Who Put That There? Temporal Navigation of Spatial Recordings by Direct Manipulation

Klemen Lilija
University of Copenhagen
Copenhagen, Denmark
lilija@di.ku.dk

Henning Pohl
University of Copenhagen
Copenhagen, Denmark
henning@di.ku.dk

Kasper Hornbæk
University of Copenhagen
Copenhagen, Denmark
kash@di.ku.dk

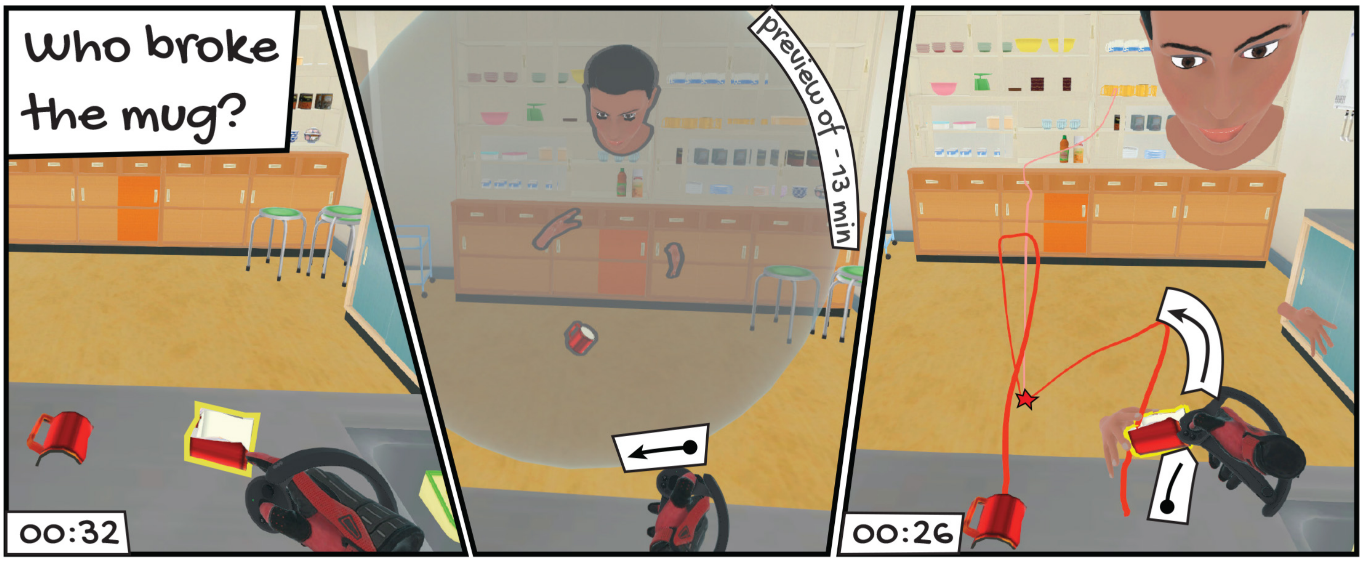

Figure 1. The viewer is in a spatial recording of a virtual kitchen and would like to find out who broke the mug on the kitchen counter (left). By touching one of the mug pieces and stepping through its changes, the viewer previews the accident which happened 13 minutes ago (middle). To see the full story of the mug, the viewer investigates its trajectory and drags the broken piece along it to navigate to the moment it was put on the kitchen counter, 26 minutes into the recording. The *Who Put That There* system enables viewers to preview and navigate by an object's changes.

ABSTRACT

Spatial recordings allow viewers to move within them and freely choose their viewpoint. However, such recordings make it easy to miss events and difficult to follow moving objects when skipping through the recording. To alleviate these problems we present the *Who Put That There* system that allows users to navigate through time by directly manipulating objects in the scene. By selecting an object, the user can navigate to moments where the object changed. Users can also view trajectories of objects that changed location and directly manipulate them to navigate. We evaluated the system with a set of sensemaking questions in a think-aloud study. Participants understood the system and found it useful for finding events of interest, while being present and engaged in the recording.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376604

Author Keywords

temporal navigation; spatial recordings; navigation techniques; virtual reality

INTRODUCTION

Spatial recordings include virtual reality (VR) captures, volumetric recordings, motion captures, and 3D video game replays. In contrast to traditional video, the viewer is *in* a spatial recording and can move within it. Currently, more and more spatial recordings are being created. For example, modern 3D video games such as Fortnite come with the ability to record spatial recordings[1], which can be shared with others. Media companies have also started experimenting with production of VR movies[2] and volumetric captures[3]. Furthermore, future instrumented rooms [21, 14] and augmented reality (AR) devices will likely increase the amount of spatial recordings.

[1] Fortnite replay system, https://www.epicgames.com/fortnite/en-US/news/fortnite-battle-royale-replay-system

[2] Disney's *Cycles*, https://doi.org/10.1145/3214745.3214818

[3] Zero Days VR, https://www.zerodaysvr.com/

Spatial recordings are viewed differently than traditional video. For example, while videos impose a fixed viewpoint, spatial recordings allow viewers to choose their own view and location within them. This gives viewers more control over the experience, but also puts a burden on them to use this freedom effectively. While spatial recordings are consumed differently, the main controls for actively exploring them are the same as for traditional video: a timeline as well as play, stop, fast-forward, and rewind. These controls limit how well viewers can interact with spatial recordings and can cause issues, such as missing of events, motion sickness, or a break of immersion. Further, limiting interaction to the control of *time* complicates exploration of a recording. Navigating by control of time requires skipping to find the moment when an interesting change occurred in the scene. This is especially burdensome for long recordings where it is easy to miss a relevant moment and where the resolution of the timeline is limited.

We propose to structure navigation through spatial recordings along the *changes of objects* therein. Viewers can explore and navigate recordings by directly interacting with objects. Instead of scrubbing through time, viewers can preview an object's changes and go to the moment a change occurred. Continuous changes (e.g., change of location) are previewed in the form of trajectories, while discrete changes (e.g., shape change) are previewed as an animated loop of the change. Figure 1 shows an example of such previews. Navigating by manipulation of objects places control *into* the scene and ensures that viewers do not miss events of interest. This enables more efficient access to interesting parts of a recording and supports sensemaking by allowing users to infer relations among objects. For example, a viewer of a spatially recorded basketball game can inspect the ball's trajectory to see if a player touched it before it went out of bounds. In a spatial recording of a game the viewer might wonder why, when and how a house was build. By stepping through house's changes the viewer can see or infer the answers.

We exemplify the concept via the *Who Put That There* temporal navigation system for spatial VR recordings. The system contains a set of direct manipulation techniques which allow users to step through objects' changes and navigate by objects' trajectories. The system also allows scrubbing through time by using a timeline, as well as rewind, fast-forward, play and pause of the recording by using conventional media controls. We evaluated the system in a think-aloud study using a spatial recording of a virtual kitchen and a set of sensemaking questions. The results confirmed the usefulness of the system and provided insights for designing future object-based temporal navigation systems.

RELATED WORK

Earlier work on spatial recordings focused on ways of capturing them [8, 23, 21, 14] and less so on systems and techniques that improve the viewing experience. In the next subsections we first present conventional ways to temporally navigate recordings and their application to spatial recordings. Then we present the related work on direct manipulation techniques for temporal navigation of recordings.

Conventional Temporal Navigation

Conventional temporal navigation systems typically consist of a timeline as well as play, stop, fast-forward, and rewind controls. These help users find events of interest or get a quick overview of what happened in the video. Such tools can be complimented with automatically generated video summaries (e.g., a teaser) and video abstractions (e.g., a storyboard). Borgo et al. referred to these techniques as *video visualization* and provided a good overview of them in his recent report [2]. It is unclear how the techniques and research findings related to viewing of conventional videos apply to spatial recordings. For example, a study of fast-forward speeds for conventional videos showed that a speed of 1:64 works well for the viewers to still comprehend changes in the scene [32]. However, using such fast-forward speed in spatial recording might make the viewers nauseous or make them miss important events in their periphery. Similarly, conventional techniques for automatic video abstraction are not easily transferable to spatial recordings.

Conventional Temporal Navigation for Spatial Recordings

Current user interfaces for temporal navigation of spatial recording typically use the same controls as the conventional video user interfaces (e.g., [19, 18, 29]). Using conventional tools to navigate spatial recordings can cause issues such as missing of events, motion sickness and breaking of immersion. To combat the motion sickness caused by speeding through the recording, Nguyen et al. proposed a technique which shrinks users' field of view (FOV) dynamically, depending on the motion in a recording [19]. While this technique does mitigate the motion sickness, it might break the immersion or make it more likely to miss events of interest. To avoid missing of events Liu et al. introduced view-dependent video textures [15] which loop a part of the scene until the viewer turns to a right direction to notice an important event. This technique is more suitable for passive viewing than active exploration of spatial recordings. Furthermore, it requires that the producer marks the important events in advance. To avoid braking of the immersion (e.g., while scrubbing the timeline) researchers have investigated more natural ways to interact with recordings. One way to accomplish this is by using gestures for temporal navigation [25, 24, 11, 16]. However, these gestures are often detached from the scene and mostly act as a substitute for play, pause, and rewind buttons.

Temporal Navigation by Direct Manipulation

In 1999, Satou and colleagues introduced the idea of using direct manipulation to temporally navigate videos [26]. They placed polygonal lines over a video to trace objects' trajectories. These lines could then be used as spatio-temporal sliders. For example, by clicking on tennis ball trajectory, the video frame changed to the time when the tennis ball was at the clicked location. The idea to couple the temporal control to video's spatial information gained traction a few years later when researchers investigated how to automate the technique [5, 7, 9, 10, 20]. They developed methods for extracting objects' trajectories, dragging of objects in the video, and ways of dealing with a moving background and change of camera's perspective.

Many of these challenges are specific to using objects' trajectories in a traditional video. Since a conventional video recording is a 2D representation of 3D information, the spatial information is skewed. For example, the distance of an object can only be gauged by its size. Furthermore, the enforced static viewpoint limits the possible interactions as the users cannot move to, for example, look behind an occluder or to follow an object's trajectory out of the current view.

There have been few explorations of direct manipulation techniques in spatial recordings. We are only aware of work by Kuhlen and colleagues, exploring navigation of scientific visualization [33, 35, 12]. Those systems were designed to navigate blood cell simulations, the workings of an impeller and similar. Therefore the interface was designed to assist detailed observation by controlling animation speed, zooming into sub-spaces and marking the spatio-temporal areas of interest. While such techniques incorporate aspects of object-based navigation, there were designed for a specialized use case. Furthermore, the techniques focused only on movement based changes (i.e., trajectories) without considering changes of appearance, configuration, topology and interactions between objects. We believe that temporal navigation by a variety of objects' changes is beneficial for a multitude of spatial recordings, not only for scientific visualizations. Next, we present our interaction concept and a system that exemplifies it.

DIRECT MANIPULATION FOR SPATIAL RECORDINGS

Consider a spatial recording of a capture the flag match. A common way to watch it is to follow a player from the beginning of the recording to the end. However, not all of the recording is equally interesting. Rather, there are pivotal moments in a match, such as when a flag is stolen or a player enters the enemy compound. We posit that structuring and navigating recording by changes is especially useful for spatial recordings. Viewers are able to explore and navigate through a recording by directly interacting with changing objects, instead of having to use a timeline.

Concept

Figure 2 illustrates the concept. The fundamental idea is that viewers of spatial recordings are interested in finding moments of change. Instead of focusing on manipulation of time, we make *changing objects* and their *direct manipulation* central to viewing and navigation.

Change concerns the properties of objects or people (we use the former as a placeholder for both). For example, objects can move, change size or color, split, disappear, change internally or change relation to another object. Depending on the specific form of spatial recording and the nature of the scene, different properties are relevant for the viewer. For example, in a football match, the players' movement, passes, strikes, and penalties can be crucial. While in a poker game the change of emotion or relation between two gazes could interest the viewer. Moreover, changes can be brief, constituting an event, such as a goal being scored, or they can happen over a longer duration of time (e.g., a player dribbling a ball from one side of the field to the other). For spatial recordings, such changes may be given as part of the recording (e.g., objects' events in a VR application). They may also be derived from the recording, for instance by activity-recognition algorithms [13, 27, 28], or through techniques for finding moments of interest (i.e., keyframes) in a recording [30]. Changes structure spatial recordings into meaningful sequences and moments. In comparison to timeline navigation, the focus on changes in many situations aligns better to viewers' questions and interests. With timeline controls, changes are secondary to the movement through time; users need to scrub the timeline while looking around the scene to see the changes occurring.

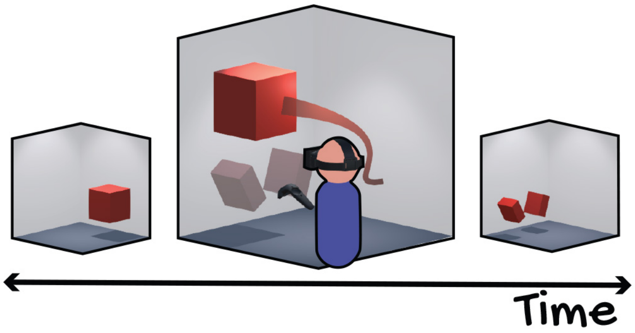

Figure 2. Our concept for interacting with spatial recordings builds on two ideas: (1) The user is interested in changes (e.g., movement of the red cube). (2) Those changes are shown in the scene as previews, for instance as a trajectory showing the cubes movement throughout the recording or as an insert showing that an object has split. Both of these may be used to navigate to the corresponding moment in time.

The second key idea is that a viewer can *directly manipulate* the objects of change to view the spatial recording. For instance, a user might select an object to inspect how it changed over time. Figure 2 shows a user selecting a cube to *preview* its changes; the trace of the cube's movement and the moment it split in half. The previewed changes are shown directly in the scene, keeping the context of the spatial recording. How such changes are shown depends on their nature. If an object moved, this can be visualized as a trajectory (as in [5]). If an object changed shape, the changed object can be shown at the location of the change. With *navigation*, users can directly manipulate objects to move between episodes of change by acting upon a preview and being transported to that moment in the spatial recording. When changes are gradual (e.g., movement, scaling, color change) users can navigate via small units of change. For example, an object's trajectory can act as a non-linear time slider on which users can scrub. If the change is momentary (e.g., object being cut in half), then users are transported to the beginning or end of the change. The change-to-time mapping hence can be discrete or continuous, depending on the nature of the change itself.

In comparison to timeline navigation, direct manipulation of objects places the navigation directly in the spatial recording. For example, viewers can select an object and jump to the moment in time it was last moved. Instead of manipulating *time*, viewers manipulate *objects* for the same effect.

Benefits

Navigating spatial recordings by direct manipulation of objects within the scene has a number of benefits.

Easy Mapping of Intentions to Actions

Embedding navigation actions in a scene makes it easier to find out how to navigate by making the actions visible and concrete. When navigating, users are interested in changes within a scene. To navigate they manipulate the scene. This means that the input and output vocabularies are similar which allows users to easily transfer their intentions into actions. For example, if a user is interested in putting *that* over *there* then pointing to *that* and then *there* is an easily discoverable and executable action [1]. Similarly, when viewers are interested in who put *that* there they can point to it and find out.

Changes are Visible In-Scene

Previewing changes in the scene guides the users' view during navigation. For example, a viewer watching a spatial recording of a theater performance might wonder how a prop entered the stage. Instead of scrubbing through the timeline, the viewer can directly query the object. Once the trace of the object is revealed, the viewers can use it to temporally navigate by manipulating the prop's position, knowing in advance where to focus their gaze. This helps users avoid missing events of interest (e.g., where the prop came from) and being overwhelmed by irrelevant changes when scrubbing the timeline.

Integrated Coarse and Fine-Grained Navigation

Our concept integrates coarse and fine-grained navigation. The temporal resolution of a timeline control is constant and limited by its width. This limitation does not apply to navigation via objects' changes. For example, consider a viewer who is interested in the movement of a tiger in a ten-hour spatial recording. The tiger might be asleep for the first five hours, then wake up and walk around slowly for a bit, sprint to the boundary of the enclosure, and then go back to sleep for another four hours. Viewers scrubbing through such a recording could easily miss the sprinting tiger. Yet, when navigating by changes of the tiger's state, this is easily found. At the same time, fine-grained navigation is supported. With our concept, the tiger's movement trace could be shown as preview and allow users to scrub *along the trace*. Hence, temporal scrubbing is replaced by spatial scrubbing (with each position linked to a point in time). While timelines are limited in space, such trajectories can be much longer. They naturally have higher resolution when more change is happening (e.g., movement).

Sensemaking

Our concept supports sensemaking by allowing key questions (e.g., who, what, when) to be answered in ways timelines cannot support. Users can inspect the subjects in question and receive quick answers (e.g., "who left the burger on the table?"). Changes happening to objects are more likely to relate to viewers' questions and interests. With a timeline, questions are secondary to control of time. Users need to look around the scene while scrubbing through time to find the moment or absence of change.

Navigating a recording by using objects' changes also allows for easy discovery of causal sequences. For example, consider a recording where you see a broken coffee mug (as in Figure 1). With direct manipulation, the viewer can navigate to the moments the mug was used (i.e., touched, moved, filled or broken). At that point, the viewer might notice other things in the recording, such as a person interacting with the coffee drinker. The viewer could then follow that person to go back in time to when they entered the room, uncovering what they came in for (and why they disturbed the coffee drinker).

THE WHO PUT THAT THERE SYSTEM

We designed a system that enables users to temporally navigate spatial VR recordings by direct manipulation of objects within the scene. The *Who Put That There* system tracks changes of location, size, appearance (e.g., a pot getting dirty), configuration (e.g., a stove being turned on), and change of shape or topology (e.g., an object breaking), as well as interactions with an avatar (i.e., object being thrown). Users can select an object and preview its changes in two ways: (1) by stepping through notable changes, or (2) by showing the object's trajectory, indicating the change in location. Users can then navigate via the previewed changes by either selecting one of them, or by scrubbing the object's trajectory. The system also includes a timeline and conventional media controls to play, pause, fast-forward, and rewind the spatial recording.

Apparatus

We implemented the *Who Put That There* system using *Unity3D*. We used a *HTC Vive Pro* with *Valve Index Controllers* for all testing and evaluation of the system. These controllers allowed us to implement direct grasping interactions with objects. They also contain triggers and a thumbstick, which we mapped system commands to.

Stepping Through Objects' Changes

To access the moments where an object changed, the users can select objects within the scene and step through their notable changes (see Figure 3). The changes we define as notable are: object going from static to moving or the other way around, object being grasped or released, change of configuration (e.g., microwave being opened or closed), appearance (e.g., a plate getting dirty), and shape or topology (e.g., burger being bitten in). Users can select an object by hovering the controller over it and then use the thumbstick to preview what happened to the object before or after their current time in the spatial recording. For example, Figure 3 shows what happened to the burger before and after being put on the kitchen counter. Once the preview is active the user can continue stepping through all the changes by using the thumbstick.

The preview is rendered in transparent sphere to separate it from the rest of the scene. It shows the selected moment of change as a looped animation. The objects outside the preview sphere stay static at the current time in the recording. To further distinguish the previewed objects, we render a gray outline around the previewed ones. Scale and position are retained and viewers can move to observe the preview from different angles. This preserves the context and supports viewers' sensemaking activities.

Some objects stay static throughout the whole recording. To identify those that did change we included a *reveal objects* functionality which flashes the objects within the scene that changed at any point throughout the recording. Further, when

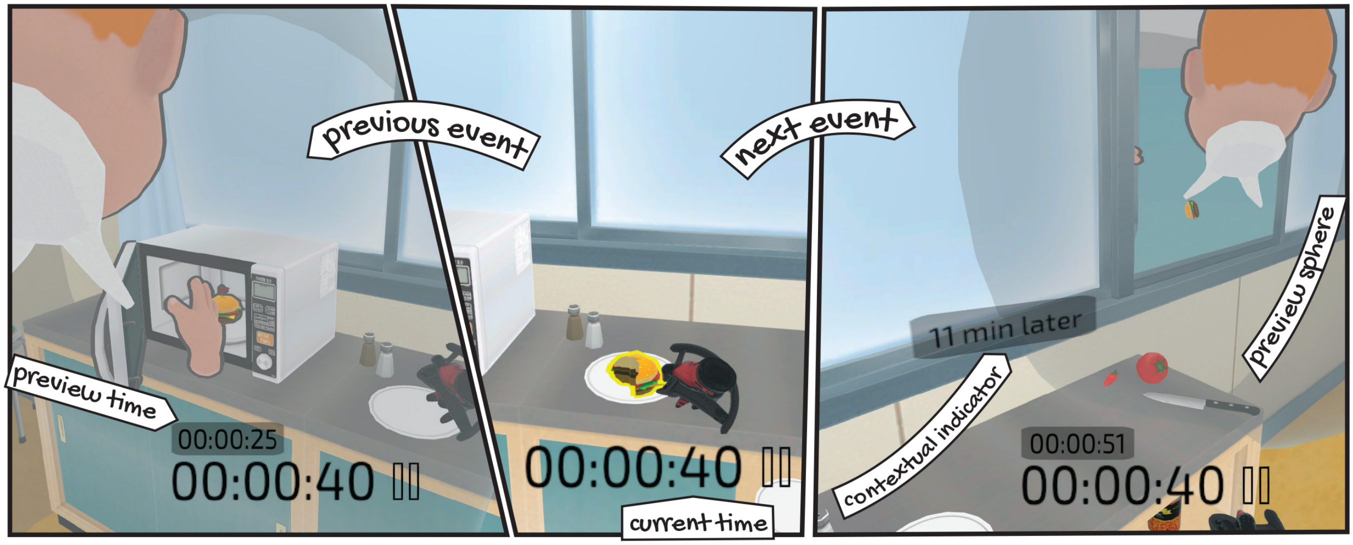

Figure 3. The stepping through objects' changes technique allows users to preview the moments the object had changed. The user can select an object (middle - burger) and step through previous (left - burger being taken out of the microwave) or next changes (right - burger being thrown out of the window) relative to the current point in the recording (00:00:40). The preview is shown within a sphere around the queried object. Objects outside of the sphere are at the current point in the recording, while the objects within the sphere are shown as they are at the previewed moment (00:00:25 and 00:00:51 into the recording). When stepping through changes a contextual indicator floats towards the shown preview while showing the relative time distance to it.

stepping through an object's changes we show a contextual indicator (similar to [4]) which directs the user to the previewed change and indicates the relative time to it.

Skipping through previews allows viewers to quickly skim through all of an object's changes without changing the whole scene around them. Essentially, viewers can peek into the past or future while staying grounded at their current position in the recording. However, we also allow viewers to fully transition to a previewed moment by selecting it. The spatial recording then skips to that moment in time and hence updates the rest of the scene. Where previewing allows viewers to select interesting moments, transitioning to these moments enables a broader set of activities. For example, in a spatial recording of a kitchen a viewer might find the moment the knife was last used. To see exactly how it was used and what else was happening in the other parts of the kitchen they can navigate to that moment and inspect the rest of the scene. For example, they might discover that other people were in the kitchen as well, or that the pan was already on the stove when the vegetables were still being cut.

Scrubbing Objects' trajectory

While stepping through changes works well for events, it is limiting when change occurs over longer duration. For example, when an object is carried around, this results in many location changes. For this case, we developed a trajectory-based navigation technique. To toggle an object's trajectory, viewers select the object and, as shown in Figure 4, add the trajectory to the current scene. Viewers can enable multiple objects' trajectories at the same time. Trajectories are color-coded per object to help distinguish them and to identify the object they belong to.

Trajectories enable a time-agnostic view of an object's history. Hence, they allow viewers to answer a variety of questions about an object, such as: (1) where it came from, (2) whether it was used at a specific location, or (3) whether it covered an area with its movement. For example, users could be wondering where the pot was taken from or if the whole floor was mopped.

To navigate the spatial recording by using a trajectory the viewers are able to scrub along it or jump to any points on the trajectory (see Figure 4). By doing this the user moves to the time in the spatial recording when the object was at the selected position. The trajectory here acts as a form of timeline, yet differs from conventional timeline in many ways: (1) it is anchored in the world instead of being an overlay, (2) constant movement along a trajectory generally results in non-linear movement through time, instead of a fixed time-step, (3) their resolution is dependent on the movement speed of the object instead of the total duration of the recording, and (4) they are specific to one object instead of the whole recording. Trajectories hence take up little space when an object was static, yet allow for fine-grained navigation when objects were moving around a lot or at high speed.

Navigating by scrubbing an object's trajectory concurrently updates the whole spatial recording. In other words, viewers always see a consistent world, in contrast to previewing discrete changes by stepping through an object's changes. Further, trajectories give a snapshot of objects' whole story and help viewers notice events they would not with a more limited preview, or when skipping through the recording. This is valuable for sensemaking as it allows viewers to discover what happened alongside the selected object's changes. For example, as shown in Figure 4, they may discover that the canned beans were put in the pot but not the tomato.

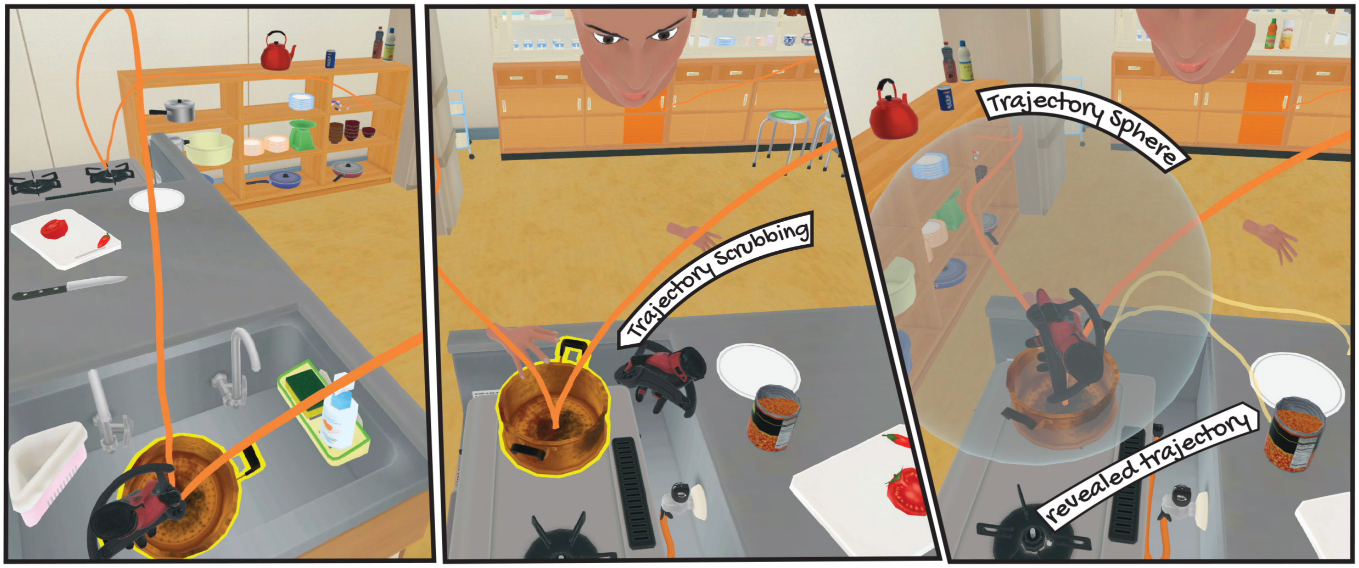

Figure 4. The trajectory scrubbing techniques allows viewers to trace the path of an object and discover events along its history. The user can scrub the trajectory by dragging the object (pot) on it and navigate to the moment the object was at specific location (e.g., stove). To find all the trajectories in a region of space the user can trigger a trajectory sphere which reveals trajectories that passed through it (e.g., trajectory of canned beans leading above the pot).

The techniques we discussed so far only work on visible objects. To show trajectories of absent objects we implemented a trajectory search technique which allows users to find trajectories that passed through a region of space at any point in the recording. The users can enable a *trajectory sphere* which is anchored to one of their controllers. The sphere then acts as a magic lens revealing all the objects' trajectories that passed through it. Users can move around with the sphere to scan the environment and uncover trajectories of interest. When the user disables the sphere the revealed trajectories disappear. To keep a revealed trajectory, the user needs to enable it while the *trajectory sphere* is active. Once the trajectory is enabled they can use it for scrubbing or jumping to any point on it. In addition to facilitating exploration the *trajectory sphere* allows users to inspect and enable objects' trajectories from a distance.

Timeline Scrubbing

In addition to our direct manipulation techniques, we implemented a conventional timeline control (shown in Figure 5). When activated the timeline appeared at the bottom of the user's view and was view-stabilized. Users can scrub the timeline via a raycasting pointer. While our direct interaction techniques allow users to quickly jump to and explore moments of change, conventional timeline controls are still useful. For example, when the user wants to jump to a specific point in time or to the end or beginning of a spatial recording.

Media Controls

To give users the basic control over the spatial recording we included the conventional media controls. The user can rewind, fast-forward, play and pause the recording. These controls enable users to fine-tune the playback while finding or viewing the moment of interest.

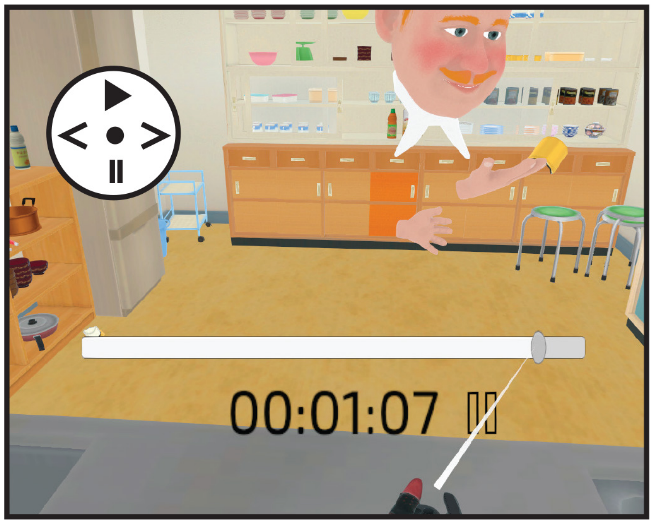

Figure 5. In addition to navigation by direct manipulation of objects, the system also includes a timeline and media controls. The viewers can scrub the timeline via a raycast from the controller to the desired point in time, and use the thumbstick to rewind, fast-forward, play or pause. The current state of the replay system is indicated by an icon next to the time.

EVALUATION

To learn how users experience object-based navigation, we evaluated the *Who Put That There* system in a think-aloud study. In particular, we were interested in whether the system was easy to use, understandable, and useful. Participants were given a spatial recording and a set of sensemaking tasks. To complete the tasks, participants had to view the relevant parts of the recording. We encouraged them to verbalize their thoughts while they carried out the tasks. After they completed the tasks, we conducted an open-ended interview.

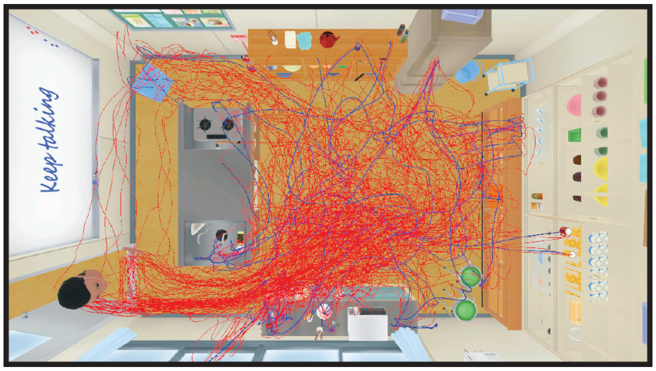

Figure 6. We evaluated the "Who Put That There" system in a 22-minute spatial recording of a kitchen. The blue trajectories show all the objects' movement through the recording and the red trajectories the two avatars' movement.

Spatial Recording

We produced a 22-minute spatial VR recording of a kitchen containing two avatars. The virtual kitchen was 6x4 meters in size and contained two kitchen counters, two shelves, a fridge, microwave, sink, stove, as well as an assortment of kitchen and food items (plates, cups, boxes, cans, vegetables, drinks, etc.). The spatial recording contained a variety of events that commonly occur in real kitchen. For example, in the recording the two avatars arranged kitchen items, cooked, ate, cleaned, as well as dealt with kitchen accidents such as the breaking of a plate. Figure 6 shows a top-down view of the kitchen with the trajectories of all objects and avatars.

Sensemaking Task

To motivate a concrete use of our system, we asked participants to answer a set of sensemaking questions while viewing the spatial recording. We had two types of questions varying in complexity. One type started with a "*What*" referring to an object (e.g., "What happened to the beer can?"). Participants could answer these by observing the object in question while viewing the recording (e.g., the beer was taken off the shelf and put on the counter by one of the avatars, then poured in a glass by another avatar). The *"What"* questions were 1) *what happened with the blue beer can*, 2) *what happened with the orange pot* and 3) *what happened with the red mug*. Questions of the second type started with "*Why*" and were asking for a reason an event had happened (e.g., "Why was the sliding door left open?"). The participants could not answer these questions by solely focusing on the question's subject (i.e., sliding door), but had to understand the context around an event (e.g., one of the avatars broke a plate while preparing a meal, she then hid the broken pieces in the closet leaving the sliding door slightly open before rushing out of the kitchen). The *"Why"* questions were 4) *why was the burger thrown out of the window* 5) *why was the trash bin kicked over* and 6) *why was the sliding door left open*.

Study Design

We split the think-aloud study into three blocks, each of them containing one *"What"* question and one *"Why"* question. In the first block, the participants navigated temporally only by using *media controls* and by *stepping through objects' changes*. In the second block the participants used only *media controls* and *objects' trajectories*. In the third block the participants used only *media controls* and *timeline*. We split the functionality of the system into three blocks to ensure that participants used all the techniques during the study, and could not rely only on those they were most comfortable with. Furthermore, limiting the available interactions allowed us to shorten the training phase and keep the study duration under an hour. We randomly assigned the questions to the blocks per participant. Participants went through a training phase before each of the blocks (e.g., training for *media controls* and the *objects' trajectories* before starting the first block).

Protocol

We recruited 11 participants (4 male, age 23–48, M = 31.9, SD = 6.7) with limited VR experience. We first introduced the participants to the VR setup and the concept of spatial recordings. After, we gave them an overview of the study and introduced them to the think-aloud protocol. To help them familiarize themselves with the protocol we engaged the participants in a simple think-aloud math task. Once the general introduction was over, the participants went through a training phase for the first block, where they were shown how to use the *media controls* and the *stepping through objects' changes* to navigate a spatial recording. For this we used a training recording and in-application controller hints to guide the participants to the right controller inputs.

Once the participants understood the techniques and were comfortable with using them, they were put in the main spatial recording and started with the first *"What"* question. During the task we encouraged participants to keep talking with verbal prompts as well as a large "Keep talking" sign on one of the walls in the virtual kitchen. After they answered the first question they moved on to the *"Why"* question. When they finished with both of the questions they were offered to take a break or to continue with the next block. The procedure for the second and third block was similar except for the techniques being used (i.e., in the second block they used *media controls* and *objects' trajectories*). At the end of the study we conducted an open ended interview to gain additional insight into their comments and the perceived advantages and disadvantages of the individual techniques. On average, the study took 60 minutes to complete.

Data Collection and Analysis

The experimenter took notes of participants' interactions and comments during the think-aloud study. Furthermore, we recorded the participants' comments by using the microphone of the headset and screen captured the participants' view while they were interacting with the spatial recording. We analyzed the collected data by following the approach of Braun and Clarke [3]. We first identified meaningful bits in the data and coded them with shorthands, focusing on participants language, as well as the temporal navigation concepts described in the previous sections. From the codes we identified broad topics and reviewed them, omitting the ones irrelevant to our research focus and merging the related ones into larger themes.

RESULTS

Three topics emerged from the thematic analysis of the think-aloud study and the end-interviews: *being there*, *making sense*, and *having fun*. All participants were able to understand and use all the functionality of the system to answer the questions. We note the observations directly related to usability in the *usability* subsection and discuss the improvements to the system in the discussion section.

Being There

Depending on the used technique, the participants expressed different level of involvement with the spatial recording. P1 stated that when *stepping through objects' changes* and dragging on *objects' trajectories* he felt more "like being there". Similarly, P6 compared scrubbing the global timeline to using the objects' trajectories, and said that the first one is "more like watching a movie" while the second one is "more like playing a game". The physicality of walking and manipulating objects gave the impression of being in an interactive world even when the possible manipulations were constrained to the recorded ones. However, this physicality also disturbed some participants as they would rather interact with objects at a distance and with less effort.

The *Timeline* was experienced as less interactive. Most often, the participants keep the view "pretty steady when using" the timeline (P1). A common pattern was to lock their view on the region of interest, start scrubbing from the beginning of the timeline, overshoot, and then correct to close in on the event of interest (e.g., a beer can being picked up). P4 mentioned that such scrubbing felt like waiting.

Making Sense

When using objects' trajectories to complete the sensemaking tasks participants mentioned greater understanding of what had happened and the context around it. P5 compared using the timeline to reading the ending of a book to find out what happened, and using the objects' trajectories to reading the full book and seeing how it happened. P11 experienced the timeline as efficient, however, it gave her the feeling of "not seeing the whole story" and "missing of events" even when she was sure she missed none. Similarly, P6 experienced trajectories as giving her more knowledge and allow for better reasoning when trying to interpret them.

The additional knowledge encoded in the trajectories could also confuse participants, especially when the trajectories were entangled or were of similar color (P10). On the other hand, many participants were quick to understand the associations between objects and trajectories. P8 mentioned it was easy to notice when an object broke or to which object the trajectory belongs just by the color. Stepping through objects' changes also facilitated active exploration. Participants were quick to inspect changes of objects in question as well as the ones that might be related. For example, when asked to find out "Why the burger was thrown out of the window?" they inspected the fridge and the tomato, to see if those two objects played a role in it.

Having Fun

Most participants mentioned that using the object-based navigation was fun, interesting and "cool". Part of this is the novelty effect which the timeline lacks. P9 mentioned that while the timeline was "definitely the simplest" it was also "the most primitive and not as much fun". However, participants mentioned more than just novelty when interacting with objects' trajectories. P2 especially liked the *trajectory sphere* and felt like it gave her "some kind of superpower". P5 drew parallels to "seeing the code behind the matrix" and "realizing the canvas" on which the interactions happen.

Usability

Participants were quick to learn and use the full functionality of the system, despite most of them having little or no prior VR experience. Object-based techniques required more explanation than timeline and media controls, as participants were familiar with the latter two from traditional video recordings. Except for minor issues, such as miss-clicks because of difficulty using the thumbstick, the participants did not experience any major usability problems when using object-based techniques. Similarly, participants had no troubles when using the media controls, while a few experienced problems with precision when using the timeline. However, in general all participant were able to use all of the systems functionality to navigate the recordings and answer the sensemaking question.

DISCUSSION

We have presented a concept for interaction which helps users navigate spatial recordings, such as recordings of VR experiences, volumetric captures, motion captures, and 3D video games. The concept departs from the idea that viewers of spatial recordings are interested in finding moments of change. Therefore we structure the navigation of spatial recordings by objects' changes. Viewers are able to directly manipulate the objects to preview changes and navigate through time. We illustrated the concept with the *Who Put That There* system, which enables such navigation for spatial VR recordings. We have argued that this way of interacting with spatial recordings helps users navigate in a meaningful manner while being closely integrated with the content of the recording. Next, we discuss the limitations, advance over previous work, potential extensions and future application scenarios.

Limitations

There are limitations to the presented concept, the *Who Put That There* system that exemplifies it, and the evaluation of the system. First, direct manipulation techniques might not be suitable for all scenarios. Our concept requires active engagement with the scene which can become exhausting or cumbersome in case of complex queries. Second, the *Who Put That There* system is only one instance of the presented concept. It does not track, preview and allow navigation by all possible changes that could happen to objects and avatars. Furthermore, the system does not support filtering of the tracked changes or compound operators for multiple objects. Third, the evaluation has limited generalizability as we only tested one scenario and recording duration. For a more comprehensive evaluation of the usability of the specific techniques, a comparison of them in different scenarios would be needed.

Advances over Previous Work

Current interfaces for navigating spatial recordings mainly use timelines, both in commercial interfaces (e.g., the Fortnite Replay System[4], or VR Player[5]) and in research (e.g., Vremiere [19], or [25]). As argued, timeline scrubbing has drawbacks when used for spatial recordings. For example, it makes it easy to miss events and difficult to follow moving objects, it can cause nausea, and can lower immersion. Timelines are excellent for questions about time, but less useful for making sense of spatial recordings when users ask questions about patterns (where is this object usually), causality (why is this here), and agency (who put this here).

We draw inspiration from previous work on direct manipulation of video content [5, 26, 9, 7, 20] and expanded on their concepts. Systems like the one by Dragicevic [5] and Wolter [33] focused on changes in the form of movement. We allow for any type of change and examplify the concept in a *Who Put That There* system, allowing navigation by change of appearance, configuration, and shape, among others. Furthermore, we include previewing and navigating by discrete as well as continuous changes.

We believe that our focus on sensemaking is unique. Earlier work has outlined how people relate to past in real [31] and virtual worlds [17]. It appears that relations to past are rarely solely about time, and are often focused around objects, activities and events. We choose *change* as the unit that connects them; partly because it is easy to integrate it and operate with it in a temporal navigation system. Furthermore, we demonstrate the usefulness of the our system for sensemaking with a qualitative study identifying additional qualities of interaction such as increased sense of presence.

Potential Extensions

The *Who Put That There* system could be extended in several ways. First, support for additional kinds of changes in spatial recordings could be added. For instance, objects might change their proximity to a specific area (e.g., knife leaving the kitchen area), or relative location to another object or an avatar (e.g., a cap being worn backwards instead of forwards). The system could also support person related changes such as change of emotion and other internal states.

Second, objects' changes could be previewed and controlled in several additional ways. For example, changes could be visualized as multicolored sculptures (e.g., similar as in [34]) to be able to preview continuous changes of appearance, shape and size. Another improvement would be to encode more information in trajectories to show their direction, time differences between its points, or interactions between multiple trajectories. They could also visually highlight notable changes such as change of configuration. The previews could also be controlled in more ways. For example, the user could scale the object by pulling it apart and navigate by change of size (e.g., of a plant growing).

Third, many spatial recordings contain a notion of viewpoint or viewpoints; this is the case for VR, AR, and many game recordings. The *Who Put That There* system currently does not handle switching between viewpoints and viewers need to move to a new location on their own. Future systems could enable changing of viewpoints as well as stepping into viewpoints of an avatar. The latter would require ways to mitigating motion sickness, however, if possible this could further increase the sense of "being there".

Fourth, in some cases changes in objects occur over large scales or far away from the viewers location. For example, consider a spatial recording that spans multiple rooms or buildings. To preview changes within such large space, the system could include a mini-map with abstracted preview and teleport transitions to navigate to them.

Further Application Scenarios

Using direct manipulation within spatial recordings to control time is a concept applicable to various settings. However, in our study we have focused on a daily life scenario demonstrated in VR. We envision similar interactions to be possible in AR in the near future. Apart from daily life, there are several other scenarios where the concept could be useful. Game replay systems could be complimented with navigating by changing objects. For example, in a capture the flag game the player could use a trajectory sphere to reveal the flag. Spatial sports recordings such as basketball games could benefit by keeping the viewer immersed while navigating through the highlights. Training material for tasks (e.g., engine assembly) that come in the form of spatial recordings (e.g., AR instructions [22], VR simulations [6], or educational application [6]) could be enhanced to allow trainees to navigate through instructions step by step, and to allow the instructors to quickly review their performance.

CONCLUSION

We proposed structuring the navigation through spatial recordings by objects' changes. The changing objects can be directly manipulated to preview and navigate to moments of interest. We examplify the concept with the *Who Put That There* system, which among other, allows stepping through an object's changes (e.g, change of appearance, configuration, grasp, size), and scrubbing on objects' trajectories. We evaluated the system with a sensemaking task and demonstrated its usefulness. Furthermore we identified qualities such as increased sense of presence, engagement and understanding of activity in the recording.

ACKNOWLEDGEMENTS

We would like to thank Jonas Schjerlund for discussions and help with figures, as well as Valve Corporation for generously providing the Index Controllers. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 648785).

[4] https://www.epicgames.com/fortnite/en-US/news/fortnite-battle-royale-replay-system

[5] http://www.vrplayer.com/

[6] Labster, https://www.labster.com/

REFERENCES

[1] Richard A. Bolt. 1980. “Put-that-there”: Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’80)*. ACM, New York, NY, USA, 262–270. DOI: http://dx.doi.org/10.1145/800250.807503

[2] Rita Borgo, Min Chen, Ben Daubney, Edward Grundy, Gunther Heidemann, Benjamin Höferlin, Markus Höferlin, Heike Leitte, Daniel Weiskopf, and Xianghua Xie. 2012. State of the art report on video-based graphics and video visualization. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 2450–2477.

[3] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological*, H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (Eds.). American Psychological Association, 57–71.

[4] Luca Chittaro and Stefano Burigat. 2004. 3D Location-pointing As a Navigation Aid in Virtual Environments. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI ’04)*. ACM, New York, NY, USA, 267–274. DOI: http://dx.doi.org/10.1145/989863.989910

[5] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video Browsing by Direct Manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’08)*. ACM, New York, NY, USA, 237–246. DOI: http://dx.doi.org/10.1145/1357054.1357096

[6] Nirit Gavish, Teresa Gutierrez, Sabine Webel, Jorge Rodriguez, Matteo Peveri, Uli Bockholt, and Franco Tecchia. 2015. Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments* 23, 6 (2015), 778–798. DOI: http://dx.doi.org/10.1080/10494820.2013.815221

[7] Dan B. Goldman, Chris Gonterman, Brian Curless, David Salesin, and Steven M. Seitz. 2008. Video Object Annotation, Navigation, and Composition. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST ’08)*. ACM, New York, NY, USA, 3–12. DOI: http://dx.doi.org/10.1145/1449715.1449719

[8] Bernhard Kainz, Stefan Hauswiesner, Gerhard Reitmayr, Markus Steinberger, Raphael Grasset, Lukas Gruber, Eduardo Veas, Denis Kalkofen, Hartmut Seichter, and Dieter Schmalstieg. 2012. Omnikinect: real-time dense volumetric data acquisition and applications. In *Proceedings of the 18th ACM symposium on Virtual reality software and technology*. ACM, 25–32.

[9] Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. 2008. DRAGON: A Direct Manipulation Interface for Frame-accurate In-scene Video Navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’08)*. ACM, New York, NY, USA, 247–250. DOI: http://dx.doi.org/10.1145/1357054.1357097

[10] Don Kimber, Tony Dunnigan, Andreas Girgensohn, Frank Shipman, Thea Turner, and Tao Yang. 2007. Trailblazing: Video playback control by direct object manipulation. In *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 1015–1018.

[11] Jarrod Knibbe, Sue Ann Seah, and Mike Fraser. 2014. VideoHandles: replicating gestures to search through action-camera video. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*. ACM, 50–53.

[12] Torsten W. Kuhlen and Bernd Hentschel. 2014. Quo Vadis CAVE: Does Immersive Visualization Still Matter? *IEEE Computer Graphics and Applications* 34, 5 (Sep 2014), 14–21. DOI: http://dx.doi.org/10.1109/MCG.2014.97

[13] Jinna Lei, Xiaofeng Ren, and Dieter Fox. 2012. Fine-grained Kitchen Activity Recognition Using RGB-D. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp ’12)*. ACM, New York, NY, USA, 208–211. DOI: http://dx.doi.org/10.1145/2370216.2370248

[14] David Lindlbauer and Andy D. Wilson. 2018. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, New York, NY, USA, Article 129, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173703

[15] Sean J. Liu, Maneesh Agrawala, Stephen DiVerdi, and Aaron Hertzmann. 2019. View-Dependent Video Textures for 360° Video. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST ’19)*. Association for Computing Machinery, New York, NY, USA, 249–262. DOI: http://dx.doi.org/10.1145/3332165.3347887

[16] Andrés Lucero, Dzmitry Aliakseyeu, Kees Overbeeke, and Jean-Bernard Martens. 2009. An Interactive Support Tool to Convey the Intended Message in Asynchronous Presentations. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology (ACE ’09)*. Association for Computing Machinery, New York, NY, USA, 11–18. DOI: http://dx.doi.org/10.1145/1690388.1690391

[17] Carman Neustaedter and Elena Fedorovskaya. 2009. Capturing and sharing memories in a virtual world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1161–1170.

[18] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017a. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST ’17)*. ACM, New York, NY, USA, 267–277. DOI: http://dx.doi.org/10.1145/3126594.3126659

[19] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017b. Vremiere: In-Headset Virtual Reality Video Editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5428–5438. DOI:http://dx.doi.org/10.1145/3025453.3025675

[20] Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2013. Direct Manipulation Video Navigation in 3D. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1169–1172. DOI: http://dx.doi.org/10.1145/2470654.2466150

[21] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, and others. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 741–754.

[22] Riccardo Palmarini, John Ahmet Erkoyuncu, Rajkumar Roy, and Hosein Torabmostaedi. 2018. A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing* 49 (2018), 215 – 228. DOI:http://dx.doi.org/https://doi.org/10.1016/j.rcim.2017.06.002

[23] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. ACM, 1716–1725. DOI: http://dx.doi.org/10.1145/2818048.2819965

[24] 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)*. ACM, New York, NY, USA, 217–218. DOI: http://dx.doi.org/10.1145/2735711.2735785

[25] 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)*. ACM, New York, NY, USA, 4077–4086. DOI:http://dx.doi.org/10.1145/2556288.2557113

[26] Takashi Satou, Haruhiko Kojima, Akihito Akutsu, and Yoshinobu Tonomura. 1999. CyberCoaster: Polygonal line shaped slider interface to spatio-temporal media. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*. ACM, 202.

[27] L. Sun, U. Klank, and M. Beetz. 2009. EYEWATCHME—3D Hand and object tracking for inside out activity analysis. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 9–16. DOI: http://dx.doi.org/10.1109/CVPRW.2009.5204358

[28] M. Tenorth, J. Bandouch, and M. Beetz. 2009. The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. 1089–1096. DOI: http://dx.doi.org/10.1109/ICCVW.2009.5457583

[29] Balasaravanan Thoravi Kumaravel, Cuong Nguyen, Stephen DiVerdi, and Björn Hartmann. 2019. TutoriVR: A Video-Based Tutorial System for Design Applications in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 284, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300514

[30] Ba Tu Truong and Svetha Venkatesh. 2007. Video Abstraction: A Systematic Review and Classification. *ACM Trans. Multimedia Comput. Commun. Appl.* 3, 1, Article 3 (Feb. 2007). DOI: http://dx.doi.org/10.1145/1198302.1198305

[31] Doménique van Gennip, Elise van den Hoven, and Panos Markopoulos. 2015. Things That Make Us Reminisce: Everyday Memory Cues As Opportunities for Interaction Design. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3443–3452. DOI: http://dx.doi.org/10.1145/2702123.2702460

[32] Barbara M. Wildemuth, Gary Marchionini, Meng Yang, Gary Geisler, Todd Wilkens, Anthony Hughes, and Richard Gruss. 2003. How Fast is Too Fast?: Evaluating Fast Forward Surrogates for Digital Video. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '03)*. IEEE Computer Society, Washington, DC, USA, 221–230. http://dl.acm.org/citation.cfm?id=827140.827176

[33] Marc Wolter, Irene Tedjo-Palczynski, Bernd Hentschel, and Torsten Kuhlen. 2009. Spatial input for temporal navigation in scientific visualizations. *IEEE computer graphics and applications* 29, 6 (2009), 54–64. DOI: http://dx.doi.org/10.1109/MCG.2009.125

[34] Xiuming Zhang, Tali Dekel, Tianfan Xue, Andrew Owens, Qiurui He, Jiajun Wu, Stefanie Mueller, and William T. Freeman. 2018. MoSculp: Interactive Visualization of Shape and Time. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 275–285. DOI: http://dx.doi.org/10.1145/3242587.3242592

[35] Daniel Zielasko, Sven Horn, Sebastian Freitag, Benjamin Weyers, and Torsten W. Kuhlen. 2016. Evaluation of hands-free HMD-based navigation techniques for immersive data analysis. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*. 113–119. DOI:http://dx.doi.org/10.1109/3DUI.2016.7460040