Virtual Reality Conferencing: Multi-user immersive VR experiences on the web

Simon N.B. Gunkel
TNO, Den Haag, Netherlands
simon.gunkel@tno.nl

Hans M. Stokking
TNO, Den Haag, Netherlands
hans.stokking@tno.nl

Martin J. Prins
TNO, Den Haag, Netherlands
martin.prins@tno.nl

Nanda van der Stap
TNO, Den Haag, Netherlands
nanda.vanderstap@tno.nl

Frank B. ter Haar
TNO, Den Haag, Netherlands
frank.terhaar@tno.nl

Omar A. Niamut
TNO, Den Haag, Netherlands
omar.niamut@tno.nl

ABSTRACT

Virtual Reality (VR) and 360-degree video are set to become part of the future social environment, enriching and enhancing the way we share experiences and collaborate remotely. While Social VR applications are getting more momentum, most services regarding Social VR focus on animated avatars. In this demo, we present our efforts towards Social VR services based on photo-realistic video recordings. In this demo paper, we focus on two parts, the communication between multiple people (max 3) and the integration of new media formats to represent users as 3D point clouds. We enhance a green screen (chroma key) like cut-out of the person with depth data, allowing point cloud based rendering in the client. Further, the paper presents a user study with 54 people evaluating a three-people communication use case and a technical analysis to move towards 3D representations of users. This demo consists of two shared virtual environments to communicate and interact with others, i.e. i) a 360-degree virtual space with users being represented as 2D video streams (with the background removed) and ii) a 3D space with users being represented as point clouds (based on color and depth video data).

CCS CONCEPTS

• **Information systems → Web conferencing**; *Multimedia information systems*; • **Human-centered computing → Virtual reality**;

KEYWORDS

Virtual Reality, VR, Social VR, WebRTC, WebVR, interactive content, immersive virtual environments, WebGL, Point-cloud

ACM Reference Format:
Simon N.B. Gunkel, Hans M. Stokking, Martin J. Prins, Nanda van der Stap, Frank B. ter Haar, and Omar A. Niamut. 2018. Virtual Reality Conferencing: Multi-user immersive VR experiences on the web. In *MMSys'18: 9th ACM Multimedia Systems Conference, June 12–15, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3204949.3208115

MMSys'18, June 12–15, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5192-8/18/06.
https://doi.org/10.1145/3204949.3208115

1 INTRODUCTION

The last few years have seen a major uptake of virtual reality technology, enabling the creation of immersive video-games and training applications, but also paving the way for new forms of video entertainment. One key challenge that many of those VR experiences face is the social barrier. That is, the apparent discrepancy between the physical separation of wearing a head mounted display (HMD) and the human need for sharing experiences. This can also be seen by large investments into social VR from key industry companies such as Facebook, Microsoft and HTC. However, currently the efforts of the big companies mainly focus on artificial and avatar-based representations for people to use in communication. Even though this is good for some use cases, avatar-based approaches may be too restrictive for interactions where non-verbal communication is important, such as conferencing, presentations, watching 360-degree videos together, remote collaboration and many more [9]. Particularly, when talking to familiar people (i.e. family and friends) you like to have a detailed and accurate representation.

To address this problem, we developed a VR framework that extends current video conferencing capabilities with new VR functionalities [3, 4, 10]. Our framework is modular and based on web technologies, and allows to both easily create VR experiences that are social and to consume them with off-the-shelf hardware. With our framework, we aim to allow users to interact or collaborate while being immersed in interactive VR content. In this demo, we focus on a communication use case, where multiple people (up to three) can sit around a table to communicate within VR. We evaluated this experience with 54 people, in unstructured testing sessions with a duration of 3-10 minutes. Each testing session was followed by a questionnaire and informal discussions. These tests where done in a static 360-degree VR environment. Furthermore, in this demo we like to compare such static 360-degree environments with new media formats and full 3D environments. We present an approach to enhance our system with color+depth video data, allowing to show users as 3D point cloud in a 3D VR environment and evaluate this approach on a technical level. Although we believe that such representation will add value, a QoE comparison with 2D is not yet performed and out of scope of this paper.

The paper is structured as follows. In section 2, we give an overview of the current state of the art. Section 3 describes our experiments with 3 users communicating in VR. Section 4 elaborates on the color + depth 3D user representation, including performance measures (4.1) and a demo description (4.2). The paper ends with a conclusion and major challenges of Social VR in Section 5.

Figure 1: Three-person experience: Virtual Reality design (left), VR User View (middle) and Setup diagram (right).

2 RELATED WORK

In the 90s, a lot of work went into creating high-end shared virtual environments. Various universities set up cave automatic virtual environments (CAVE) and CAVE-like systems which could be used to communicate remotely, of which [5] and [8] are good examples (same authors), using what they call "video avatars". These environments typically used back projection and large calibrated camera rigs to produce a coherent virtual environment. Other examples such as [6] use large screens, again together with calibrated camera rigs, to also offer a sense of togetherness. Other work from this era consists of using graphical avatars to create large shared virtual environments, of which [1], [7] and [11] give some overview. The impact of avatar realism was studied as well [2].

Virtual reality saw renewed interest with the rise of high-quality but affordable HMDs, of which the Oculus Development Kit, which carried the promise of bringing high-quality VR to the masses. This has led to new initiatives in shared and social VR experiences as well. Nowadays, social VR is mostly associated with graphical avatars in a graphical environment. Main examples are currently Facebook Spaces[1], AltSpaceVR (recently bought by Microsoft)[2], BigScreen[3] and SteamVR[4]. All these consist of a shared environment, using graphical avatars to represent the users, and offering various things to do, such as sharing games, screens and shared web browsing, including shared video watching, shared exploring, etc. The transfer of user motion to avatar motion is achieved by employing HMD and controller tracking. These environments offer a compelling experience of togetherness, in which immersive video is combined with spatial and 3D audio. Still, they are limited in the way users can interact and communicate with each other. For example in the user study of [9] a user of Facebook spaces stated: *"The social cues that you would normally have about someone .. weren't there"*.

In our preliminary work [3, 4, 10] our focus was on using video streaming and web technologies as a basis to bring people together in virtual environments. We have shown that shared and social VR experiences can be created by using current of-the-shelf equipment and by using a WebVR-based framework. Our previous work however, is limited to connecting two users together in a a static 360-degree virtual environment with 2D video streams to represent users. In this work, we demonstrate two improvements, i) adding a third user into our system and ii) utilizing a 3D virtual environment with a depth-based point cloud user representation.

[1]https://www.facebook.com/spaces
[2]https://altvr.com/
[3]http://bigscreenvr.com/
[4]https://steamcommunity.com/steamvr

3 MULTI-USER VR COMMUNICATION

In this section we present a three-people experience based on our TogetherVR framework [3, 4, 10]. TogetherVR is a completely web-based framework to build and consume shared and social VR experiences. Our main motivation to utilize web technology is to allow an easy and widespread deployment and low entry burden for end users and developers. TogetherVR is utilizing many established web frameworks (like WebVR, AFrame, three.js, WebRTC, WebAudioAPI, WebGL, soket.io, Angular, Dash) to create virtual spaces and to enable users to communicate in such spaces. To allow communication via audio and video via our system, we alpha-blend people into the environment based on WebGL shaders. In our system users are recorded with a Kinect 2 RGB-plus-depth camera, and the background of the users is replaced with a uniform green color before transmission (see Figure 2, bottom left). After receiving the video, the background is removed via alpha-blending in the receiving browser, thus leaving a transparent image showing just the user without his/her physical background. For capture and transmission we use a resolution of 960x540 pixels with 25fps.

In this three-people experience, three people sit around a round table in VR. The view of each user is the same; that is, each user sees the other two users on the opposite side of the table and a video playing on the top of the table (see Figure 1). In our setup (Figure 1, right), we have three specific and similar user places, each with a laptop (MSI GT62VR), Oculus Rift HMD (CV1), Kinect camera, headset (Pioneer SE-M631TV) and unidirectional microphone (Power Dynamics PDT3). Users are recorded from the front so that they are able to see and interact with two people at the same time. We held a 1-day experiment trial in an informal and uncontrolled setting at our lab facilities. In this experiment, we collected feedback through a short questionnaire from 54 participants (avg. age of 32,5 and 42% Female), who communicated and watched a video with our system.

We evaluated the quality of the (visual) experience and the sense of (co-)presence, asking participants to rank their experience on a 5-point Likert-type scale. We performed a Cronbach's Alpha test to check the consistency of our questionnaire items (i.e. whether the scale used for our questionnaire would yield reliable measurements). The test reveals an overall good consistency, all questions show

Figure 2: Different Video Capture modes A (top left) full color, B (bottom left) green-screen cut-out and C (right) color cut-out (top) with depth (bottom).

a high alpha value (avg. of >0.7). People appreciated the overall quality of the experience (98% of the participants scored 4 or higher, avg. 4.3, SD 0.51). They also felt involved in the virtual environment (experience, avg. 4.1, SD 0.76). Further, they reported a sense of being in the VR scene (presence, avg. 4.2, SD 0.80) and the sense of being with the other users (co-presence, avg. 4.1, SD 0.86).

Within the experience we observed more interaction between participants compared to our previous co-watching experience [3, 4], with 2 users sitting beside each other, watching a movie. This is not only because more people were involved; while seeing people from the front as well as seeing the video playback and the other people at the same time, people expressed a more natural conversation setting. Further, no participant reported on any experience of motion sickness. However, participants reported not seeing themselves (self-view) as well as not being able to stand up or get close to the camera as diminishing factor for the experience.

4 3D USER REPRESENTATION

In this section, we present our ongoing effort to include new media formats. This is, using RGB+depth (RGBD) information, to display users as a 3D point cloud. To do so, we mainly changed 2 components from the three-person experience (see Section 3): (A) the Kinect v2 camera capture to record both color and depth, and (B) the WebGL shader to display the user video. Regarding the camera capture (A), we combined the color image with the depth image (see Figure 2, right). However, as the browser and current WebRTC implementation does not support depth encoding we need to use an intermediate step. We map the 16 bit depth value from the Kinect into the RGB color space using the green and red color only. We use green and red because these colors are prioritized in an YUV color mapping. An example of such a mapping can be seen in Figure 2 (bottom right). To display this image as a point cloud (B) we changed the WebGL shader, in order to not display a flat image, but each pixel with coordinates in the 3D space, based on the depth image. Our shader is based on the work by George MacKerron[5].

4.1 Performance

We run performance tests, to evaluate our depth based point cloud approach on a technical level. This is to compare the resource usage under different video transmission to better understand the strict performance requirements and limitations of our web-based solution. In this analysis, we compared 3 conditions (see Figure 2 and Table 1), transmitting the full color image (top left), the image with a replaced green background (bottom left) and the image with color + depth information (right). We ran 2 computers (measuring CPU, GPU and bandwidth usage), one with a Kinect sensor to capture and send the video over WebRTC and one to receive and display the video. Additionally, we encoded 10 sec clips in all three conditions with FFmpeg (VP8) to show the differences in encoding sizes of the three approaches. Table 1 shows the measurements of our tests. The differences in CPU/GPU usage are minimal in all 3 cases, showing only little overhead for the color+depth video. Further, the bandwidth used is the same for all conditions (in relation to the resolution), which is based on the automatic WebRTC settings. Looking at the encoding size of the 3 clips, the differences are more complicated. Particularly, that condition B (cut-out) is larger than condition A (full color) seems counter intuitive, as a large part of B consist of a uniform color. In this regard, it is good to mention that most encoders (i.e. h264 and VP8) are not optimized for such video files, but for color scenes. Thus, this result is not surprising.

[5] http://blog.mackerron.com/2012/02/03/depthcam-webkinect/

Table 1: Performance of different video encoding, decoding and rendering.

Performance matrix	Encoding	Decoding	Bandwidth	10 sec P8 file
A: RGB Full Colour (540x960)	34%CPU 15%GPU	16%CPU 30%GPU	2.7Mbps	356KB
B: Greenscreen cut-out (540x960)	40%CPU 16%GPU	17%CPU 30%GPU	2.7Mbps	449KB
C: Cut-out + depth (1080x960)	38%CPU 24%GPU	24%CPU 30%GPU	5.2Mbps	466KB

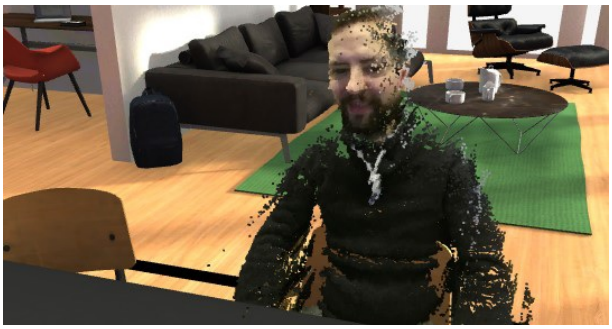

Figure 3: 3D room with RGBD based point cloud.

4.2 Demo Experience

To show the difference between the 2 approaches (360-degree static background with 2D user representations and a 3D scene with point cloud user representations), we integrated our depth based approach with a 3D living room (see Figure 3) into our system. In this demo experience, we can exchange between both the 3D environment and the 360-degree environment in order for people to compare both cases. Furthermore, it allows us to showcase that also more complex 3D VR applications can easily be enhanced with photo-realistic user communication in web-based applications.

5 CONCLUSION & FUTURE WORK

In this demo we present both a 3-user Social VR experience as well as a browser-based 2 user experience where users are represented as point clouds, based on color+depths 2D video streams. Even though we show that such experiences are technically possibly with little performance overhead in comparison to the static 2D video case, there are still many technical and user related work to be done in the future. Based on our current findings we see the following problems as the biggest challenges for social VR applications, which should be addressed in the future:

Being able to see yourself. Currently participants cannot see their own body. A common approach taken in VR is rendering virtual arms that follow the movement of the real arms movements. We are investigating to what extent we can create a self-view which allows users to see (photo-realistically) their own bodies.

High immersion and presence for communication. Having photo-realistic representations of other people in our system has helped provide high immersion and presence so far, however in our current approach, the HMD is still visible to the users. In the future, we intend to apply HMD removal techniques.

Being able to interact and engage in VR with more users simultaneously. In this paper, we presented social VR experiences for up to three users. In the future, we will look at accommodating more users at the same time, with all the associated communication challenges, both in terms of network and user interactions.

First timer effect and real-world use cases. In our current tests, people often only spent a short time in VR, and sometimes experience VR or social VR for the first time. In the future, we plan to do user evaluations in a more controlled setting, and incorporate use cases from the real world where users follow a specific task for a longer period of time, for example half an hour to an hour. We also plan to have a group of users use it on several occasions, to see if repeated experiences would change their perception of our social VR environment. We are particularly interested in topics related to education, collaboration and live events (e.g. sports). With this we like to investigate the benefits of VR in general and social VR, together with industrial partners. Importantly, we like to compare avatar-based vs. photo-realistic social VR in different use cases.

ACKNOWLEDGMENTS

This paper was partly funded by the European Commission as part of the H2020 program, under the grant agreement 762111 (VRTogether, http://vrtogether.eu/).

REFERENCES

[1] Steve Benford, Chris Greenhalgh, Tom Rodden, and James Pycock. 2001. Collaborative virtual environments. *Commun. ACM* 44, 7 (2001), 79–85.
[2] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. 2003. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 529–536.
[3] Simon Gunkel, Martin Prins, Hans Stokking, and Omar Niamut. 2017. WebVR meets WebRTC: Towards 360-degree social VR experiences. In *Virtual Reality (VR), 2017 IEEE*. IEEE, 457–458.
[4] Simon NB Gunkel, Martin Prins, Hans Stokking, and Omar Niamut. 2017. Social VR Platform: Building 360-degree Shared VR Spaces. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 83–84.
[5] Michitaka Hirose, Tetsuro Ogi, and Toshio Yamada. 1999. Integrating live video for immersive environments. *IEEE MultiMedia* 6, 3 (1999), 14–22.
[6] Peter Kauff and Oliver Schreer. 2002. An immersive 3D video-conferencing system using shared virtual team user environments. In *Proceedings of the 4th international conference on Collaborative virtual environments*. ACM, 105–112.
[7] Jason Leigh, Andrew E Johnson, Thomas A DeFanti, Maxine Brown, M Dastagir Ali, Stuart Bailey, Andy Banerjee, P Benerjee, Jim Chen, Kevin Curry, et al. 1999. A review of tele-immersive applications in the CAVE research network. In *Virtual Reality, 1999. Proceedings., IEEE*. IEEE, 180–187.
[8] Tetsuro Ogi, Toshio Yamada, Ken Tamagawa, Makoto Kano, and Michitaka Hirose. 2001. Immersive telecommunication using stereo video avatar. In *Virtual Reality, 2001. Proceedings. IEEE*. IEEE, 45–51.
[9] J. Outlaw and B. Duckles. 2017. Why Woman Don't Like Social Virtual Reality. https://extendedmind.io/social-vr
[10] MJ Prins, S Gunkel, and OA Niamut. 2017. TOGETHERVR: A FRAMEWORK FOR PHOTO-REALISTIC SHARED MEDIA EXPERIENCES IN 360-DEGREE VR. In *Technical Papers International Broadcasting Convention (IBC)*.
[11] Ralph Schroeder. 2012. *The social life of avatars: Presence and interaction in shared virtual environments*. Springer Science & Business Media.