IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 1, 2021, accepted April 19, 2021, date of publication April 29, 2021, date of current version May 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076488

Voxelvideos for Entertainment, Education, and Training

HOLGER REGENBRECHT[1], (Member, IEEE), CLAUDIA OTT[1], NOEL PARK[1], STUART DUNCAN[1], AND JONNY COLLINS[2]

[1]Department of Information Science, University of Otago, Dunedin 9054, New Zealand
[2]Education Perfect, Dunedin 9016, New Zealand

Corresponding author: Holger Regenbrecht (holger.regenbrecht@otago.ac.nz)

This work was supported by the National Science Challenges Science for Technology and Innovation (SfTI) Ātea Grant.

ABSTRACT Volumetric videos allow for a true three-dimensional experience where users can freely choose their viewing angles and be actually immersed in a video clip. High quality video productions are gaining attention and first volumetric video recordings are commercially provided at select places. Unfortunately, the production process is very time, labour, and technology-resource intensive, which requires specialist hardware, software, and production experts. A ''youtube-like'' production, distribution, and experience system would be desirable. Here we present an approach which allows for the creation and interactive replay of three-dimensional videoclips using a novel voxel-based platform—voxelvideos. We can show that our voxelvideos experienced in virtual and augmented reality are effective, enjoyable, and perceived as useful. We hope that our approach and findings will encourage researchers, media experts, and hobbyists to experiment with voxelvideos as a new form of affordable media production and experience.

INDEX TERMS Volumetric video, real-time, 3D video capture, 3D reconstruction, voxels.

I. INTRODUCTION

Short video clips shared over the internet are an omnipresent phenomenon thanks to the availability of affordable recording, processing, editing, distribution, and viewing technologies. At the same time, virtual reality is emerging as an immersive experience platform mainly because of the availability of affordable head-mounted displays, interaction devices, and computers. Unfortunately, the ease with which 2D video clips can be produced is not achieved with virtual reality yet and therefore widespread content creation and sharing is difficult.

We suggest the simplification of the production process, easing the quality requirements, and using voxels with low resolutions for virtual reality video clips. Voxels—volumetric pixels—are three-dimensional points in space which are perceived as gap-less the same way as pixels are perceived as without gaps in two-dimensional pictures and videos. If voxels are recorded in a streaming fashion then they can be placed as three-dimensional videos within a virtual environment to be experienced as voxelvideos. Figure **??** shows a flexible recording setup and users' views of the resulting voxelvideo in virtual reality (VR) and augmented reality (AR), respectively.

For the research presented here, we produced three voxelvideos exemplifying application scenarios for this new kind of medium: (1) entertainment: two musicians playing a folk song, (2) education: a language learning lesson, and (3) training: a yoga instructor teaching poses.

In a user study with 16 participants, we could show that even our very coarse voxel representations, achievable with today's technology, led to believable experiences for the respective desired effects (to entertain, to educate, to instruct) and to a sense of presence when watching the voxelvideos. In addition, we were interested in whether an immersive VR system would produce a different user experience than an AR system where one would see the surrounding environment and their own body for the three application domains. Therefore, our voxelvideos have been studied in three technological settings: immersive VR, optical see-through AR, and, as a baseline, 2D ''youtube-like'' video clips of the same scenes. Our findings suggest that the 3D scenes (VR and AR) are perceived as more effective, with the yoga training scene scoring highest. AR voxelvideos scored higher than the

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

FIGURE 1. Playback of voxelvideos in laboratory environment (centre) experienced in AR (North star HMD left) or VR (Oculus rift HMD right).

VR versions for entertainment (music) and education (language) but not for the training (yoga) scene. Also, the 3D scenes were rated higher for enjoyment and achieve a sense of presence and believability.

Our main contributions are (a) the presentation of the idea of voxelvideos as a new medium of creative expression and experience, (b) a replicable description of a prototype system for recording and playback of voxelvideos (cf. figure 1), and last but not least (c) the report on a user study showing the effectiveness of voxelvideos in three different domains.

We proceed by briefly discussing the related work in the field, in particular volumetric videos and voxel-based VR/MR research. We then describe the recording and playback procedures to produce voxelvideos with our VIMR system, and report in detail a user study on the voxelvideo experience in 2D, VR, and AR conditions.

II. RELATED WORK

The concept of voxelvideos builds on two main areas of previous research: (1) volumetric videos and (2) voxel-based approaches. We will briefly discuss both here.

A. VOLUMETRIC VIDEO PRODUCTION AND PLAYBACK

Producing holographic illusions with the help of virtual and mixed reality techniques has been researched for some time; either for telepresence purposes (e.g. [1]–[3]) or for giving video clips true three-dimensionality (not to be confused with stereoscopic movies). In any case, the scene to be played back, normally featuring people, needs to be captured, stored and/or transmitted, and finally rendered, preferably with free viewpoint control.

A popular example would be Intel TrueView [4]; a system for producing a volumetric reconstruction of an entire sports field. The system consists of multiple high-performance on-site servers connected over fibre to many (up to 38) 5K stationary mounted cameras surrounding the field which produce the volumetric reconstruction which can be rendered from arbitrary points of view after a short processing time (not real-time). They also constructed a four-story 10,000 square foot geodesic dome for 3D filmmaking, called Intel Studios [51]. The whole space is equipped with 100 8K resolution cameras (270 GB/sec of raw footage) where the footage is also processed over their high-performance servers.

A similar approach is followed with Volucap [5], [6] but for the production of high quality 3D movie assets and entire scenes. Here, a diffuse back-lit dome is equipped with 16 stereo camera pairs, each pair connected with a dedicated high-end PC. The resulting volumetric reconstructions achieve (almost) the fidelity of classic 2D movie productions, but in 3D! However, this high quality target requires hundreds of hours of post processing per minute of produced volumetric video (and about 1.5 TB/min for the meshed content).

Microsoft's Mixed Reality Capture Studio [7] uses 106 cameras for capturing (raw footage is 600GB/min, compressed to custom MP4 format at about 400MB/sec to 1.8GB/sec depending on the target device). Their target is a wide-ranging variety of platforms and devices (Unreal, Unity, Windows (native support), ARKit/SceneKit).

The same stationary cameras approach with non-real-time post processing steps is used by 8i [8] and 4D View Solutions [9]. They target end users and developers in the consumer and computer game sectors.

While the above solutions produce high quality results, they are unavailable for "youtube"-like productions. They require specific, static setups, specialists, and time- and resource-intensive post processing.

B. VOXEL-BASED MIXED REALITY

If lower levels of visual fidelity are sufficient, then an ad-hoc volumetric video capturing system can be built with off-the-shelf components (e.g. [10], [11]). In this context either the mesh quality must be reduced or a simplified visual representation must be used, such as voxels [12]. Voxel grids with varying degrees of (low) resolutions scale with the capabilities of hardware and software, and with modern hardware voxels have shown promise as a visual primitive for rendering [13]–[15].

Voxel data structures are often used internally; Microsoft's KinectFusion [16], [17] uses voxels for internal data representation and RemixedReality [18] use a voxel grid to store spatial and temporal transformations which are applied to

the virtual environment, where the virtual environment is reconstructed and rendered as mesh models.

SLAMCast [19] produces voxel reconstructions of static scenes, and uses a memory-efficient signed distance function encoding to transmit the reconstructions to a rendering client. The client then renders the reconstruction as a surface model.

Voxel structures are also useful in point cloud compression [20], and to accelerate operations on point clouds [21]–[23]. Other structures have been used to manage large numbers of voxels in large scenes with OpenVDB [24], however the implementation does not target real-time rendering performance. Nvidia's GVDB [25] library is a GPU accelerated voxel system inspired by OpenVDB. Construction and rendering of the GVDB data structure can be done in real-time, however the included rendering system has no scenegraph features and integration with existing frameworks and game engines is challenging.

Finally, Atomontage [26] is a pure voxel game engine, where both the internal data structure and the visual representation are voxels. Objects and scenes are rendered as gapless voxel reconstructions where the voxels are rendered as cubes rather than using voxels to derive a surface model. This is the rendering paradigm we follow with our voxelvideos.

III. RECORDING AND PLAYBACK

Based on work published in [13], [27] we have developed a voxel-based, immersive mixed reality (VIMR) system which facilitates recording voxelvideos from one or more commodity RGBD cameras, and replaying them with consumer VR systems such as the Oculus Rift. Our voxelvideos consist of a '.vox' file comprising a timestamped sequence of serialize-encoded octrees, accompanied by a JSON metadata file and optionally a number of WAVE-encoded audio tracks. In addition to video metadata (recording time, title, runtime, etc) the JSON file contains a lookup table of semantic labels to apply to voxels, and an audio track configuration block for each audio track. The semantic label scheme allows us to attach audio tracks to an audio-labeled voxel to enable spatial audio rendering. The audio source direction is encoded as Euler angles where regular voxels store their colour information.

For convenience these files are collected in a zip archive with a '.voxz' extension, which we call a voxelvideo, while we refer to the extracted collection of files as a raw voxelvideo.

The following sections describe the methods for reconstructing a spatially coherent scene from multiple RGBD cameras, including a convenient camera registration method, and recording and playback with synchronised audio.

A. RECORDING SYSTEM

Our current recording system consists of four computers and three Kinects pointed at the center of a 2.56m^3 volume, and spaced at roughly equal radial intervals. This enables us to reconstruct and record almost the whole volume at a voxel resolution of 8mm (the camera frustums exclude some of the corners). Other configurations are possible: single-camera single-computer systems, smaller volumes at finer voxel resolutions (down to 1mm), and laptop-based portable systems with multiple tripod-mounted cameras are examples.

In our three-Kinect system three of the computers run camera client software which produces one world-aligned voxel reconstruction (in an octree structure) from one camera's point of view, converts the octree to a serial encoding, and sends it to the fourth computer - the server. On the server, the data is merged, and can either be rendered directly or re-serialized for recording as voxelvideos (or for streaming to a remote renderer for telepresence scenarios). An operator control program, typically running on the server computer, connects to each component (client, server) to control the recording process and modify system parameters at run-time (voxel resolution, background subtraction, etc). The operator control program also handles audio recording, primarily because this simplifies an operator interface for selecting which microphones to use for audio recording.

1) VOXEL RESOLUTION

According to their definition, voxel reconstructions should be gap-less. Camera sensor resolutions impose a limit on the minimum voxel size for gap-less reconstruction at a given distance from the camera, and the height of the capture volume (here 2.56m) imposes a minimum camera distance from the capture volume. Our configuration aims to have true gap-less reconstructions in the centre of the capture space, which gives us our nominal voxel size of 8mm for Kinects reconstructing our 2.56m^3 space.

Smaller spaces with finer resolutions (down to 1mm) and larger spaces with coarser resolutions (reasonably not more coarse than 32mm) are also possible with our system. Our choice of capture volume is driven by a trade-off between achieving finer voxel resolutions and being able to reconstruct a standing person with some room to move.

For a recording of a single person at 30 FPS our three-Kinect system produces about 20 k voxels which gives us a data rate of about 2.74 Mb per second. This is more than an order of magnitude less than what high-quality volumetric video systems afford (by trading in volumetric resolution).

2) CAMERA REGISTRATION

Our system is typically integrated with an Oculus Rift VR system, and the world space is defined by the Oculus HMD and controller tracking coordinate system. This makes the calibration of the entire system easy so that the calibration process can be performed by a lay person. While the resulting calibration precision is not as high as e.g. [28] it is sufficient for our voxel resolutions.

The cameras are registered to world space by tracking an Oculus touch controller (tracked in world space) and a rigidly attached visual marker (tracked in camera space). From that set of point correspondences we estimate a coarse transform from each camera space to world space. For systems which are not integrated with a VR system (e.g. a number of

independent Intel Realsense RGBD cameras) we use a checkerboard to define the world coordinate space and estimate the coarse transform. We then capture a world-aligned point cloud (aligned using the coarse transform) of the capture space from each camera, and use ICP (iterative closest point) to compute a refinement transform to minimize the alignment error from each camera to a designated reference camera. The coarse and fine transforms are combined, saved, and then loaded by each camera client and used to align each camera reconstruction to world space. This arrangement distributes the computational load of producing world-aligned voxel reconstructions, and simplifies spatial filtering to exclude voxels which are outside the bounds of the capture volume.

3) AUDIO SYNCHRONIZATION

We compare voxel frame timestamps to the system clock to ensure real-time drift-free playback even for long voxelvideos, and we render audio using the built-in subsystems in Unreal and Unity, respectively. For synchronous audio playback it is sufficient to ensure that the audio and voxel playback begin at the same time. Therefore we begin audio recording shortly before voxel recording, and then trim the beginning of the audio file to match the first voxel frame. The trim-time for an audio track is stored in its configuration block, and on first playback a trimmed version of the original file is produced. Raw voxel videos typically include both the original and the trimmed file, while the zipped format only uses the trimmed audio file.

B. PLAYBACK SYSTEM

We developed our playback system on a Unity based voxel system [13], [27]. We support the wide field of view optical see-through North Star [29] (AR), as well as the Oculus Rift [30] (VR) for 3D viewing of our voxelvideos. The Unreal engine based VIMR system used for recording voxelvideos is not used for AR playback yet, since the North Star only supports Unity. Users are able to look and walk around freely within a limited space (3m x 3m x 2.5m). Related North Star and Oculus assets were imported into Unity which contained the necessary game objects for HMD integration.

1) VR AND AR HMDs

We equipped the North Star with an Intel RealSense T265 SLAM tracker for determining the device's pose [31]. For each eye, the North Star has two screens which project images onto their corresponding half mirror concave lenses. The North Star was calibrated using a 6×9 checkerboard (27mm) to correct lens distortion of the mirrors. Before startup the device is placed in a fixed location and orientation (on the right corner of the system desk facing left) relative to the world origin. In the Unity editor, two virtual cameras capture images of the scene. These two images are then rendered as textures onto separate quads, which are distorted according to the mirror calibration files. A single virtual orthographic camera in Unity acts as the ''main camera'' viewing the two quads to provide a stereoscopic view. Because the North Star display extends the screen space instead of having its own dedicated viewport, we maximise Unity's ''Game view'' tab as a separate window on this extended display. A standalone program acquires the pose of the SLAM camera, which is communicated over a loopback socket interface to Unity and is applied to the virtual cameras. The SLAM tracker is initialized by ensuring the North Star HMD is moved around (including large translations) when the Unity playback system is started.

The Oculus Rift uses its own outside-in IR constellation tracking, which was calibrated using the provided Oculus Runtime setup. Three Oculus tracking sensors were set up in a triangulated position for 360° tracking coverage.

2) AUDIO

Headband-less, over-ear audio headphones are used for providing playback stereo sound (Koss Clip-On KSC21). Unity's spatial audio is enabled on all audio sources to mimic realistic sound. All imported audio assets are positioned relative to their expected sound source, since the Unity playback system does not support voxel labels or audio voxels. For example, if sound comes from a musical instrument, then it would be manually positioned around the center of the recorded voxelized instrument. Unlike the Oculus Rift, the North Star doesn't have its own audio output.

IV. USER STUDY

A user study was conducted to explore to what degree people perceive 3D voxelvideos as effective when compared with 2D ''youtube''-like video clips. Furthermore we wanted to investigate whether people rate voxelvideos differently when watching in a closed, immersive environment (VR), or an open, augmented environment (AR). Both conditions were applied for three different scenarios of using video clips, chosen from top ranked youtube video categories [32]: music and entertainment as the highest ranked, and education and how-to (training/instructions) as two growing categories. Our concrete examples are a two-person band playing a folk song for music entertainment, a German language learning scene for education, and yoga instruction for training.

In the user study we focused on a) the perceived effectiveness of the 3D videos compared with the 2D videos, b) the general enjoyment and effectiveness when comparing 3D and 2D, and c) the perceived usefulness of the 3D videos for the different scenarios (music, language, yoga). We also asked the participants about their preference (AR or VR) for one chosen scenario (either music, language or yoga) and investigated the degree of presence and co-presence in the two systems (AR and VR).

A. METHODOLOGY

Sixteen participants (8 male, 8 female) between 23 and 60 years (average 36 years) of different ethnicity (13 Caucasian, one Chinese, and two Mixed and NZ Maori/NZ

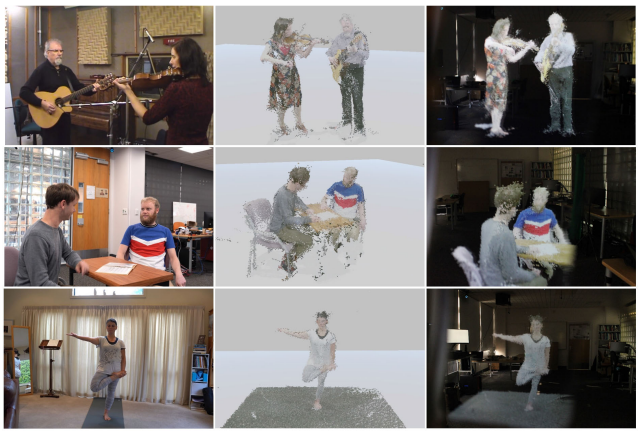


FIGURE 2. Study viewing conditions—left column: 2D video, centre column: VR, right column: AR; top row: entertainment/music, middle row: education/learning, bottom row: training/yoga.

European) took part in the study. Thirteen of the participants had previous experiences with VR and HMDs.

As preparation for the study, three scenes (each between 2.5 and 3 minutes long) were recorded in 2D and 3D, either in one take (language) or in two different takes but with very similar content (music, yoga). The sound for the 3D video was recorded by using headsets and/or a directional shotgun microphone. For the 2D video clips we intentionally chose an amateur recording set-up (lighting and environment) to establish a ''home-made'' look and feel as opposed to a polished, professional video clip. Figure 2 shows all nine resulting conditions with the baseline videos in the left hand column and the two 3D conditions next to them.

To replay the 2D video clips we used a standard 24'' 1080p computer monitor and a pair of closed, quality over-ear headphones. For watching the 3D video clips an Oculus Rift HMD was used for the VR condition or a North Star HMD for the AR condition. Both HMDs were equipped with the same band-less clip-on headphones.

In our study we used a within-group design for the three scenes (music, language, yoga) and a between-group design for the 3D viewing conditions (AR, VR). Therefore, all 16 participants watched all three scenes in 2D and in 3D. Participants were assigned to either VR or AR as their main viewing condition. The main viewing condition as well as the viewing order of the scenes was randomised using Latin Square. As the 2D video was considered to be the ''youtube-like'' benchmark in this study we always asked the participant to watch the 2D video before the 3D video. No start, stop or pause was used while watching the videos. Participants were allowed to adjust the volume of the 2D video.

An experience questionnaire was administered after the study to assess the participants' sense of presence in the environment (based on the MREQ, co-presence questions, and the IPQ; e.g. ''Somehow I felt that the virtual world surrounded me.''), their spatial perception, and any signs of simulator sickness (based on the SSQ). The study was run by one facilitator and supported by one technical operator controlling the 3D viewing modes and scenes. The procedure was as follows:

1) Welcome: Participants were greeted, informed about the study, and asked to sign the consent sheet and complete a six-item, demographic questionnaire.

2) Effectiveness of 3D compared with 2D videos: In the first part of the study the facilitator asked the participant to watch the first, randomly assigned scene in the 2D viewing mode (computer monitor). Upon completion of the video, the facilitator asked one of the following questions (scenario dependent) and the score (not at all 0 . . . 10 very much so) was noted by the facilitator:

- Music: How entertaining did you find the video?
- Language: How educational did you find the video?
- Yoga: How instructive did you find the video?

Depending on the randomly assigned main viewing condition (AR or VR) the facilitator helped the participant to fit either the North Star HMD or the Oculus Rift HMD and asked the participant to watch the voxelvideo. Participants were informed that they were allowed to move and explore while the facilitator would be by their side to watch the cables and make sure that it is safe to move. There was no further conversation. The operator took notes about participants' actions and noted an observed ''activity score'' (not active at all 0 . . . 5 extremely active). When the participant had finished watching the video he/she was helped to take off the HMD and the applicable question (see above) was repeated and the score noted. These five steps of 1) watching the 2D video, 2) asking the question, 3) watching the 3D video, 4) asking the question and 5) asking for comments were repeated for the other two scenes in randomised order.

3) General enjoyment and effectiveness: When the participant had watched all three scenes in 2D and 3D he/she was asked to decide which of the two viewing conditions (2D or 3D) they enjoyed most and which they found more effective for the purpose. The three possible answer choices for the two aspects (enjoyment and effectiveness) were: ''2D'', ''3D'' or ''Can't decide'' followed by an invitation to discuss the reasons for each of the three scenes.

4) Experience questionnaire: After completing the tasks, participants filled in a combined questionnaire with a total of 31 items. The first 8 items were chosen from the igroup presence questionnaire IPQ [33]. The IPQ is an instrument to measure a person's sense of presence in a virtual environment assessing spatial presence, involvement, and realism, which are also relevant factors for mixed reality environments. We left out six items which are only applicable to pure virtual environments. In addition to the application of the IPQ, we administered a four item sub-set of the Mixed Reality Experience Questionnaire (MREQ) [27]. Co-presence was measured by choosing the three co-presence items from [34] and we added three items asking about the perceived usefulness of the 3D video viewing mode for each scenario (entertainment, learning, instruction). All questions used Likert-like scales (7- point). We also included 16 items to

record any signs of simulator sickness [35]. Before filling in the questionnaire, each participant was informed that all questions related to the 3D viewing experience only and that the term ''others'' was related to the characters in the video and not to the facilitator or the operator in the room.

5) Comparison of AR and VR: For the last part of the study the participant was asked to select one scene to watch again under the HMD. The selected scene was not used for their main 3D viewing condition. After the participant had finished watching he/she was asked which of the two 3D viewing conditions they found more effective for the purpose (entertaining, learning, training). The answer choices were: ''AR'', ''VR'' or ''Can't decide'' followed by a quick inquiry into why.

6) Wrap-up: Participants were thanked for the participation and all non-staff members were rewarded with a grocery voucher. Staff members received two blocks of chocolate. The procedure as described above took 40 to 55 minutes.

B. RESULTS AND DISCUSSION

In this section we will report on the results for the seven aspects we wanted to explore.

1) PERCEIVED EFFECTIVENESS (2D AND 3D) FOR THE DIFFERENT SCENARIOS (MUSIC, LANGUAGE, YOGA)

This aspect is investigated by using the scores (0 to 10) which participants reported after watching the 2D video and the 3D video (always second).

The majority of the ratings (36 out of 48) indicate that the 3D viewing mode was more favourable for the purpose with an average increase of 2.0. The remaining quarter of the ratings (12 out of 48) indicate no improvement (9 ratings) or a decrease (3 ratings with an average of -1.5) for the 3D version of the scene (Table 1).

TABLE 1. Perceived effectiveness for conditions and scenes.

Scene	Music			Language			Yoga		
Mode	2D	3D	Diff	2D	3D	Diff	2D	3D	Diff
mean AR	6.6	8.1	1.5	4.3	6.1	1.8	6.1	7.4	1.3
stdev AR	*2.0*	*1.0*	*1.9*	*2.1*	*1.2*	*1.2*	*1.6*	*2.2*	*1.5*
mean VR	4.9	6.0	1.1	4.1	5.2	1.1	6.8	8.3	1.6
stdev VR	*1.8*	*2.1*	*1.3*	*2.5*	*2.2*	*1.4*	*1.6*	*1.0*	*1.6*
mean All	**5.8**	**7.0**	**1.3**	**4.2**	**5.7**	**1.5**	**6.4**	**7.8**	**1.4**
stdev All	***2.0***	***1.9***	***1.6***	***2.2***	***1.8***	***1.3***	***1.6***	***1.7***	***1.5***

When using a paired t-test (Bonferroni correction applied to compensate for multiple comparisons) to explore if those observations are due to chance only we found that the 3D viewing condition was rated statistically significantly higher than the 2D viewing condition for all three scenes (test statistics not included) with the largest effect size for the yoga scene (Cohen's $d = 0.86$).

Given the small sample size further conclusions need to be treated with caution. We could observe that the yoga instruction scene attracted the highest average ratings for the 2D and the 3D viewing mode. In contrast, the language scene seemed to be the least effective scene attracting the lowest average ratings for both the 2D and the 3D viewing mode. Results regarding the comparison of the AR and VR viewing conditions are inconsistent: the AR viewing mode scored higher than the VR viewing mode for music and language but not for the yoga scene. None of the observed differences were statistically significant when using an unpaired t-test and applying Bonferroni correction to compensate for multiple comparisons.

Participants were asked about their ratings and experiences. Regarding the music scene it was mentioned that while it was great to move around (P0: ''I don't have to sit there.'') and find the viewpoint you liked (P3: ''It's great that you get to pick your viewpoint.'') the VR 3D video did not offer a lot of atmosphere (P4: ''Being in a white room doesn't work for me.''), and the sound was not coherent with the wide, white space. In both viewing conditions participants noted the lack of detail when watching the players' hands (P5: ''If it was high definition, it would have been very interesting.'') or the actors' facial expressions (P6 ''Although I could see them from different angles, I could not see their [facial] expressions.''). Most participants however commented on the liveliness (P12: ''Gives it a live feel about it.'', P7: ''It felt like a live performance opposed to people playing in a recording studio.''), the sense of being there (P11: ''I felt like I was part of the band.'') and the ability to see more details (P15: ''I liked how you could see their movements. It's more immersive I suppose.''). For the language scene the participants' comments about the 3D scene were mainly addressing the increased feeling of being there (P13: ''Felt like I was part of the conversation.''), the increased sound quality (P15: ''Sound was better, you could go right up.''), the advantage of moving towards the different sound sources (P0: ''I can move closer to hear their voices better.''), the desire to read the menu on the table (P15: ''Would have been nice to see the menu.'') and the sense of being more immersed was mentioned (P7: ''It was easier to absorb. It was more immersive and engaging.''). The yoga scene attracted a lot of comments on how it was easier to follow the instructor's movements (P1: ''It was easier to see what she was talking about.'') and how it was possible to see what she was doing behind her as well as picking up more details about her body posture (P11: ''I could see more detail. I could also see what she was doing behind her.''). Again not seeing her facial expressions was noted negatively leading to a more impersonal feeling (P13: ''Not being able to see her face made it less real.''). Participants also mentioned wishing they could pause and rewind the video.

2) PERCEIVED GENERAL ENJOYMENT (2D OR 3D) FOR THE DIFFERENT SCENARIOS (MUSIC, LANGUAGE, YOGA)

After watching all scenes in 2D and 3D, participants were asked to choose the viewing mode which they enjoyed most for the three different scenes. Table 2 presents the frequencies of answer choices ''2D'', ''3D'' or ''Can't decide'' (UD for undecided).

TABLE 2. Perceived enjoyment for conditions and scenes.

Scene	Music			Language			Yoga		
Mode	2D	3D	UD	2D	3D	UD	2D	3D	UD
count AR	0	7	1	1	7	0	1	5	2
count VR	2	5	1	0	7	1	1	6	1
count All	**2**	**12**	**2**	**1**	**14**	**1**	**2**	**11**	**3**

The data shows that the majority of participants found the 3D viewing mode more enjoyable than the 2D viewing answering 37 times out of 48 that they preferred the 3D viewing mode for enjoyment. On 6 occasions participants were not able to decide which of the viewing modes they enjoyed more. A preference for the 2D viewing mode was given on 5 occasions with no concentration on a particular scene or viewing condition (AR or VR). We conclude that the viewing condition or the scene did not influence participants' perceived enjoyment as we don't observe large variations in the answer choices.

When inquiring about why 3D was chosen over 2D participants mainly reiterated their remarks initially made about being able to move (P9: ''You are able to explore.'', P10: ''You are not stuck to your chair.'') and the novelty aspect of the 3D video (P9: ''Cause it was different. Adds another level to the enjoyment.''). The low voxel resolution was mentioned as not taking away too much from the experience (P11: ''Even though it was pixelated, it was still lively.'') and on the rare occasion that 2D was chosen over 3D comments were questioning the gain from the 3D experience (P13: ''I don't know if the 3D added much more to the experience than the 2D did.'')

3) PERCEIVED GENERAL EFFECTIVENESS (2D OR 3D) FOR THE DIFFERENT SCENARIOS (MUSIC, LANGUAGE, YOGA)

In a follow-up question, participants were asked to choose the viewing mode which they found more effective (as opposed to enjoyable) for the purpose of the three different scenes. Table 3 presents the frequencies of answer choices ''2D'', ''3D'' or ''Can't decide'' (UD for undecided).

TABLE 3. Perceived general effectiveness for conditions and scenes.

Scene	Music			Language			Yoga		
Mode	2D	3D	UD	2D	3D	UD	2D	3D	UD
count AR	1	5	2	1	6	1	0	8	0
count VR	2	4	2	0	6	2	2	4	2
count All	**3**	**9**	**4**	**1**	**12**	**3**	**2**	**12**	**2**

When asked about the effectiveness, participants seemed more undecided than when asked about the enjoyment. On 9 occasions out of 48 they could not decide resulting in a slightly lower overall count for the 3D viewing mode (33 out of 48). The number of occasions that the 2D viewing mode was preferred remained on a similar level with 6 out of 48. It can be noted that effectiveness ratings for the yoga scene seem to be influenced by the viewing condition (AR or VR). All participants using the AR viewing condition found the 3D viewing mode more effective than the 2D viewing mode as compared with 4 participants who used the VR viewing condition.

Comments on participants' choice of 3D over 2D were similar to the observations made earlier about being immersed (''P5: Because it just felt like you're part of it.'') and closer to the action (P6: ''More close to the interaction.''). When focusing on the effectiveness participants became more aware of the sound quality (P4: ''The acoustics was better. Can hear them properly if I go closer.'' [3D]) and the visual quality of the videos (P12: ''Just double the resolution may enhance the experience.'' [3D], P5: ''Because of better visuals.'' [2D]) and found it more difficult to decide for one or the other viewing mode (P15: ''Can't decide because of resolution stuff.'').

4) PERCEIVED USEFULNESS (VR AND AR) FOR THE DIFFERENT SCENARIOS (MUSIC, LANGUAGE, YOGA)

The experience questionnaire included three questions to rate the usefulness of the 3D video on a 7-point Likert-like scale ranging from -3 not at all to 3 very much.

The average ratings indicate that the participants found the 3D viewing mode useful (Table 4), however to varying degrees between the scenes. The yoga scene scored highest for the AR viewing condition ($M = 2.5$) and highest overall ($M = 2.2$). None of the participants rated the 3D viewing mode for the yoga scene as not useful and only one participant picked a neutral rating for the VR viewing condition. In contrast the language scene attracted the lowest scores with more ratings around the midpoint. Four participants did rate the usefulness negatively with scores < 0 and two participants gave a neutral rating. These results are comparable with the music scene (3 negative and 2 neutral ratings). However 6 participants found the 3D music scene very useful (5 of the AR viewing condition), whereas no participants rated the language scene as high.

TABLE 4. Perceived usefulness for conditions and scenes.

Scene	Music	Language	Yoga
mean AR	1.88	0.75	2.50
stdev AR	*1.76*	*1.30*	*0.71*
mean VR	0.88	0.25	1.88
stdev VR	*1.73*	*1.75*	*0.99*
mean All	**1.38**	**0.50**	**2.19**
stdev All	***1.82***	***1.55***	***0.91***

When tested against the midpoint of 0 in a one sample t-test (Bonferroni correction applied to compensate for multiple comparisons) we found that the music scene as well as the yoga scene were rated statistically significantly above the midpoint (test statistics not included) with the largest effect size for the yoga scene using the AR viewing condition (Cohen's $d = 3.54$).

5) PRESENCE, CO-PRESENCE, AND SIMULATOR SICKNESS MEASURES FOR 3D CONDITIONS (AR AND VR)

The experience questionnaire included 15 items to measure co-presence and presence in the 3D viewing condition.

Unfortunately, there is no absolute measure for presence and co-presence, so we have to use significance tests against the mid-point, assuming that ''0'' actually means neutral. The means of the presence questionnaire (IPQ) and the mixed reality experience questionnaire (MREQ) are significantly above the midpoint as tested with a one-sample t-test assuming unequal variances (df = 15). With a t-critical of 1.73 all t-stat are higher ($p<0.05$) than t-critical (IPQ: 3.30, MREQ: 4.69). The three items measuring co-presence (BAIL) produced a test result significantly below the mid-point (t-critical BAIL: 163.96) as 14 out of 16 participants reported that the virtual characters were not aware of their presence, which is not surprising. We conclude that the 3D voxelvideos induce a sense of presence (IPQ), that objects and characters were reported to be convincing (MREQ), but that the characters in the voxelvideos are not perceived as being people who were actually in the room.

The 16 item simulator sickness questionnaire (SSQ) [35] revealed no serious issues with the system. The measures were computed and the symptoms were classified following Kennedy *et al.*'s work [36]. We computed the SSQ by summing the symptom scores for each participant and the overall mean was computed (M = 0.6875). This SSQ score shows that our system fits the <5 category, which reflects negligible symptoms.

6) PREFERENCE (AR OR VR) FOR ONE CHOSEN SCENARIO (EITHER MUSIC, LANGUAGE OR YOGA)

In the last part of the study participants could choose one scene to watch again using the HMD which was not their main viewing condition. Afterwards they were asked to indicate their preference.

The results suggest no particular preference for either the AR or the VR viewing condition as 8 participants preferred the AR condition and 7 participants the VR condition (Table 5). One participant was undecided when watching the language scene again. Given those results one might speculate whether the participants were biased towards (or against) their main viewing condition, but this is not the case as 6 participants preferred the opposite viewing condition.

TABLE 5. Participants' preference for conditions and scenes.

Scene	Music	Language	Yoga	Sum
count AR	6	1	1	8
count VR	3	1	3	7
count UD	0	1	0	1
count All	**9**	**3**	**4**	**16**

The music scene attracted the most votes to be watched again with over half of the participants (9 out of 16) doing so. We did not inquire why participants chose a particular scene, so we cannot comment on this aspect. We gathered however other interesting comments when asking why a particular viewing condition was preferred.

As indicated by the data, the participants overall did not prefer either viewing condition. One group of participants commented that they liked the AR viewing mode better because they could see the environment (P5: ''Felt more present in the real world.''), experience the real space (P13: ''I think it was because it wasn't in the white space. It felt more realistic.'') and felt safer to move around (P0: ''In AR I can still see [the environment].''). In contrast other participants commented the VR viewing condition was more immersive (P6: ''The immersion was higher.''), less distracting (P2: ''Really shut off any external stimulus.'') and that they enjoyed ''being somewhere else'' (P6: ''AR felt like I was still in the lab. VR felt like I was transferred to another place.''). These contradicting remarks might been caused by the fact that participants did not all choose to watch the same scene, and we may find that depending on the scene, the AR vs. VR choice would be different. However it is also possible that personal preferences in terms of being present and feeling safe as opposed to being fully immersed and ''transported'' play a greater role in their decision making.

7) OBSERVED ''ACTIVENESS'' OF PARTICIPANTS WHEN WATCHING VOXELVIDEOS

When the participants used the HMD to watch the 3D videos the operator observed what they saw on his screen and took notes regarding their movements while the facilitator kept close to the participants making sure that it was safe to move. At the end of each 3D video an activity score was noted between 1 not active at all (standing and not moving) and 5 extremely active (walking around a lot, investigate closely).

In general, participants were not very active apart from moving to find an ideal viewing position. Often this viewing position was not altered a lot during the video watching. We observed that almost all participants were cautious not to go closer than it would be socially acceptable. Nobody walked through the characters, and trying to touch characters was rarely observed. When watching the music scene participants often walked to a position to face them and then did not move any further (score 2) as people would do when watching musicians on a stage. In the yoga scene it was much more common to observe people going around the instructor when movements were happening behind her back (score 3). Nobody actively tried to follow the instructor, which would have been difficult because of the HMD and the cables but also because of participants' awareness that they were being observed. In the language scene we saw different behaviours, from standing stationary after finding a good view point (score 2) to walking around a lot to listen to the different characters (score 3 or 4) and trying to sit at the table at times by kneeling or getting down to the characters' eye-level (score 4).

Differences between the viewing conditions are hard to judge. We would have expected that people move more in the AR viewing condition where they can see the environment and might feel safer to do so. From the average scores (Table 6) it can be noted that the AR viewing condition resulted in higher average activity scores for all three scenes. However, given the small sample size, the variation in the

TABLE 6. Observed activeness for conditions and scenes.

Scene	Music	Language	Yoga
mean AR	2.88	2.88	3.25
stdev AR	*1.25*	*0.99*	*1.04*
mean VR	2.38	2.63	3.13
stdev VR	*0.74*	*0.92*	*0.64*
mean All	**2.63**	**2.75**	**3.19**
stdev All	***1.02***	***0.93***	***0.83***

data, and the coarse subjective measurement of the score itself, no firm conclusions can be drawn.

In general the collected activity scores were a bit disappointing as we had hoped that participants would explore more of the visual and audio effects in the 3D scene (scores between 3 and 5). We expected that people would have become more comfortable and explore more after the first or second viewing, but we could not observe such an increase in activity. The question of why people were not more active would require further investigation. Note: None of the participants showed any signs of activity when watching the 2D video.

C. LIMITATIONS

We are aware that our study design results in a number of threats to validity. Conducting the study in the same lab as the members that produced the 3D voxelvideos inherently results in a bias of being "expected to comment positively" about the new developments. We tried to mitigate this effect by informing the participants that we are genuinely interested in their unbiased opinion and by recruiting mainly mature, critically outspoken people.

Another threat to validity is the non-randomised order of the 2D and 3D viewing condition. We are unable to comment to what degree the watching of the 2D video before the 3D video influenced the participants' ratings. On one hand we think that gathering a baseline rating was necessary to judge the 3D rating and we wanted to avoid the situation where viewing the 2D video after the 3D video would have been experienced as being rather boring. On the other hand we are aware that there was a learning effect as well as the effect of being already familiar with the "plot" which may have resulted in a feeling of being at ease, feeling more relaxed, in control, and being able to pick up more details in the 3D environment.

Bringing people into a virtual or augmented environment has a certain "wow" effect which might overshadow flaws and shortcomings of the system. This novelty effect may wear off after multiple usages of the system and repeating the same study with the same participants may produce lower ratings for the 3D viewing condition. We tried to mitigate this effect by mainly recruiting people who were exposed to AR/VR systems in the past. Only 3 out of the 16 participants had no prior experience with virtual reality and HMDs.

An aspect mentioned by the participants was the different audio qualities of the 2D and 3D versions of the videos. In fact the audio quality of the 3D videos was more favourable because the audio a) was of better general quality for the language and yoga scene (headsets and a good quality microphone was used as opposed to a built-in camera microphone), and b) provided spatial sound. Where the second aspect is an inherent feature of the 3D system, we tried to mitigate the first aspect by professional post processing to reduce the noise level.

Self-reported measures are a tricky instrument as they are very subjective and people have different baselines for ratings. However given the exploratory character of this study we believe that we collected rich and valid data by asking the participants about their experience, thoughts, and inviting comments. More objective measures should be applied in follow-up studies with a focus on hypotheses testing using the results from this study as guidance.

Small sample sizes can always confound results. With only 16 participants we were risking not to discover any tendencies. However we are pleasantly surprised that most of the aspects we explored showed consistent patterns of participants' responses throughout the study. We highlighted the cases where we could not see those patterns in the discussion of the results.

V. CONCLUSION AND FUTURE WORK

We presented the idea of voxelvideos as a new medium for three-dimensional, creative expression. We showed that voxelvideos can be produced in an affordable way and that they can be effective and enjoyable. In a user study we demonstrated a) the perceived effectiveness of the 3D videos compared with the 2D videos, b) the general enjoyment and effectiveness comparing 3D and 2D, and c) the perceived usefulness of the 3D videos for all different scenarios. We also showed that the participants in general preferred 3D over 2D and AR over VR and that the participants felt present in the 3D conditions.

To achieve real-time performance during recording and replay on available, of-the-shelf hardware we tolerated relatively noisy and low resolution voxelvideos. We could show that this fidelity is sufficient for effective volumetric video experiences and therefore opens up a wide range of feasible dissemination options. No computational or bandwidth performance was needed to "prettyfy" the results—all performance went into the delivery of a high enough quantity and quality of voxels.

We used three representative application scenarios to (a) illustrate our concept and (b) evaluate our voxelvideos. Voxelvideos are neither limited to the specific examples we used for each of the three applications—language learning for education, musicians for entertainment, and yoga for training—nor is the list of application areas exhaustive. Voxelvideos are relevant in all scenarios where traditional 2D videos are used today (cf [32]), e.g. music and entertaining dramatic performance; gaming; people & celebrities; sports; comedy; film & animation; science & technology; and even pets & animals.

For instance, it would be possible that the latest chart-breaking pop band not only produces an artistic 2D video clip of their newest song, but that they produce a 3D ''holographic'' experience for their fans to be interactively walked through, perhaps even together with other fans and with some form of mimicked interaction with the band members. Such a pop music voxelvideo would include representations of the musicians, real or artistically designed environment and objects, novel 3D effects, as well as new forms of interactions and communication between band and immersed ''visitors''.

Another promising usage for voxelvideos would be off-site and on-site training in manual tasks in industrial contexts with virtual or augmented reality. Imagine the operation of a complex piece of machinery in a manufacturing context which traditionally has to be learned by way of written and illustrated documentation, 2D videos, and labour-intensive training with instructors. Voxelvideos would allow an instructor to record once and a trainee to watch and practice as many times as needed and wanted. This would include the ability to actually be in the position of the instructor—impossible in real-world situations.

A third, very relevant use case is utilising voxelvideos in the context of therapy and rehabilitation. This is an area of application where virtual reality has proven to be effective in the last two decades. For instance, in the treatment of social phobias, clients could immerse themselves in social situations (a meeting, a party, a talk) populated with recorded and interactive people in a meaningful contextual environment without the actual, massive fear of exposure. Rather, a controlled, systematic desensitisation can be applied, e.g. fewer people at a certain, safe proximity first and later a real crowded scene. Similarly, in stroke rehabilitation, patients can practice lost motor movement capabilities, including neuroplastic recovery, with the immersive, 3D experience of voxelvideos.

Those use cases are just some of many possible applications and fields for voxelvideos. Thinking about transitions from pixels to voxels leads to the exploration of new territory for affordable and scalable volumetric video and interactive experiences.

Voxelvideos also open up novel ways of experiencing virtual and mixed reality environments. Real or computer modelled environments and objects can be integrated into the voxelvideo environment in any form of rendering, including in voxelvideo style (cf [13]) by artificial voxelisation (spatial and temporal). E.g. Atomontage [26] always uses computer models for (high resolution) voxelisation. So, either coherence [37] can be achieved by voxelising everything or intentional non-coherence can be used artistically or otherwise by way of hybrid rendering, e.g. voxels and meshes. The same is true for characters in the scene, either stemming from real-world recordings, as in our user study example, or from virtual avatars. Both forms can be human or algorithmically controlled, e.g. through a game engine or AI behaviour.

Using voxels instead of meshes or point clouds for volumetric videos has advantages which can be exploited for different purposes. The coarseness of the voxel representations achievable today can be seen as an advantage as it explicitly does not ask the viewer for a judgement on graphical realism. Rather it directs attention to behavioural realism and therefore is less likely to suffer from uncanny valley effects [38]. It can also be used for artistic effects akin to non-photorealistic rendering techniques ([39]). Also, the voxel resolution ''naturally'' scales with desired effect and with developing hard- and software; the same techniques can and will be used for different resolutions: in the end voxels are voxels. Computationally, thanks to developments in GPU capabilities and to the decades of voxel-based research, simplicity and elegance can be achieved today which allows for very efficient processing, storage, and transmission of voxels. Readily available computing hardware, periphery, and software can be used to produce voxelvideos.

Many groups can benefit from the novel concept and our implementation of voxelvideos: Designers and researchers in HCI, virtual and augmented reality, psychology, etc., can explore new ways of expression and interactivity and end users can express themselves in new, alternative ways. Those end users might come from the ''youtube'' communities, but also from other creative groups.

Our future work concentrates on research and development of variable voxel resolutions and renderings, on the exploitation of assigning semantic labels to individual voxels and groups of voxels, on merging recorded and live voxelvideos for rich telepresence experiences, on the integration of more interactivity within voxelvideos, and on meaningful environments into which voxelvideos can and will be integrated.

ACKNOWLEDGMENT

The authors would like to thank their participants and the group member of the HCI Lab for supporting their work, in particular Elora Chang, Jonathan Sutton, and Jacob Young. Also, they thank their colleagues for interesting and ongoing discussions about voxelvideos and for comments and revisions on the manuscript, in particular Tobias Langlotz, Oliver Schreer, Kevin Romond, Michael Wagner, and Sandy Garner. A big thanks also goes to our additional ''voxelvideo creatives'': Anna Bowen, Mike Moroney, John Egenes, Farrel Burns, and Goeknil Meryem Biner.

REFERENCES

[1] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, and K. Strehlke, ''blue-c: A spatially immersive display and 3D video portal for telepresence,'' *ACM Trans. Graph.*, vol. 22, pp. 819–827, 2003.

[2] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, and V. Tankovich, ''Holoportation: Virtual 3D teleportation in real-time,'' in *Proc. 29th Annu. Symp. User Interface Softw. Technol.* Tokyo, Japan: ACM, Oct. 2016, pp. 741–754.

[3] S. Beck, A. Kunert, A. Kulik, and B. Froehlich, ''Immersive group-to-group telepresence,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 4, pp. 616–625, Apr. 2013.

[4] I. T. View, ''Get closer to the game with Intel true view,'' Tech. Rep., 2017. [Online]. Available: https://www.intel.com/content/www/us/en/sports/technology/true-view.html

[5] O. Schreer, I. Feldmann, T. Ebner, S. Renault, C. Weissig, D. Tatzelt, and P. Kauff, "Advanced volumetric capture and processing," *SMPTE Motion Imag. J.*, vol. 128, no. 5, pp. 18–24, Jun. 2019.

[6] O. Schreer, I. Feldmann, S. Renault, M. Zepp, M. Worchel, P. Eisert, and P. Kauff, "Capture and 3D video processing of volumetric video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4310–4314.

[7] Microsoft Corporation. (2019). *Microsoft Mixed Reality Capture Studio*. Accessed: Mar. 2021. [Online]. Available: https://www.microsoft.com/en-us/mixed-reality/capture-studios

[8] (2019). *8i—Third Dimensions Storytelling*. Accessed: Mar. 2021. [Online]. Available: https://8i.com

[9] (2019). *4DViews—Volumetric Video Capture Technology*. Accessed: Mar. 2021. [Online]. Available: http://www.4dviews.com

[10] M. Dou, H. Fuchs, and J.-M. Frahm, "Scanning and tracking dynamic objects with commodity depth cameras," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Adelaide, SA, Australia, Oct. 2013, pp. 99–106.

[11] B. Kainz, D. Schmalstieg, S. Hauswiesner, G. Reitmayr, M. Steinberger, R. Grasset, L. Gruber, E. Veas, D. Kalkofen, and H. Seichter, "OmniKinect: Real-time dense volumetric data acquisition and applications," in *Proc. 18th ACM Symp. Virtual Reality Softw. Technol. (VRST)*, New York, NY, USA, 2012, pp. 25–32.

[12] G. K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Dec. 2000, pp. 714–720.

[13] H. Regenbrecht, J.-W. Park, C. Ott, S. Mills, M. Cook, and T. Langlotz, "Preaching voxels: An alternative approach to mixed reality," *Frontiers ICT*, vol. 6, p. 7, Apr. 2019.

[14] K. H. Sing and W. Xie, "Garden: A mixed reality experience combining virtual reality and 3D reconstruction," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2016, pp. 180–183.

[15] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 169:1–169:11, 2013.

[16] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

[17] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 343–352.

[18] D. Lindlbauer and A. D. Wilson, "Remixed reality: Manipulating space and time in augmented reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.* Montreal QC, Canada: ACM, 2018, pp. 1–13.

[19] P. Stotko, S. Krumpen, M. B. Hullin, M. Weinmann, and R. Klein, "SLAM-Cast: Large-scale, real-time 3D reconstruction and streaming for immersive multi-client live telepresence," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 2102–2112, May 2019.

[20] J. Kammerl, N. Blodow, R. B. Rusu, S. Gedikli, M. Beetz, and E. Steinbach, "Real-time compression of point cloud streams," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 778–785.

[21] D. Holz, A. E. Ichim, F. Tombari, R. B. Rusu, and S. Behnke, "Registration with the point cloud library: A modular framework for aligning in 3-D," *IEEE Robot. Autom. Mag.*, vol. 22, no. 4, pp. 110–124, Dec. 2015.

[22] M. Vlaminck, H. Luong, and W. Philips, "Multi-resolution ICP for the efficient registration of point clouds based on octrees," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 334–337.

[23] D. Eggert and S. Dalyot, "Octree-based SIMD strategy for ICP registration and alignment of 3D point clouds," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 3, pp. 105–110, Jul. 2012.

[24] K. Museth, "VDB: High-resolution sparse volumes with dynamic topology," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 27:1–27:22, 2013.

[25] R. K. Hoetzlein, "GVDB: Raytracing sparse voxel database structures on the GPU," in *Proc. High Perform. Graph.* Goslar, Germany: Eurographics Association, 2016, pp. 109–117.

[26] Atomontage. (2019). *Atomontage Inc*. Accessed: Mar. 2021. [Online]. Available: https://www.atomontage.com

[27] H. Regenbrecht, C. Botella, R. Baños, and T. Schubert, "Mixed reality experience questionnaire (MREQ)-reference," Univ. Otago, Dunedin, New Zealand, Inf. Sci. Discuss. Papers Ser. 2017/01. [Online]. Available: https://ourarchive.otago.ac.nz/handle/10523/7151

[28] S. Beck and B. Froehlich, "Sweeping-based volumetric calibration and registration of multiple RGBD-sensors for 3D capturing systems," in *Proc. IEEE Virtual Reality*, Mar. 2017, pp. 167–176.

[29] Leap Motion. (2019). *Project Northstar*. [Online]. Available: https://developer.leapmotion.com/northstar

[30] Facebook Technologies. (2018). *Oculus Rift*. Accessed: Mar. 2021. [Online]. Available: https://www.oculus.com/

[31] Intel Corporation. (2019). *Intel Realsense T265*. Accessed: Mar. 2021. [Online]. Available: https://www.intelrealsense.com/tracking-camera-t265/

[32] X. Che, B. Ip, and L. Lin, "A survey of current YouTube video characteristics," *IEEE Multimedia Mag.*, vol. 22, no. 2, pp. 56–63, Apr. 2015.

[33] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence, Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 266–281, Jun. 2001.

[34] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich, "The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments," *Presence, Teleoperators Virtual Environ.*, vol. 14, no. 4, pp. 379–393, Aug. 2005.

[35] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, Jul. 1993.

[36] R. S. Kennedy, J. M. Drexler, D. E. Compton, K. M. Stanney, D. S. Lanham, and D. L. Harm, "Configural scoring of simulator sickness, cybersickness and space adaptation syndrome: Similarities and differences," in *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues*, p. 247.

[37] J. Collins, H. Regenbrecht, and T. Langlotz, "Visual coherence in mixed reality: A systematic enquiry," *Presence, Teleoperators Virtual Environ.*, vol. 26, no. 1, pp. 16–41, Feb. 2017.

[38] J. Seyama and R. S. Nagayama, "The uncanny valley: Effect of realism on the impression of artificial human faces," *Presence, Teleoperators Virtual Environ.*, vol. 16, no. 4, pp. 337–351, Aug. 2007.

[39] W. Steptoe, S. Julier, and A. Steed, "Presence and discernability in conventional and non-photorealistic immersive augmented reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Sep. 2014, pp. 213–218.

[40] F. Biocca, C. Harms, and J. K. Burgoon, "Toward a more robust theory and measure of social presence: Review and suggested criteria," *Presence, Teleoperators Virtual Environ.*, vol. 12, no. 5, pp. 456–480, Oct. 2003.

[41] S. R. Fussell, R. E. Kraut, and J. Siegel, "Coordination of communication: Effects of shared visual context on collaborative work," in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*, 2000, pp. 21–30.

[42] R. Komiyama, T. Miyaki, and J. Rekimoto, "JackIn space: Designing a seamless transition between first and third person view for effective telepresence collaborations," in *Proc. 8th Augmented Hum. Int. Conf.*, New York, NY, USA, Mar. 2017, pp. 14:1–14:9.

[43] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock, "Mixed reality remote collaboration combining 360 video and 3D reconstruction," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2019, pp. 201:1–201:14.

[44] J. Young, T. Langlotz, M. Cook, S. Mills, and H. Regenbrecht, "Immersive telepresence and remote collaboration using mobile and wearable devices," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 1908–1918, May 2019.

[45] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, "RoomAlive: Magical experiences enabled by scalable, adaptive projector-camera units," in *Proc. 27th Annu. ACM Symp. User Interface Softw. Technol.* Honolulu, HI, USA: ACM, Oct. 2014, pp. 637–644.

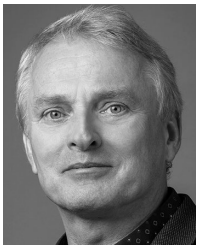
[46] D. Meagher, "Geometric modeling using octree encoding," *Comput. Graph. Image Process.*, vol. 19, no. 2, pp. 129–147, Jun. 1982.

[47] Y. Kim, B. Ham, C. Oh, and K. Sohn, "Structure selective depth superresolution for RGB-D cameras," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5227–5238, Nov. 2016.

[48] Epic Games. *Unreal Engine*. Accessed: Mar. 2021. [Online]. Available: https://www.unrealengine.com/en-US/what-is-unreal-engine-4

[49] S. Birlinghoven and S. Augustin, "Improving the AVANGO VR/AR framework—Lessons learned," in *Proc. Workshop Virtuelle Und Erweiterte Realität*, 2008, pp. 209–220.

[50] Intel Corporation. *Intel NUC*. Accessed: Mar. 2021. [Online]. Available: https://www.intel.com/content/www/us/en/products/boards-kits/nuc.html
[51] Intel Studios. *Intel DevMesh*. Accessed: Mar. 2021. [Online]. Available: https://devmesh.intel.com/projects/the-future-of-immersive-filmmaking-behind-the-scenes-at-intel-studios#about-section
[52] J. Amanatides and A. Woo, "A fast voxel traversal algorithm for ray tracing," *Eurographics*, vol. 87, no. 3, pp. 3–10, 1987.
[53] H. Regenbrecht, K. Meng, A. Reepen, S. Beck, and T. Langlotz, "Mixed voxel reality: Presence and embodiment in low fidelity, visually coherent, mixed reality environments," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2017, pp. 90–99.



HOLGER REGENBRECHT (Member, IEEE) received the Dipl. Inf. and Dr. Ing. degrees majored in computer science with a minor in civil engineering from Bauhaus University Weimar, Germany, and the Ph.D. degree in applied computer science and architecture.

He worked as a Freelance Software Developer and a Research Assistant, until he joined DaimlerChrysler Research and Technology, as a Developer, and later a Group Leader. Since 2004, he has been a Teacher and a Researcher with the Department of Information Science, University of Otago, New Zealand, where he is the currently the Head of the Department. He has been working in the fields of virtual and augmented reality for 25 years. His research interests include human–computer interaction (HCI), applied computer science and information technology, (collaborative) augmented reality, 3D teleconferencing, psychological aspects of mixed reality, three-dimensional user interfaces, and computer-aided therapy and rehabilitation.

CLAUDIA OTT received the Ph.D. degree in computer science education from the University of Otago, in 2015. Before this, she worked as a Freelance Software Developer for a range of companies in Germany. After two years working as a Vision Systems Engineer, she was appointed as a Lecturer at Otago University, teaching at the Department of Computer Science and the Department of Information Science, in 2017. Her research interests include education and learning analytics, and the application of augmented and virtual reality.

NOEL PARK received the degree in computer and information science from the University of Otago. He is currently pursuing the Ph.D. degree, under the supervision of Holger Regenbrecht, working to help Māori in New Zealand, to reconnect their dispersed whanau (family) as part of the Ātea Project. His research interest include virtual/mixed reality, telepresence, HCI, Māori and technology, and has four years of experience in the field.

STUART DUNCAN received the bachelor's and master's degrees in electrical engineering from the University of Canterbury, Christchurch, New Zealand, with a focus on machine vision and embedded systems.

He has worked for several years in industrial automation and research and development, before enrolling as a Ph.D. Student at Otago University with a research topic in computer-aided rehabilitation.

JONNY COLLINS received the bachelor's degree in computer science and the master's and Ph.D. degrees in information science from the University of Otago, New Zealand.

He was responsible for building VR/AR applications under research and development for an industrial automation company before moving to an education technology firm, where he currently works as a Software Engineer. He has been involved in other research areas, such as augmented reality (AR) systems and rehabilitative AR/VR. His primary research interest includes virtual reality (VR) for learning and education.