

How participants use tools and how they talk to each other reflects how they are thinking about the video and the environment. Tourgether360 seems to engender and enable the possibility to navigate the video through a spatial mental model, rather than a strictly time-based model.

6.3 Use, Persistence, and Ownership of Markers

While participants did not encounter major challenges using the markers to complete the study tasks, we observed that how participants seemed to want to use markers encompassed a broad class of usages that we had not envisioned at the outset. We discuss how groups used the markers to coordinate activity. Then, we discuss the challenges they encountered with the fact that the markers are persistent, rather than ephemeral in their presentation, which leads to clutter in the space. Finally, we discuss the tensions around ownership that occur because of this persistence.

6.3.1 Use of Markers: Detailed Allocentric Coordination

The markers served as an important mean of allocentric navigation, coordination, and communication. Participants used them to mark places that they would return to, mark places that they wanted to talk about, and discuss the architecture under the markers. As an example, we observed how [P3] requested his partner to put a marker to understand the particular position that the former wanted to discuss: "*Just put a marker, and I'll come there.*" Subsequently, the participants discussed the location together and elected it as a location of choice. The markers helped the collaborators to understand each other's perspective when referring to places in the video they deemed interesting. They anchored conversation, where long verbal exchanges tended to occur around and about the markers. For example, the markers helped [P5] to describe his opinion about the location he thought was particularly relevant and interesting: "*Do you see*

my marker? I think that this one could be the [main shrine in the shrine complex]. This (while referring to the other marker) should be the main gate."

Participants also used markers as a coordination mechanism to maintain awareness of others' activities when they were out of view—as a form of visual feedthrough. This happened most frequently when participants split apart during loosely coupled parts of the task. They monitored one another's activities through the markers that were visible both in the minimap and main view, which served as a way of illustrating work being done. For example, in one study session we observed how one of the participants [P2] jumped to the end of the video immediately after the session had started and proceeded marking several potentially relevant locations. The pair then had discussed whether these markers are meaningful or not, even though one of the participants [P1] had never visited this part of the video himself and viewed them via the minimap.

Although the markers were fundamentally grounded in the space, participants would sometimes refer to an individual marker with a combination of spatial and temporal speech. Typically, this was to clarify which specific marker they were referring to (if a spot had multiple markers). For example, at the end of the study session, [P1] said to his partner: "*OK, so let's [choose] the markers placed at 35 seconds, 52 seconds, and 3 minutes [as our final choice of relevant locations]*". Thus, the markers enabled the participants to meaningfully organize their work, dividing the tasks between them. It also served as a powerful support for communication, allowing the participants to ground their discussions while understanding what precisely the other users referred to.

6.3.2 Persistence of Markers: Lost Ephemeral Context

While all the participants used the markers as the part of their coordination strategy, they sometimes had trouble identifying the locating specific markers further along in the task once many markers had been created. For example, when [P4] wanted to direct his partner's attention to a specific marker in the environment among a series of markers that were already there, his partner was confused which specific marker was under discussion. After spending roughly 15-20 seconds trying to locate the marker, they gave up and continued reviewing other locations. Thus, while the markers served as opportunities for coordination and discussion, participants still needed ways of clarifying *which* marker their partners should look at. Beyond this, the markers in of themselves were sometimes insufficient to clarify *what* partners should look at. For instance, [P2] reports, *"It was mostly verbal descriptions [to clarify]. The markers were not so useful because you cannot see them from everywhere.*" Similarly, [P3] reported, *"Even when you see the marker, it is not always clear what exactly it refers to".* In this respect, the markers served to get partners to roughly the right location, and then the participants would need to clarify what to look at with verbal prompts. We observed all the participants actively referencing the points of interest in the video using verbal expressions, particularly verbal deictic referencing, like: *"look to the right"* or *"go forward and then slow down"* [P9].

The challenge is that while it was easy to create markers, once created, it would take an equal effort to remove them. Thus, markers would be left throughout the video during the study session—regardless of whether they had been placed to just attract someone's attention temporarily as part of conversation (e.g., "Look at this!") versus markers that were intended to mark significant points of interest (e.g., parts of the video that the group expected to return to). Participants commented on their confusion around understanding the markers that were put into the environment, especially toward the end of the study sessions, when large number of markers were already in the video. The participants explained that because some of the markers have already lost their significance or simply due to many markers crowding the view, they were confused as to what each marker meant. For example, [P2] stated that after he encountered the markers that were put there by his partner in the experiment, *"It was hard to understand what this marker refers to."* [P2].

Markers served as a monolithic "spatial communication" mechanism, with participants being unable to distinguish their intention, or what they referred to. Several participants suggested being able to color-code markers to signify intention. Others suggested adding the ability to provide verbal annotation with markers to *"describing the context behind the reason for putting this marker" [P3].* In practice, it may also be useful to include mechanisms that support temporary, ephemeral deictic reference, such as telepointers, as described by [8]. Such a visual mechanism could support conversation without the need to clutter the environment in a persistent way.

6.3.3 Ownership of Markers

We rarely saw markers deleted, regardless of the group. In Tourgether360, markers denote the creator with a label; however, as described above, the reason for a marker's creation was not clear. Participants described feeling reluctant to delete others' markers—even when the environment was extremely cluttered with markers. [P7] explained this as a problem of ownership: *"This marker is not mine, I had not created it, so I don't think that I should have the right to delete it. I don't know why [another person] created it, so I will not meddle with it".* Similarly, P12 mentioned, *"Towards the end of the session, we had marked a lot of points because of which it became confusing. Although I found some of the markers placed by [my partner] to be clumsy, I was reluctant to delete them. Maybe it could be implemented such that if I delete a marker, it gets deleted for me only, so that [my partner] can revisit the marker he had bookmarked."*

Thus, while the markers served a coordinating role for our participants, they still created situations of ambiguity that needed to be resolved verbally. Furthermore, their persistence caused clutter—particularly later on during task sessions—even when their presence was only intended for ephemeral purposes.

7 DISCUSSION AND FUTURE WORK

As ongoing COVID-related restrictions continue to disrupt tourism and travel, social experiences in virtual spaces may become an important way of contributing to people's social and personal well-being. Experiences around 360° videos provide rich, engageable, and realistic content that is well-suited for a range of touristic experiences. However, the current design of 360 video players makes it hard to comfortably enjoy and navigate such videos with others—particularly when the goal is to communicate, coordinate and socializing with other people. Tourgether360 extends prior work on providing spatial means of understanding and watching 360° videos [21] by supporting spatial navigation on an architectural minimap, and simulates a co-habited space with other collaborators. The findings from our study reveal several new questions and issues that are worthy of future study:

Spatial/Semantic Metaphors for Navigating 360 Video. The approach illustrated in Tourgether360 encourages navigation and operation with the data through spatial means. In practice, we observed that references were made to the architecture in the environments (e.g., "Look at the entrance") as well as in relation to participants and partners' avatars (e.g. "Look to my left," or "Look to your left"). In effect, the minimap provides a spatial overview that is akin to the semantic transcripts that can be used to cross-reference into video (e.g. [11]). While it is possible to imagine many semantic layers being applied to videos to support navigation (e.g., labeling buildings; labeling egress; labeling people or cars; transcripts; labeling businesses, etc.), it is interesting to consider what kinds of semantic metaphors are important to effectively navigate video. To some extent, this likely depends on the nature of the task, and the intention for navigating and studying the video in the first place. By studying the mental maps that people develop as they watch and discuss 360° videos, we may be able to uncover the most effective types of labels and metaphors for navigating 360° video.

"Being in a Place Together" Rather than "Watching Together." While considerable work has explored remote, social TV watching experiences (e.g. [12]), one of the enduring challenges has been to design experiences that are enjoyable for people to use—as if they were collocated and watching together. The approach that Tourgether360 takes tries to take that a step further—rather than considering 360° videos something that can be *watched together*, Tourgether360 allows making 360° videos something people can *inhabit together*: an experience that was made evident by participants' remarks. The flexibility to move around in the video space independently enhanced this sensation of control and immersion. It is interesting to consider what other kinds of media we might be able to design immersive experiences where people feel as if they are "together." We know, for instance, that text conversation with other viewers of livestream broadcasts (and potentially people in the live video) can help people feel "together" (e.g. [12]). One way to conceptualize these conversations is that they are creating virtual conversational places that viewers occupy together [13]. What are other ways that we can create "places" that viewers can cohabit together without creating undesirable burden on the ways they can interact?

Focus+Context as a Metaphor for Collaborative Coordination. In our study, participants move between periods of loosely coupled and tightly coupled collaborative work [10,11] . To coordinate these shifts, participants used the minimap to understand where their partner was, as well as to develop a quick understanding of what they were looking at, and where they were to go. This "context gathering" step is provided by the minimap through the avatar representation (along with the avatar's orientation cone) and the markers. We noted that participants typically studied this minimap first before teleporting themselves to their destination. This resonates with Schneiderman's adage to support overview before providing details on demand. [28]. Our study suggests that beyond simply providing ongoing "workspace awareness" of other's activities, awareness tools can also provide this sort of "context" information for when a collaborator needs to shift their view of the workspace dramatically. It also suggests that when collaborators need to dramatically shift their position in the workspace, doing so smoothly so as to provide them with some overall understanding of the target destination (and the work collaborators have done there) is useful.

Extensions for Head-Mounted Displays. This work explores 360° video viewing from the perspective of desktop computer use, where providing a minimap in the periphery of the display is an accepted practice (from video games). Yet, it is unclear how to modify this approach for head-mounted displays. We are actively pursuing how to provide awareness of others' activities in 3D workspace when collaborators are wearing head-mounted displays.

Extensions for Non-Tour 360° Videos. Our approach relies on 360° video tours, where a single camera moves through a relatively fixed architectural space. Yet, while many videos on popular platforms are recorded as tour videos, not all 360° videos are recorded in this way. We need to understand how non-tour 360° videos are explored and watched to understand what kinds of approaches would be appropriate for viewing collaboratively with others.

Scaling the Experiences for Groups. One expected use case for viewing 360° tour videos collaboratively is in the primary or elementary school context, where a teacher might take their class on a virtual field trip (e.g., of a museum or a wildlife reserve). In principle, this could be done with a 360° tour video, yet how do awareness mechanisms work for such videos, and for larger groups? As we saw, even with a small group (i.e., a pair of participants) operating in tasks lasting no longer that 15 minutes, the space would be cluttered with markers. New approaches may need to be explored for scaling the experience for groups.

8 CONCLUSION

Tourgether360 provides a way for collaborators to explore and navigate 360° tour videos together with the metaphor of a shared space. We observed in our study that users can easily adopt spatial metaphors for navigation, but that additional

communication and coordination mechanisms may be necessary for a truly effective experience. In spite of these limitations, participants enjoyed inhabiting and exploring new shared space together. Tourgether360 demonstrates that even simple augmentations of 360° videos can change the nature of collaborative experiences, and sheds new light on how we can further improve such collaborative experiences.

REFERENCES

[1] Steve Benford, John Bowers, Lennart E. Fahlén, Chris Greenhalgh, and Dave Snowdon. 1995. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '95), ACM Press/Addison-Wesley Publishing Co., USA, 242–249.

[2] Steve Benford and Lennart E. Fahlén. 1993. Awareness, focus, and aura: A spatial model of interaction in virtual worlds. *ADVANCES IN HUMAN FACTORS ERGONOMICS* 19, (1993), 693–693.

[3] Samuel Dodson, Ido Roll, Matthew Fong, Dongwook Yoon, Negar M. Harandi, and Sidney Fels. 2018. Active Viewing: A Study of Video Highlighting in the Classroom. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (CHIIR '18), Association for Computing Machinery, New York, NY, USA, 237–240.

[4] Nicolas Ducheneaut, Robert J. Moore, Lora Oehlberg, James D. Thornton, and Eric Nickell. 2008. Social TV: Designing for Distributed, Sociable Television Viewing. *International Journal of Human–Computer Interaction* 24, 2 (February 2008), 136–154.

[5] Jeff Dyck and Carl Gutwin. 2002. Groupspace: a 3D workspace supporting user awareness. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '02), Association for Computing Machinery, New York, NY, USA, 502–503.

[6] Diana Fonseca and Martin Kraus. 2016. A comparison of head-mounted and hand-held displays for 360° videos with focus on attitude and behavior change. In *Proceedings of the 20th International Academic Mindtrek Conference* (AcademicMindtrek '16), Association for Computing Machinery, New York, NY, USA, 287–296.

[7] Mike Fraser, Steve Benford, Jon Hindmarsh, and Christian Heath. 1999. Supporting awareness and interaction through collaborative virtual interfaces. In *Proceedings of the 12th annual ACM symposium on User interface software and technology* (UIST '99), Association for Computing Machinery, New York, NY, USA, 27–36.

[8] Saul Greenberg and Gutwin Carl Roseman Mark. 1996. Semantic telepointers for groupware. In *Proceedings Sixth Australian Conference on Computer-Human Interaction*, ieeexplore.ieee.org, 54–61.

[9] Chris Greenhalgh and Steven Benford. 1995. MASSIVE: a collaborative virtual environment for teleconferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 2, 3 (September 1995), 239–261.

[10] Carl Gutwin and Saul Greenberg. 1998. Design for individuals, design for groups: tradeoffs between power and workspace awareness. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (CSCW '98), Association for Computing Machinery, New York, NY, USA, 207–216.

[11] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work* 11, 3–4 (2002), 411–446.

[12] William A. Hamilton, John Tang, Gina Venolia, Kori Inkpen, Jakob Zillner, and Derek Huang. 2016. Rivulet: Exploring Participation in Live Events through Multi-Stream Experiences. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (TVX '16), Association for Computing Machinery, New York, NY, USA, 31–42.

[13] Steve Harrison and Paul Dourish. 1996. *Re-place-ing space.* DOI:https://doi.org/10.1145/240080.240193

[14] Kyoungkook Kang and Sunghyun Cho. 2019. Interactive and automatic navigation for 360° video playback. *ACM Transactions on Graphics (TOG)* 38, 4 (July 2019), 1–11.

[15] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell me where to look. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA. DOI:https://doi.org/10.1145/3025453.3025757

[16] Ville Mäkelä, Tuuli Keskinen, John Mäkelä, Pekka Kallioniemi, Jussi Karhu, Kimmo Ronkainen, Alisa Burova, Jaakko Hakulinen, and Markku Turunen. 2019. What Are Others Looking at? Exploring 360° Videos on HMDs with Visual Cues about Other Viewers. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video* (TVX '19), Association for Computing Machinery, New York, NY, USA, 13–24.

[17] Raúl Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE transactions on robotics* 33, 5 (October 2017), 1255–1262.

[18] Luís A. R. Neng and Teresa Chambel. 2010. Get around 360° hypervideo. In *Proceedings of the 14th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '10*, ACM Press, New York, New York, USA. DOI:https://doi.org/10.1145/1930488.1930512

[19] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. Vremiere: In-Headset Virtual Reality Video Editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 5428–5438.

[20] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (UIST '17), Association for Computing Machinery, New York, NY, USA, 267–277.

[21] Gonçalo Noronha, Carlos Álvares, and Teresa Chambel. 2012. Sight surfers: 360º videos and maps navigation. In *Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia* (GeoMM '12), Association for Computing Machinery, New York, NY, USA, 19–22.

[22] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (UIST '17), Association for Computing Machinery, New York, NY, USA, 289–297.

[23] Benjamin Petry and Jochen Huber. 2015. Towards effective interaction with omnidirectional videos using immersive virtual reality headsets. In *Proceedings of the 6th Augmented Human International Conference* (AH '15), Association for Computing Machinery, New York, NY, USA, 217–218.

[24] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[25] Sylvia Rothe, Mario Montagud, Christian Mai, Daniel Buschek, and Heinrich Hußmann. 2018. Social Viewing in Cinematic Virtual Reality: Challenges and Opportunities. In *Interactive Storytelling*, Springer International Publishing, 338–342.

[26] Gustavo Alberto Rovelo Ruiz, Davy Vanacken, Kris Luyten, Francisco Abad, and Emilio Camahort. 2014. Multi-viewer gesture-based interaction for omni-directional video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), Association for Computing Machinery, New York, NY, USA, 4077–4086.

[27] Mehrnaz Sabet, Mania Orand, and David W. McDonald. 2021. Designing Telepresence Drones to Support Synchronous, Mid-air Remote Collaboration: An Exploratory Study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), Association for Computing Machinery, New York, NY, USA, 1–17.

[28] Ben Shneiderman. 2003. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *The Craft of Information Visualization*, Benjamin B. Bederson and Ben Shneiderman (eds.). Morgan Kaufmann, San Francisco, 364–371.

[29] Anthony Tang and Omid Fakourfar. 2017. Watching 360° Videos Together. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA. DOI:https://doi.org/10.1145/3025453.3025519

[30] Anthony Tang, Omid Fakourfar, Carman Neustaedter, and Scott Bateman. 2017. Collaboration in 360° Videochat: Challenges and Opportunities. DOI:https://doi.org/10.11575/PRISM/31064

[31] Audrey Tse, Charlene Jennett, Joanne Moore, Zillah Watson, Jacob Rigby, and Anna L. Cox. 2017. Was I There? Impact of Platform and Headphones on 360 Video Immersion. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), Association for Computing Machinery, New York, NY, USA, 2967–2974.

[32] Cheng Yao Wang, Mose Sakashita, Upol Ehsan, Jingjin Li, and Andrea Stevenson Won. 2020. Again, Together: Socially Reliving Virtual Reality Experiences When Separated. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[33] Nelson Wong and Carl Gutwin. 2010. Where are you pointing? the accuracy of deictic pointing in CVEs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1029–1038.

[34] Nelson Wong and Carl Gutwin. 2014. Support for deictic pointing in CVEs: still fragmented after all these years'. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (CSCW '14), Association for Computing Machinery, New York, NY, USA, 1377–1387.

[35] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. 2020. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE J. Sel. Top. Signal Process.* 14, 1 (2020), 5–26.

[36] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S. Fels. 2019. Can you believe [1:21]?! In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA. DOI:https://doi.org/10.1145/3290605.3300719