Computers & Graphics (2019)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

New Interactive Strategies for Virtual Reality Streaming in Degraded Context of Use

Lucile Sassatelli[a,*], Marco Winckler[a], Thomas Fisichella[a], Antoine Dezarnaud[a], Julien Lemaire[a], Ramon Aparicio-Pardo[a], Daniela Trevisan[b]

[a] *Université Cote d'Azur, CNRS, I3S, 06900 Sophia Antipolis, France*
[b] *Universidade Federal Fluminense, Niteroi-RJ 24210-346, Brazil*

ARTICLE INFO

Article history:
Received May 2nd, 2019
Accepted Sep. 25, 2019

Keywords: 360 degrees video streaming, virtual reality, interactive systems, degraded context of use, limited bandwidth

ABSTRACT

Virtual reality videos are an important element in the range of immersive contents as they open new perspectives for story-telling, journalism or education. Accessing these immersive contents through Internet streaming is however much more difficult owing to the required data rates much higher than for regular videos. While current streaming strategies rely on video compression, in this paper we investigate a radically new stance: we posit that degrading the visual quality is not the only choice to reduce the required data rate, and not necessarily the best. Instead, we propose two new impairments, Virtual Walls (VWs) and Slow Downs (SDs), that change the way the user can interact with the 360 degree video in an environment with insufficient available bandwidth. User experiments with a double-stimulus approach show that, when triggered in proper time periods, these impairments are better perceived than visual quality degradation from video compression. We confirm with network simulations the usefulness of these new types of impairments: incorporated into a FoV-based adaptation, they can enable reduction in stalls and startup delay, and increase quality in FoV, even in the presence of substantial playback buffers.

1. Introduction

Contents and equipment's for Virtual Reality (VR) have been developing fast in the last couple of years, both from a technological and commercial point of view. The technology is benefiting from major progresses in VR headset design (such as the announced Google-LG new 18-megapixel display) and compression [32]. From a business perspective the sales of VR headsets are foreseen to reach a yearly 40 million in 2022 and the market $215B [9]. With games and AR applications, cinematic contents and 360° videos in particular are important elements in the range of immersive contents. These are spherical videos which are meant to be watched in a VR headset for the user to get immersed into the content's world. They open new perspectives for story-telling, journalism or education.

As it is currently the case for regular videos, their preferred mode of consumption will remain Internet streaming. However, a major obstacle to stream 360° videos is their required data rate, or bandwidth. Owing to the distance between the user's eye and the screen when wearing a VR headset, the data rate must be two orders of magnitude higher than that of 4K videos. Given the resolution of the human fovea, a full impression of reality from the sight would require 5Gpbs, even with the latest H.265 video coding standard [4]. These data rates are not available in standard Internet accesses, and the network challenges entailed by massive distribution of immersive content are substantial.

*Corresponding author:
e-mail: lucile.sassatelli@univ-cotedazur.fr (Lucile Sassatelli)

A major question therefore arises: how to stream immersive content under limited bandwidth? This article contributes in this direction. The general principle in existing research is to send in high quality (i.e., with high encoding rates) the sector of the video the user faces, and the rest in lower quality. This therefore makes the transmission decisions dependent on the user's behavior in the virtual environment. Deciding which part of the sphere to send in high quality from the remote streaming server hence requires to predict the future user's Field of View (FoV). Such prediction is only partly possible over very short time horizons (order of a second or less) owing to the complex dependency on previous motion and content, and inherent randomness [33]. For a given constrained bandwidth, the greater the discrepancy between the bandwidth and the highest video rate, the narrower the sector sent in highest quality, and the greater the probability the user will face a low quality sector.

This article investigates a radically new stance on the problem: assuming that the goal of an immersive experience is to make the user feel as in a real-world thanks to the sight, and given the impact of visual degradation on the vestibular system (as compared with watching a regular screen) and the feeling of presence, we posit that degrading the visual quality is not the only way to reduce the required data rate, and not necessarily the best choice. Based on the knowledge of the human attentional process, we identify new dimensions in which to impair the content to absorb the lack of bandwidth, complementarily to the visual quality. Specifically, we design two types of impairments and show that, when triggered in proper time periods, they can be better perceived than visual quality degradation from video compression, for the same amount of data to transfer.

Contributions:

• **We introduce two new types of impairments, named Virtual Walls (VWs) and Slow Downs (SDs), to improve the experience of 360° video streaming under limited bandwidth.** We implement them in a streaming player compliant with the Spatial Relationship Description (SRD) amendment to the MPEG-DASH (Dynamic Adaptive Streaming over HTTP) standard for 360° video streaming.

• **We carry out user experiments with 18 users and 11 video scenes to identify whether VWs and SDs are alternative impairments acceptable to the users and that can improve the level of experience compared with quality adaptation alone**. We use a double-stimulus approach to have every VW and SD versions compared with a reference version (both versions consume the exact same data rate). The video content represents different categories and comes from reference datasets.

• **The results show that both VW and SD impairments are generally preferred by the users over the compression-only reference**. A thorough analysis of quantitative subjective assessments and objective metrics (head motion collected from logs) enables to understand the important factors involved in the user's preference. Standardized SUS and AttrakDiff questionnaires confirm the acceptability of our approach.

• Finally, we assess the gain in streaming performance VW and SD can bring to different FoV-based adaptation logics more or less prioritizing buffering over responsiveness to head motion. **We confirm with network simulations the usefulness of these new types of impairments: incorporated into a FoV-based adaptation, they can enable reduction in stalls and startup delay, and increase quality in FoV, even in the presence of substantial playback buffers.**

In our concern for reproducibility, the code made and the user experimental data collected for this work are made publicly available at [40, 39].

The article is organized as follows. Sec. 2 presents related works. Sec. 3 introduces and motivates the proposed impairments. Sec. 4 details the experimental protocol. Sec. 5 analyzes the results of the user experiments, while Sec. 6 presents network simulations. Finally, we discuss some of the questions raised by our approaches, including important perspectives, in Sec. 7, and give conclusions in Sec. 8.

2. Related works

We review below four core aspects for our goal: the main recent findings on attentional behavior in VR, then the general classes of attention guidance techniques, the perception of slow motion and finally how the problem of streaming VR has been tackled so far.

Sitzmann et al. in [44] provide an extensive study (involving 169 users) of how do people explore in static VR environment (i.e., 360° images). They show that the average exploration time, that is the time a user takes to scan the entire 360°-wide longitude span is 19 seconds. They also show how this time depends on the actual content of the static scene: the exploration time decreases with lower entropy of the saliency map. Saliency is a well explored question in the domains of human attention and computer vision. It has been identified (see, e.g., [10]) that humans first register low-level features (such as edges and motions) then get attracted by higher-level, or semantic, features of the content (such as human faces, cars or animals). The saliency map of an image is the two-dimensional probability distribution of the user's gaze direction. A lower entropy of the saliency means a lower number of well-isolated Regions of Interest (RoIs). In such cases, the user's attention gets attracted to the few salient regions faster than 19s. In David et al. in [13], the authors present a dataset of 19 360° videos of 20 seconds, along with the head and eye gaze recording of 57 participants. In this article we use videos from this dataset they make open to the community. The authors show that the exploration phase in their videos last between 5 and 10s.

One purpose of studying the user attention is to devise efficient attention guiding techniques, that can be as diverse as using subtle visual cues [21, 50], editing and rotation [42, 41, 36] or haptic signals [25]. The references we mention and describe below are only representative works in these classes. In [21, 50], the higher sensitivity of the peripheral vision to high-frequency flickers is leveraged to trigger signals enticing the user to turn her head in the flickers' direction. In [21], the flickering luminance and frequency decrease when the distance between the user's gaze and the desired region decreases, thereby remaining as unconscious as possible. In [42], Serrano et al. investigate different movie editing for cinematic VR. Indeed,

legacy video editing, meant to drive the user's attention, cannot be readily applied in VR as the user controls the camera, which in turn directly connects with her vestibular system. The impact on the user's attention of RoI alignment between successive scenes is specifically studied. Serrano et al. uncover that the time to settle down on a RoI after a scene edit varies exponentially with the angle of misalignment between the RoIs before and after the edit. In [41], the authors introduce so-called snap-changes, where the user's FoV gets re-positioned, in a snap, in front of the RoI desired by the director. These edits are perceived as fast cuts and proven not to disturb the vestibular system, as no intermediate motion is perceived, contrary to, e.g., [23]. Finally in [25], vibro-tactile signals emitted by vibrators embedded in a headband are used to help the user localize the direction of the target they are supposed to track.

The impact on perception of slowing down periods in a video, often referred to as a *slow-motion* effect, has been studied in several works. For example, Mather et al. in [31] identify that the perception of what is considered a normal speed can be altered after watching only 30s of a slowed down or accelerated scene, compared with normal playback speed. In [8], Caruso et al. show that slow motion replay increases the impression that the action was intentional.

Video slow down has also been considered as an adaptation lever for Internet streaming. Representative works on the topic are [18] and [28]. In both [18, 28], the video playout rate is varied to better absorb network variation without requiring a large playback buffer, that increases delay to fill up before starting playing. The audio track is slowed down accordingly by using the signal processing technique of the speech signal named Waveform-Similarity-based Synchronized Overlap-Add (WSOLA) [48]. This technique preserves the pitch and experiments have shown that, with slow-down or speed factors in the range 35%-230%, the impact is "inaudible" or "not annoying". In our work, we make the hypothesis that SD can be beneficial to perception in VR, and can hence be harnessed to help streaming too. However, to avoid any problem with audio, we envision using a voice activity detection technique, such as [53], to prevent using SD when someone is distinctively talking.

The general principle to cope with insufficient bandwidth to stream VR content is to lower the required data rate by making transmission decisions based on the user's FoV. The way the sphere is split into low-quality and high-quality zones has been the subject of numerous works (e.g., [37, 51]), one standard extension from adaptive video streaming to 360° (MPEG DASH-SRD [34]) and a newly released standard, MPEG Omnidirectional Media Application Format (OMAF) [32]. Existing approaches consider compression only and their efficacy depends on the correct prediction of the future FoV. In [1], the authors propose a taxonomy of 360° content by analyzing the distribution of the head positions obtained from 32 users on 30 videos of average duration 3 minutes. From these findings, the authors identify video classes on which the head position prediction task is made easier, and how the streaming algorithms can consequently be adapted. In this article, we take advantage of their classification to design new impairments to give more flexibility to the streaming algorithm. We then verify our hypotheses on their impact on the user's perception using representatives of these video categories from [1]. In [12], Dambra et al. first show that alignment of RoIs between two scenes enables lowering the user's head motion (corroborating the findings of [42]), which in turn increases the efficiency of streaming algorithms to deliver high qualities in the FoV. Importantly, they also show how film editing can be designed to better predict the head position, thereby easing streaming by consuming less bandwidth for the same level of quality in the FoV. This shows that there are more dimensions to the VR experience than the visual quality only. In the present article, we build on these last three works to devise new ways to impair the VR content in degraded environment, by leveraging how the user is susceptible to watch the content in different periods. We are able to show that more flexibility can be given to the streaming algorithm beyond adapting the compression rate only.

3. New types of impairments: VW and SD

We first present elements on the phases of the human attention when watching a 360° video, before introducing the new types of impairments we propose, each aimed at being used in one of the phases.

3.1. Background on attentional process

It has been recently shown in [44] and [1] that, when presented with a new VR scene[1], a human first goes through an exploratory phase that lasts about 10 to 20s ([1, Fig. 18], [44, Fig. 2]), before settling down on RoIs. The duration and amplitude of exploration, as well as the intensity of RoI fixation, depend on the video content itself. Almquist et al. have identified the following main video categories for which they could discriminate significantly different users' behaviors: *exploration, static focus, moving focus* and *rides*. In *exploration* videos, the spatial distribution of the users' head positions tend to be more widespread. For that reason, the homogeneous content (absence or high number of RoIs) in *exploration* videos hardly allows to predict where the users will watch and possibly focus on. *Static focus videos* are made of a single salient object (e.g., a standing-still person), making the task of predicting where the user will watch easy: an angular sector can be identified, and will remain the same over time. *Moving focus* and *Static focus* videos are similar in that the RoIs are easily identified and hence the head positions are easier to predict than in *exploration* videos. However in *moving focus* videos, contrary to *static focus* videos, the RoIs move over the sphere and hence the angular sector where the FoV will be positioned changes over time. *Rides* videos are characterized by substantial camera motion, the attracting angular sector hence remaining that of the direction of the camera motion.

[1] Hereafter, we use the term "scene" as defined by Magliano and Zacks in [30] as a period of the video between two edits with space discontinuity. In our experiments, we simply stitch atomic videos made of a single scene [1, 13], and call them "scenes".

3.2. Slow Down

During the exploration phase occurring at the beginning of every scene, the head position can hardly be predicted and hence it is difficult to ensure high quality in the FoV at any point in time. The same goes for Exploration-type scenes. Our hypothesis is that in such phases or for such scenes, the users need time to apprehend the new world they get into. Hence, providing spatially-homogeneous high quality (over the entire sphere) at the cost of slowing down the video, may improve how the users perceive the content.

Definition: A Slow Down (SD) reduces the video playback speed.

We hypothesize they can be useful in the first 5 to 10 seconds of a new scene. The idea of SD came from previous VR experiments where users would have appreciated more time to get their bearings before too much action unfolds, hence suggesting that temporarily slowing down a scene might help to better immerse into the environment. It is also corroborated by Fearghail et al. in [16] who found that longer shots were less disorientating because the user had enough time to explore the environment at their own pace. By lowering the playback speed, a SD gives more time for the playback buffer to fill with high quality video segments, thereby preserving the visual quality and improving the impression of reality. A SD might go unnoticed in scenes such as a landscape or a street with moving cars if the user does not know the normal-speed video, while it may be easily detected by the users in scenes with distinguishable humans or animal walks. Also, the soundtrack should not be merely slowed down, but may instead be looped over or replaced for the duration of the SD. Our goal is to identify whether SDs can improve the level of experience compared with the corresponding video period played back at normal speed but with a lower encoding rate (lower quality). Therefore, in this proof of concept, we set to slow down exploratory scenes and shut down the sound entirely in both the reference and SD versions of the videos the users compare.

Implementation: SD has been implemented by modifying a 360° video streaming player for Android named TOUCAN-VR and released in [11]. Our new branch is available at [40]. This player builds on the Exoplayer media player, which allows to modify the playback rate for SD through the `setPlaybackParameters` method of a `player` object. For each video we make an XML file specifying the start time, end time and slowing factor for each SD period. This XML file is parsed by the 360° video player to then enforce the SDs during the video playback.

3.3. Virtual Walls

The rule-of-thumb so far (see, e.g., [20, Sec. 3] or the Oculus Rift developer guidelines [35, p. 5]) has been to always send something for the user to watch in any part of the sphere. In this article, we hypothesize that it is possible to restrict temporarily the angular sector the user can access, in order to save sending some part of the sphere and being able to increase the quality in the accessible sector. We therefore investigate if and how, after the exploration phase, in *static focus* and *ride* scenes, placing a so-called Virtual Wall can improve the Quality of Experience (QoE) compared with a higher compression factor (lower quality).

Definition: We define a Virtual Wall (VW) as a restriction of the accessible angular sector.

There are several possible ways to restrict the visible part of the sphere to save transmitting part of the content. A simple way is to replace the non-transmitted sphere sectors with black patches. This is the approach taken by the VR180 camera of Google [19]. However, VR180 is not meant at helping streaming and is hence not adaptive, as the video is shot over 180° only, and hence the rest of the sphere is never accessible. In the "Lebron James" 360° video released by Felix & Paul studios, a 180° sector of the video has been replaced with a high-quality stereoscopic photography to increase quality. Here, we propose VW as a mechanism that can be triggered as required over time and the accessible angular sector is a parameter and does not need to be 180° only. Also, entering into a black area (i) may be unsettling as the user may loose her footing (as when one closes the eyes while standing up), and (ii) would be consciously perceived. Instead, we design the VW impairment as a subtle degradation of the user interaction with the content: when the longitude of the user's position reaches the limit of the visible sector, the FoV only refreshes in latitude until the user comes back in the visible sector. As only the longitude is affected, the user does not risk to lose her footing. Fig. 1 (e.g., Col. 2-3) shows the impact of the VW on the accessible FoV positions. VWs are positioned after the exploration phase in videos with concentrated saliency (*Static focus* and *Rides*), so that the probability that a user hits a VW by trying to look away from the interesting regions should be low. We hypothesize that a substantial fraction of users will not perceive the VW, and if they do, they will quickly learn that turning their head back to the previous position unblocks the system, and will avoid hitting a VW multiple times.

Implementation: As for SD, we modify the available code of the TOUCAN-VR Android player for 360° videos [11] to implement VWs. Our new branch is available at [40]. This Android player also builds on the Samsung Gear VR Framework. This framework enables to constantly collect the head position through calls to the `getCurrentYAngle` method on a `GVRSceneObject` object. We can then decide, based on where the head is, relatively to the VW angular limits, to arbitrarily set the displayed FoV with the `getTransform().setRotationByAxis` method called on the `GVRSceneObject` object. A custom `lastRotation` variable is used to track and control the difference between the actual coordinates of the head, and the coordinates of the displayed FoV. We describe in an XML file the periods and angular sectors where to position VWs, which is parsed by the player at the beginning of the playback to correctly enforce the VWs.

4. Hypothesis and experimental protocol

This section details the specific hypotheses we make on the VW and the SD impairments, as well as the evaluation of their overall usability and user experience. This evaluation is made

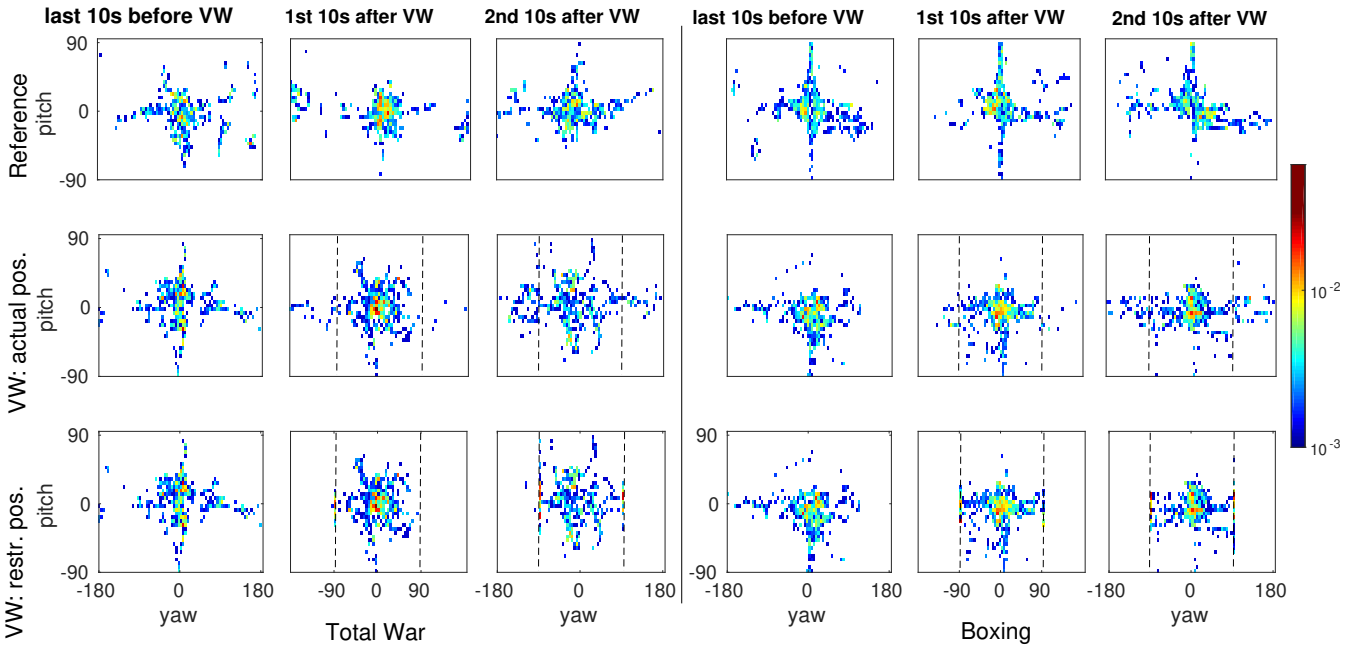

Fig. 1: Heat maps of most utilized yaw and pitch angles in 3 periods (columns). The visible sector of the VW is the 180° span between the vertical dotted lines.

using a double-stimulus approach following the guidelines of the International Telecommunications Union (ITU) [47]. We use standard and ad hoc questionnaires with specific metrics the users are asked to score. This evaluation is completed by the analysis of the head motion logs recorded during the experiments.

In our concern for reproducibility, the modifications we brought to the code of the 360° video player are available in [40] as a new branch of TOUCAN-VR [11], and the user experimental data collected for this work is made publicly available in [39].

4.1. Hypotheses to test

Slowing down the video is expected to be useful to the user experience in (i) exploration phase occurring at the beginning of a scene, and/or (ii) in scenes of Exploration type. Virtual Walls can be triggered in the focusing phase, i.e., after the exploration phase, in videos where a clear focus is likely to occur: these are videos of Ride and Static focus type. SD and VW are therefore designed as complementary to each other, additionally to be complementary impairments to a compression-only based approach (which only adapts the quality to the FoV).

We therefore position the VWs in scenes and periods where the users are much likely to focus on a known region. We position the SDs in the first few seconds (0-5s to 0-15s) of a new scene. In the user experiments we carry out, we refer to the video version with SD or VW as the *effect version*. Every *effect version* is compared with the corresponding *reference version* without SD nor VW. The specific hypotheses of the study are:

H1 Compared with a reference with no VW consuming the same amount of data, the version with VW is generally preferred.

H2 Compared with a reference with no SD consuming the same amount of data, the version with SD is generally preferred.

4.2. Design of experiment

The whole experiment was designed to be performed in 1 hour (it took in average 67 minutes). We use a double-stimulus approach as described in [47].

Video content:
The video scenes have been taken from two recent datasets [1, Table 1] and [13]. The videos and their features are detailed in Table 1. The videos named *Comb. Rides* and *Comb. Explo* are in fact compilations of 4 and 3 atomic scenes, respectively. We do so to reduce the duration of the experiment. All the scenes are freely available on Youtube with the IDs listed in [1, Table 1] or from the open dataset referenced in [13].

The VW impairment is tested on 5 scenes corresponding to two videos: *Comb. Rides*, which is made of four scenes classified as Rides in [1, Table 1], and *Boxing* made of one scene classified as Static focus in [1, Table 1]. The SD impairment is tested on 6 scenes corresponding to 4 videos: *Comb. Explo* made of 3 scenes classified as Exploration in [1, Table 1], and *Bar*, *Underwater* and *Touvet*, which can also be classified as Exploration videos. A substantial camera motion is present in 3 of the 6 scenes (Avenger, Bar and Touvet, as reflected by the Temporal perceptual Information score listed in [13, Table 1]). The purpose of distinguishing the last three short scenes as independent videos was to assess the impact of different SD durations.

In total: (i) VWs are tested on 5 scenes of both target video categories for this impairment (Rides and Static focus) with a total duration of 260s; (ii) SDs are tested on 6 scenes of category Exploration, with and without camera motion, with a total duration of 160s.

The exploration phase is estimated to 19s in [44], even though it seems significantly longer according to [1] that presents the dataset from which we take the scenes in Comb. Explo. Therefore, to be conservative, all VWs are positioned after the exploration phase, i.e., after about 20s of the start of the scene, and last a few tens of seconds until the end of the scene. We position SD between 0s-5s, 0s-10s and 0s-15s of the different scenes, as indicated in Table 1. The angular sector of the VWs is 180° in longitude (i.e., in yaw angle). No restriction is made on latitude, to maintain proper balance (i.e., in pitch angle). The slow down factor of SDs is set to 2 because it corresponds to the factor between the bitrates of the two different quality levels considered below, and it seemed reasonable when we experienced ourselves the video while preparing the user experiments.

Encoding rates and comparison fairness:
We do not consider real arbitrary network conditions for the user experiments, but instead emulate network conditions where we hypothesize the user's experience would benefit from VW or SD. Indeed, our goal in this article is to bring the proof of concept that breaking away from the sole quality adaptation is sensible from the user perspective. In real network conditions with arbitrary varying bandwidth, VW and SD would have to be finely tunable as quality levels can be, by dynamically choosing when they should be triggered, and with which parameters: angular sector for VW and slow down factor for SD. A full-fledged streaming adaptation logic would therefore decide on the qualities over time and space, and how to employ SD and VW to best adapt to the user's motion, to the current video scene and to the network conditions. This however requires to make SD and VW adaptive, which is the important perspective of this work, as discussed in Sec. 7.

We therefore make the experiments in a controlled environment, where the quality variations in the reference and the VW and SD versions are set fixed, but correspond to representative bandwidth scenarios where VW and SD are expected to be useful (i.e., in the startup phase where the user explores for SD, and in the focusing phase for VW).

Also, as our focus in this work is on the user's perspective in our double-stimulus approach (rather than adapting to arbitrary network conditions, left for future work), we consider two possible video qualities, the highest consuming twice as much data rate as the the lowest (in average, as Variable Length Coding is used).
Remark: We mention that considering only 2 quality levels is not unusual, as done for example in [24] and [2]. Also, the ratio is often 2 between two successive encoding rates in video streaming: see, e.g., the DASH manifest file available in [14, 49]. This is due to the logarithmic shape of the typical rate-distortion curves: to have a regularly-spaced quality set (using, e.g., VQM or SSIM as metric), the bitrates corresponding to the successive qualities are multiplied by a constant factor (often close to 2).
In the periods where a VW or a SD can be triggered, we consider the bandwidth to be sufficient to stream only the lowest quality over 360° at normal playback rate. The SD version therefore allows to display the highest quality over the entire 360° sphere during the SD period. The VW version allows to display the highest quality over the accessible sector (hence set to 180° here). The reference version plays at normal speed and is supposed to fetch the minimum data to have something to display over 360°, it therefore can only display the low quality. The impact of the different encodings on different videos can be seen in the screenshots of representative FoVs in Fig. 2. However, it is difficult to render the perceived level of compression artifacts in vignettes as viewed in a headset. Fig. 3 presents in a larger size two of these items. Finally, let us mention that the encoding difference is measured to bring a difference of about 35dB in Peal Signal to Noise Ratio (PSNR).

4.3. Discussion on handling sound

We have chosen to mute the sound in all the videos, for both the reference and effect versions, so as to avoid any interfering factor in the comparison. Indeed, Beerends and De Caluwe showed in [5] that the video quality heavily impacts the subjectively perceived audio quality in audiovisual stimulus. Though, reversely, the audio quality impacts the subjectively perceived video quality to a lesser extent, it impacts it nonetheless. That is why standard video quality assessment campaigns are made without sound. This also applies to 360° video quality evaluation, as recently reported by Singla et al. in [43, Sec. 4.1].

However, it is true that the handling of sound in SD is important. SDs correspond to slowing down the video playback, and are therefore not meant to be used when distinct voices can be heard. Instead, they are meant to be used in the exploration phase, that is at the start of a new scene, which often starts with background noise (or music). As presented in Sec. 2, speech signal processing techniques (such as WSOLA [48] and following, introduced in Sec. 2) exist to scale down the playback speed of the audio with that of the video if needed. Though theoretically applicable in an immersive environment such as 360° video playing, we can also envision a more conservative approach where, to make sure a SD is not triggered in a speech period, a reference system for Voice Activity Detection (VAD) shall be used. Deep Learning (in particular for home assistants) has brought significant improvements to the field, and methods such as [53] or [26] shall be employed to identify periods suitable for a SD.

Once such periods are identified and a SD can be triggered, a simple solution to make the soundtrack last for the extra time required by the SD, can then be looping on the soundtrack in the SD period's original duration. As background sound are often not listened to closely, we consider the event the user would change her preference for the more visually degraded quality is unlikely. However, confirming that is left for future work.

4.4. Participants

The user experiments were run between December 2018 and January 2019. We recruited 18 users using a convenience sample. Exact gender-balance was met. Two participants were above 40 years old, the others were between 20 and 30. Apart from 3 administrative staff, the other participants were undergraduate or graduate students. About 65% had already watched 360° videos in headsets, while less that 10% were used to play

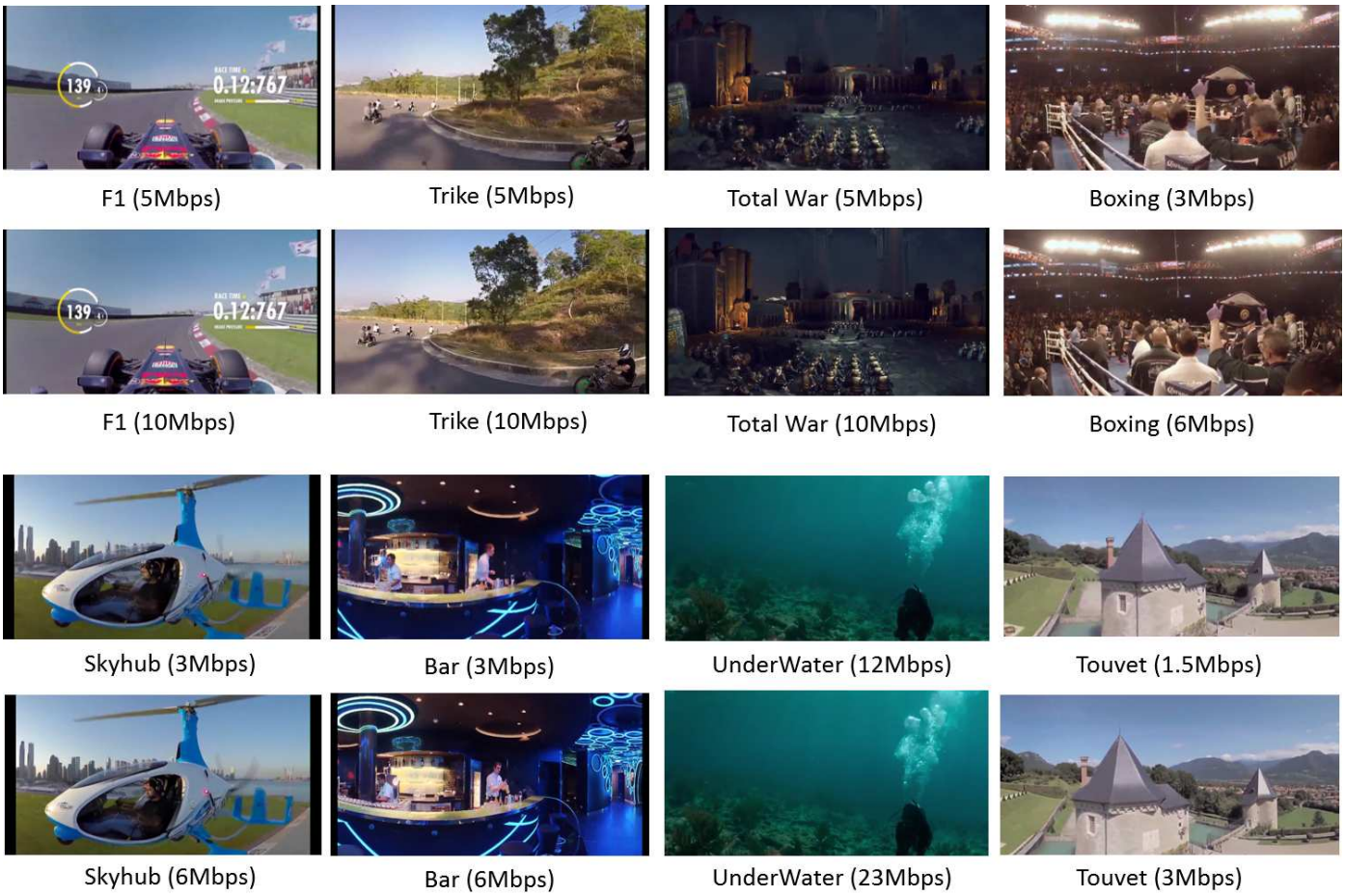

Fig. 2: Screenshots showing the impact of the different encodings on different videos for representative tiles. The upper (lower) rows are video scenes used to experiment with VW (resp. SD).

Video name	Scene, duration	Class	VW/SD (period)	Rate outside the VW/SD period (Mbps)	Rate in VW/SD period for ref. (Mbps)	Rate in VW/SD period for VW/SD version (Mbps)
Comb. Rides	F1, 31s	Ride	VW (18s-31s)	10	5	10
Comb. Rides	Trike, 51s	Ride	VW (25s-51s)	10	5	10
Comb. Rides	Assassin's Creed, 51s	Ride	VW (20s-46s)	10	5	10
Comb. Rides	Total War, 42s	Ride	VW (22s-42s)	10	5	10
Boxing	85s	Static focus	VW (25s-85s)	12	3	6
Comb. Explo	Zyed Road, 32s	Exploration	SD (0s-10s)	6	3	6
Comb. Explo	Skyhub, 35s	Exploration	SD (0s-10s)	6	3	6
Comb. Explo	Avenger, 32s	Exploration	SD (0s-10s)	12	6	12
Bar	20s	Exploration	SD (0s-5s)	6	3	6
Underwater	20s	Exploration	SD (0s-15s)	23	12	23
Touvet	20s	Exploration	SD (0s-10s)	3	1.5	3

Table 1: Description of videos (scenes from [1], classes, encoding rates) and new applied impairments.

VR games. They all had normal or corrected-to-normal vision. All participants gave written consent for participating in this study.

4.5. Equipment and setup

We used Samsung S7 Edge phones and the Samsung Gear VR headset. We used the 360° video streaming player available in [11] and compliant with the MPEG-DASH SRD standard enabling spatially-heterogeneous qualities depending on the user's FoV. VW and SD impairments have been coded within this Android application as a branch available at [40].

4.6. Procedure

The users were informed that the purpose of the experiment was to collect their preferences on a number of immersive videos. Before starting, the participants were informed that all information recorded during the sessions will be kept anonymous. They were also instructed on how to employ the *think aloud* protocol. After a video to get familiar with the gear and the virtual environment, they were shown each video in both versions, with and without SD/VW. The order of the videos was that of Table 1, while the order of the versions was picked randomly, established prior the experiments for all the user indexes by drawing a binomial random variable.

The users were hence presented back-to-back with the effect and reference versions of each video in a random order. Using a scale from 1 (the worst) to 5 (the best), users were asked to rate each version of the video w.r.t.: the *visual quality of the video*, the *perceived variation of video quality over time*, the *responsiveness of the system* to their head motion, their *comfort*, and the *time available to enjoy the video*. After seeing the two versions of the video, participants were asked to indicate which version they did prefer. As also considered in [44], the videos were watched standing up in order not to restrict motion, with the back of a chair in reach to keep balance if needed. After the series of pairwise viewing, we debriefed with the users. We asked them if they noticed the VW or SD impairments, and if they wanted to watch again some videos to see them. Participants were then asked to fill in the System Usability Scale (SUS) [6] and AttrakDiff [22] questionnaires.

4.7. Selected metrics

In addition to the verbal comments, the users' ratings of each above questions and the SUS and AttrakDiff questionnaires, we also augmented the logging threads of the player to collect objective measurements of the user's motion. In particular, we recorded the exact head position (yaw and pitch angles, or equivalently, longitude and latitude). When VWs were active, we also recorded the positions of the displayed FoV, as well as the actual head position, thereafter enabling us to compute the duration, instants and number of hits. All these metrics are analyzed next to understand how our new impairments were perceived and whether the hypotheses on their relevance to improve the experience are validated.

5. Results

We first analyze the results of the user experiments for VW, then for SD. We show in which extent they can confirm hypotheses H1 and H2. We analyze the importance of each factor (visual quality, responsiveness or comfort scores) in the expressed preference. The last part analyzes the results of the SUS and AttrakDiff questionnaires.

Fig. 3: Enlarged screenshots showing the impact of the different encodings for the F1 and Boxing scenes.

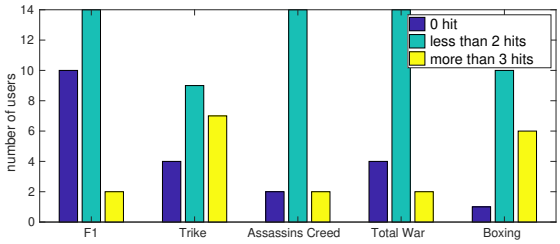

Fig. 4: Histogram of the number of users (total of 16) never hitting the wall, at most twice or more for each scene.

5.1. Results on Virtual Walls

Let us first present objective data on how the users interact with the VW by analyzing log data. Fig. 4 depicts the number of users hitting each wall never, less or more than twice. We observe that for every VW set in the different scenes, more than 50% the of users do not hit the VW more than twice. This percentage reaches 88% in all Ride scenes but one (Trike, which is the one with the least amount of camera motion). We therefore confirm that, when placed appropriately, the users seldom sense the VW, even more so in high camera motion rides.

Fig. 5.a depicts the fraction of users having declared to prefer, overall, the effect version over the reference version, for both videos featuring VWs. Let us remind that the Comb. Rides video is made of 4 Ride scenes, but the subjective comments are collected after seeing the video, not after each individual scene, for the sake of experiment duration. We observe that a majority of users tend to prefer the version with VW, over the reference version leaving the entire sphere accessible at the cost of lower quality: the VW version is preferred by 58% of the users in Rides (Comb. Rides) and 68% in Static focus (Boxing). The data is fitted to a Bernoulli distribution, whose 90% confidence interval on the probability parameter is represented on each bar (using the `fitdist` function of Matlab, which computes the confidence from the variance of the maximum likelihood estimator). The confidence intervals (returned by the fitdist Matlab function) are obtained from the variance of the maximum likelihood estimate of the probability parameter (see [45]). Owing to the width of the confidence intervals, we cannot formally conclude that H1 is confirmed. However, we can perform a breakdown analysis of the factors involved in the preference, which finally enables us to identify lines of improvement for the implementation choice of VW.

Comfort: First, Fig. 5.b shows the boxplots of the comfort score the users rated each version with (on a 1 to 5 scale, 5 being the best). Importantly, it shows that the users did not rate their overall comfort lower in the effect version than in the reference: this demonstrates that VW is acceptable.

Visual quality: Second, Fig. 6.a depicts the boxplots of visual quality scores given to each version. As expected, the visual quality score given to the VW version is significantly higer than that given to the reference version. More interestingly,

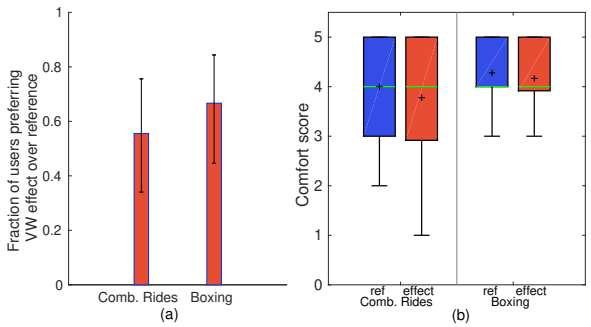

Fig. 5: (a) Fraction of users preferring VW over the reference for each video clip. (b) Comfort score. The greater, the better. The black cross marks the average, the green line the median.

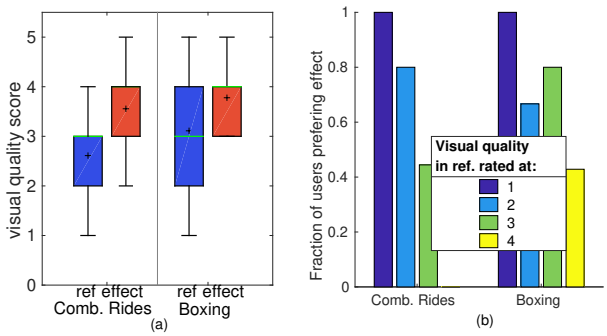

Fig. 6: (a) Visual quality score. The black cross marks the average, the green line the median. (b) Fraction of users preferring VW over reference conditionally to their visual quality score of the reference version.

Fig. 6.b shows the correlation between the preference and quality score, by plotting the fraction of users preferring the effect version given the quality score given to the reference. Visual quality therefore confirms to be a crucial parameter in the preference expressed by the user: it strongly correlates with the preference additionally to being consistently rated higher in the VW version.

Responsiveness to head motion: Third, Fig. 7.a reports the scores given by the users to the question "How much has the system been responsive to your head motion?". We expect it to reflect the consciousness the users got of VW. As expected, responsiveness to head motion is rated worse in the effect version with VW than in the reference version. In particular, the scoring difference between reference and effect is more pronounced for Boxing than for Comb. Rides. This may be simply explained by the significantly higher number of times users have hit the VW in Boxing than in Comb. Rides, as seen in Fig. 4. Then, Fig. 7.b depicts the correlation of the preference and the responsiveness. Two results are striking. First for the Boxing video, the correlation between preference and responsiveness is not clear, despite the responsiveness score between reference and effect higher for Boxing. For this video, the responsiveness is therefore not a crucial factor in preference. Second, for the Comb. Rides video however, despite the lower number of VW hits than in Boxing, likely reflecting in a lower score difference between reference and

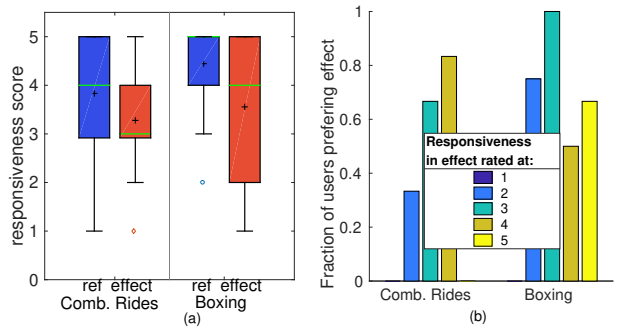

Fig. 7: (a) Responsiveness score. The black cross marks the average, the green line the median. (b) Fraction of users preferring VW over reference conditionally to the responsiveness score given to the effect version.

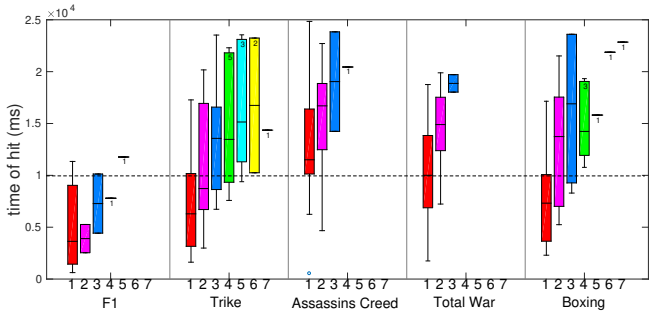

Fig. 8: Boxplots of hit times, in order of occurrence. The number of samples is indicated in each bar.

effect, the responsiveness score turns out to be highly correlated to the preference (as still seen in Fig. 7.b). The reason for this clear responsiveness-preference correlation in the Comb. Rides video can be found in the open comments made by the users. They reported that the camera motion worsens the feeling of the VW.

Lines of improvement for VW: This last observation suggests that the implementation of VW can be improved for scenes with camera-motion, for example by slowing down the playback as the user gets closer to the wall, and/or dimming the scene to alleviate the perception of motion. A second line of improvement can be identified by analyzing Fig. 8 depicting the time instants of the successive hits, in order in appearance. It is striking that generally only one hit occurs in the first 10 seconds of the VW period (below the dashed horizontal line). This phenomenon is also observed through the heatmaps in Fig. 1, where less discrepancy between head position and FoV position is visible in the first ten seconds of the wall. A simple guideline we can extract from that is to make a VW last no more that 10s uninterrupted if possible. This visible sector reduction over 10 seconds shall already give helpful slack to the streaming algorithm when the bandwidth is too low to stream high quality in the accessible area and low quality elsewhere.

Integration of the VW by the users: Finally, Fig. 9 represents the distribution of the duration of each wall hit against the index of the hit (the first till the seventh). It shows that the hit

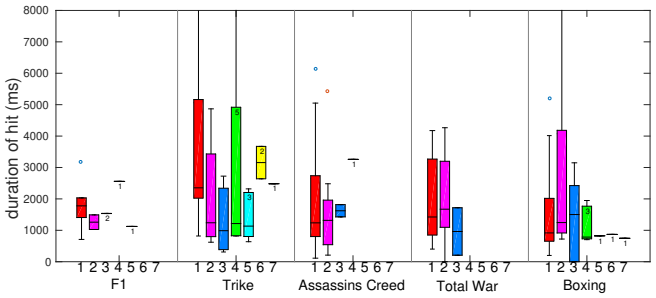

Fig. 9: Boxplots of hit duration for each ordered index of hit. The number of samples is indicated in each bar.

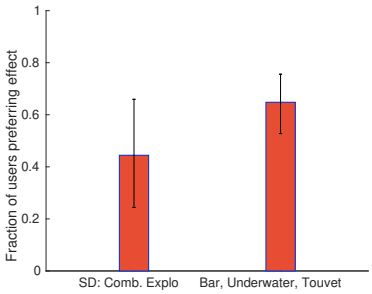

Fig. 10: Fraction of users preferring SD over the reference for each video clip. The samples for Bar, UnderWater and Touvet have been aggregated for the sake of the presentation.

duration generally decreases with the hit order. This would mean that once a wall has been hit (sensed), the subsequent hits are made shorter, if made at all. This uncovers a learning process from the user: once they understand there is a wall (lower score on responsiveness commented above), they tend to register it and avoid it.

5.2. *Results on Slow-Down*

Let us now analyze the results for SD. Fig. 10 plots the fraction of users having preferred the effect over the reference version. The data processing is the same as that for Fig. 5.a. The user's preference has been collected after each individual video has been watched in both versions, and we have gathered the results of the Bar, Underwater and Touvet video. This explains the lower confidence interval on the estimated fraction of users in the right hand-side bar. These preference results allow concluding that the users prefer the version with SD for the set of videos Bar, Underwater and Touvet. Below we provide an analysis of the factors impacting the user's preference in the case of SD. This analysis finally enables to conclude that for the scenes composing the Comb. Explo video, the SD has no impact on preference (there is an even split of the users' preferences between both versions but no degradation of the perception).

H2 is therefore clearly validated for the Bar, UnderWater and Touvet videos: SD, when appropriately positioned in Exploration phases, has the potential to improve the user's experience.

Visual quality: Fig. 11 shows the visual quality scores of each video in its effect and reference versions. First, for

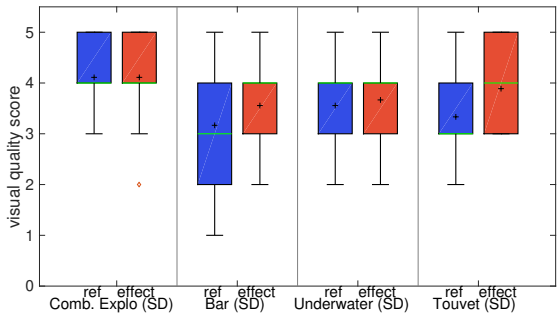

Fig. 11: Score in visual quality given to each video in the SD and reference version. The black cross marks the average, the green line the median.

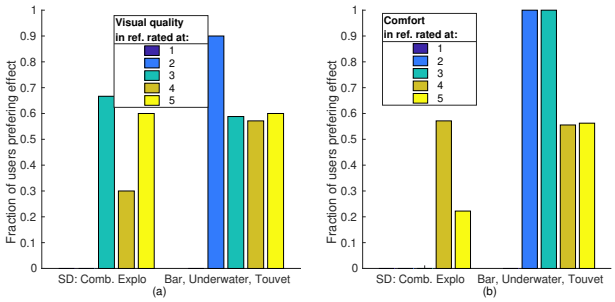

Fig. 12: (a) Fraction of users preferring SD over reference conditionally to the quality score given to the reference version. (b) Fraction of users preferring SD over reference conditionally to the comfort score given to the reference version.

Comb. Explo, we can see that there is no difference in the assessed quality of the reference and the SD version (in which the first 10 seconds of each scene, of duration of about 30s, are slowed-down by a factor of 2). The same happens for Underwater, which can be seen in Fig. 2. The Underwater scene video is a sea exploration in somewhat muddy waters, and is indeed characterized by a relatively low index of Spatial perceptual Information (SI), as defined in [46] and computed in [13, Table 1]. SI suggests that the level of scene detail is low and hence may not be impacted significantly by lower encoding rates. SI for Underwater is 42.1. Comparatively, Touvet and Bar have a much higher SI of 59.5 and 119.8, respectively, and the SD version obtains higher visual quality scores than the reference. Comb. Explo is made of 3 clips sharing the same characteristics as Underwater for which no quality difference appears in the scores: all 3 scenes have a low level of detail (Zyed Road is made exclusively of city lights, Skyhub is made of large parts of sea and grass, and Avenger is a computer-generated environment made of large parts of textures easily compressed). Let us now examine the correlation between the expressed preference and the visual quality score. Fig. 12.a shows that there is slight correlation between preference and perceived visual quality for the video group made of Bar, Underwater and Touvet.

Comfort: Fig. 13 presents the comfort ratings of the SD and reference versions. It is interesting to see that for one video, did SD bring a significant improvement to comfort: this video

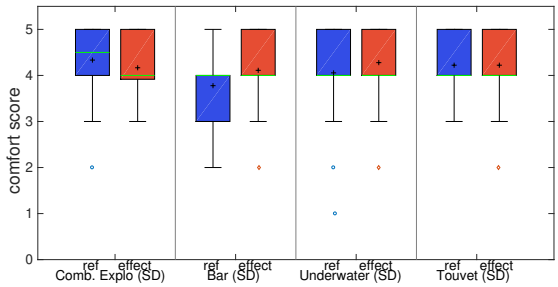

Fig. 13: Comfort scores of the SD and reference for each video testing SD. The black cross marks the average, the green line the median.

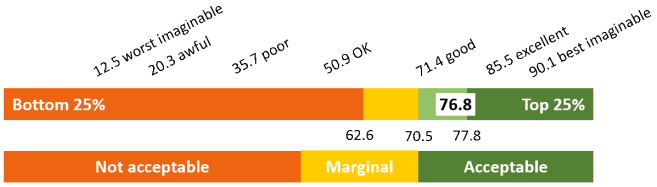

Fig. 14: Classification of the SUS final overall scores according to [3].

is that with the highest camera motion. It is Bar, in which the camera slides rapidly over the floor in a psychedelic night-club interior (see Fig. 3). Bar is indeed characterized by the highest SI and Temporal perceptual Information (TI) in the dataset [13, Table 1]. TI quantifies the level of visual changes over time in the scene. Bar has a SI of 119.8 and TI of 27.6. Comparatively, Underwater and Touvet have a (SI=42.1,TI=9.8) and (SI=59.5,TI=4.9), respectively. SD therefore reveals as a powerful tool to improve comfort in videos with substantial camera motion. Let us mention that this connects with the recent findings of Farmani et al. in [15]. They artificially reduce vection — the illusion of self-motion, which is connected to cybersickness — by snapping the viewpoint, reducing continuous viewpoint motion by skipping frames based on the speed of viewpoint rotation. Fig. 12.b confirms that there is correlation between the preference expressed by the users and the comfort score they have given. Finally, let us mention that the different SD durations set in Bar, UnderWater and Touvet did not yield significantly different perception from the users.

As VW, SD is therefore validated as a valid alternative impairment to visual quality, which enables improving the user's experience while consuming the same amount of bandwidth as the reference version.

5.3. SUS and AttrakDiff results

The results above stem from a double-stimulus approach crafted to specifically assess the proposed new types of impairments, VW and SD, in the envisioned context of use (degrading the content depending on the user's attentional behavior when the bandwidth is insufficient to stream high quality). Additionally to these specific results, we wanted to have an explicit and conscious feedback from the users on the relevance of our approach, in the form of standardized assessments. At the very end of each session with a user, we therefore motivated the rationale for our approach: whether or not they had perceived some VWs or SDs, we explained what they are meant for, and also proposed they experience them again, this time consciously. Considering this approach of possibly triggering VW or SD to preserve visual quality in VR videos as a product, we then questioned the user using the System Usability Scale (SUS) and AttrakDiff. The SUS provides a standardized and rapid assessment of the perceived usability of a system. AttrakDiff is another standardized method to assess interactive products, by specifically evaluating the pragmatic and hedonic qualities.

The overall SUS score we have computed from the users' answers reaches 76.8. As shown in Fig. 14, this score confirms a high level of usability as the minimum acceptable value starts in the second quartile and corresponds to 62,6. According to the classification of Bangor et al. [3], this SUS score of 76.8 can be qualified as "good". We have also extracted from the SUS the usability score (sum of the SUS items 1, 2, 3, 5, 6, 7, 8, and 9, multiplied by 3.125) and the learnability score (sum of the items 4 and 10, multiplied by 12.5), as defined by Lewis and Sauro [27]. We have thereby obtained an average usability score of 75.9 and a learnability score of 80.6. This high learnability score is coherent with the observations of users' comments during the test, and corroborate the analysis we made at the end of Sec. 5.1: the duration of head hits against the VW decreases over time, suggesting learning avoiding the VW.

The analysis of the user's experience using AttrakDiff [22] is depicted in Fig. 15. The results show high values for hedonic and pragmatic attractiveness, with some room for improvement in both dimensions. The portfolio diagram on the right-hand side in Fig. 15 shows the pragmatic and hedonic qualities of the system with a high confidence and reliability of the results. Fig. 15 on the left-hand side depicts the average values for the four dimensions of AttrakDiff. The pragmatic quality (PQ) is just average which means that it follows the standards. The hedonic quality is assessed in terms of identity (HQ-I) and stimulation (HQ-S). While the identity falls in the average region, the stimulation score seems to indicate that users are motivated by the experience. Regarding the hedonic quality, users seemed to be stimulated by the experience but there is room for improvement. This is in line with the very context of this work: none of the presented versions were perfect, as different types of impairments (aimed at consuming the same amount of data) were compared (we recall that most of users still prefer the effect versions). Finally, the overall impression of users (ATT) indicates the the solution is very attractive. Parsing the users' comments uttered (and noted down) when thinking aloud, we observe that many users expressed enjoying the experience, and in particular with the SD. The user N9 simply described the effect version with VW as *"...it is top!.."*. The SD received indeed most of positive comments from the participants and it was judged as *"...very good, it gives me more time for exploration..."* as reported by user N11.

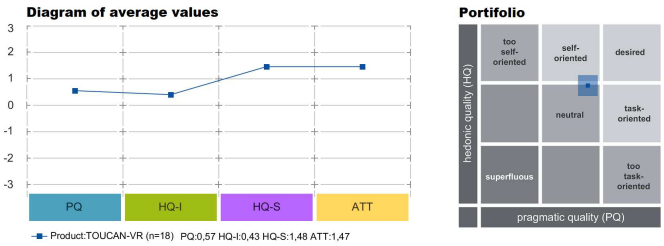

Fig. 15: Left: AttrakDiff dimensions. Right: AttrakDiff portfolio.

6. System-level impact of VW and SD

The previous section has shown results of the user experiments that were aimed at verifying whether VW and SD are alternative impairments acceptable to the users and that can improve the level of experience compared with compression alone. These alternative impairments are made to help support usage in limited bandwidth, and the experiments were made for typical scenarios where these impairments are envisioned to help (counter quality degradation in the startup exploration phase for SD or in the focusing phase for VW).

This section now aims at illustrating how much gain can VW and SD bring to application-level metrics (playback interruptions, startup delay, quality in FoV) when they enrich a reference FoV-based adaptation. FoV-based adaptation is the general class of streaming strategies that employ compression only: they decide which quality to fetch for each piece of the sphere and each time segment, based on the user's head position. Examples of such quality adaptations can be found in [38, 1]. One way, which we consider here, to enable to fetch heterogeneous spatial qualities is by tiling the sphere into so-called *tiles*, as defined by the SRD amendment to the MPEG-DASH adaptive bitrate standard for video streaming.

We simulate different streaming adaptation logics fed with head motion traces from the above user experiments. We show that, incorporated into a FoV-based adaptation, VW and SD enable reduction in stalls, startup delay and increase quality in FoV, even in the presence of substantial playback buffers. These results therefore show the potential of these complementary impairments to improve FoV-based strategies.

6.1. Problem modeling

We consider that the video is segmented in time and tiled in space, with J segments and M tiles, as described in Table 2 providing all system parameters. There is one buffer per tile. The bandwidth process is tracked by running an exponentially weighted moving average. The segments are fetched ahead of time with a look-ahead window of K. The streaming problem aims at selecting the quality l to download for each tile i of each segment j. The objective is to maximize the expected quality in the FoV, and the constraints are not to exceed the estimated bandwidth, to ensure some level of playback buffer for every tile (between B_{min} and B_{max} seconds), and to prioritize most urgent segments. An exact formulation of the problem as an Integer Linear Program is used to then derive heuristics. For conciseness, we only describe these heuristics here.

Parameter	Definition
M ($\mathcal{M}$), L ($\mathcal{L}$), J	number (set) of tiles, quality levels, segments
Δ	minimum period between 2 download decisions
K	look-ahead window in number of segments
q_l	quality rating of level l
$buf_i(t)$	num. of sec. stored in buffer of tile i at time t
$p_i(j)$	proba. that tile $i \in$ FoV at segment j
s_{ijl}	size (in B) of tile i of seg. j at quality level l
C_t	estimated bw. for download from t for dur. Δ
B_{min}, B_{max}	min and max buffer size
$\mathcal{W}_{per}$ ($\mathcal{W}_{angle}$)	set of seg. (tiles) in a VW period (visible sector)
$\mathcal{S}_{per}$ ($\mathcal{S}_{factor}$)	set of seg. in a SD period and slow factor
Decision var. $x_{ijl} \in \{0,1\}$	indicates whether tile i of seg. j is scheduled for download, $\forall i \in \mathcal{M}$, $j \in \mathcal{J}$, $l \in \mathcal{L}$

Table 2: Parameters and variable of the optimization problem

6.2. FoV-based strategies with SD, VW and competitors

Every heuristics discussed below is an iterative algorithm, such as in Algo. 1, whose rationale is to start from the highest quality for all the tiles and to decrease the quality as far away from the expected FoV as possible, as much as required, while prioritizing most urgent segments. This corresponds to a so-called pyramidal strategy, also described in [17]. Similar problem formulation and heuristics have been derived in numerous works, e.g., [1, Sec. 5.2], [38, 7]. The novelty here is the consideration of possible VWs and SDs in the adaptations. It is described for VW in Algo. 1 and named *Adaptation-VW*. Similarly for SD, it is named *Adaptation-SD* and can be described by (i) Algo. 1 removing reference to $\mathcal{W}_{angle}$ and (ii) scaling the playback duration of each segment j by the slow-down factor if $j \in \mathcal{S}_{per}$. The segments' durations are used to compute the buffer states $buf_i(t)$.

Let us detail how the probability of each tile i being watched in future segment j is estimated based on the current FoV at t, $FoV(t)$. Let $dist(FoV(t), i)$ denote the distance between the current FoV and the center of tile i. At time t, $p_i(j)$ is simply estimated with

$$p_i(j) = \frac{(\max_{i \in \mathcal{M}} dist(FoV(t), i)) - dist(FoV(t), i)}{\sum_i ((\max_{i \in \mathcal{M}} dist(FoV(t), i)) - dist(FoV(t), i))}.$$

Also, the parameters $j_i(t)$, $\forall i \in M$, denote the first segment index not in tile i's buffer, and $j(t) = \min_i j_i(t)$. We define $j_{i,min}$ so that $[j_i(t), j_{i,min}]$ is the minimum set of segments that must be downloaded from t to ensure that the buffer of tile i always has more than B_{min} seconds of playback stored.

For fair comparison we consider 2 references in each case of VW and SD: a reference adaptation unaware of the important regions (for VW) and instants (for VW and SD) where the quality should be increased to improve the probability to have high quality in the FoV (*Adaptation-unaware*), and another reference where it is aware and strives to match the quality decisions of the SD- or VW-enabled adaption (*Adaptation-aware-VW* and *Adaptation-aware-SD*). Hence, *Adaptation-unaware* is the same pyramidal strategy based on the current FoV but without any consideration of VW. It is hence described with Algo. 1 without the *if* statement in line 6 nor without reference to constraint with $\mathcal{W}_{per}$ and $\mathcal{W}_{angle}$ in line 1. *Adaptation-aware-VW* is meant to be less conservative by considering the knowledge of the VW position (i.e., of the highest saliency region), and

forcing to download high quality in this region. It is described with Algo. 1 without reference to constraint with $\mathcal{W}_{per}$ and $\mathcal{W}_{angle}$ in line 1.

Similarly, *Adaptation-SD* is compared with *Adaptation-unaware* and *Adaptation-aware-SD*. *Adaptation-aware-SD* is *Adaptation-unaware* except that we force it to take the same decisions as *Adaptation-SD* for segments $j \in \mathcal{S}_{per}$ (to force quality as high as that decided by *Adaptation-SD* though no segment is slowed down in *Adaptation-aware-SD*).

Algorithm 1: Streaming decisions with heuristic *Adaptation-VW*

Data: Buffer states $buf_i(t), \forall i \in \mathcal{M}$
Result: $\{x_{ijl}\}, \forall i \in \mathcal{M}, l \in \mathcal{L}, j = j(t), \ldots, j(t)+K$
1 For all i, j verifying constraints $buf_i(t) \geq B_{max}$ or $j < j_i(t)$ and verifying $j \in \mathcal{W}_{per}$ and $i \notin \mathcal{W}_{angle}$, allocate highest quality: $x_{ijL} = 1$;
2 Compute requested data: $data = \sum_{ijl} x_{ijl} s_{ijl}$;
3 $j = \min(j(t) + K - 1, J)$;
4 **while** $data > C_t\Delta$ *AND* $j \geq j(t)$ **do**
5 **for** i *in descending order of distance to FoV(t)* **do**
6 **if** $j \notin \mathcal{W}_{per}$ *OR* $i \notin \mathcal{W}_{angle}$ **then**
7 **if** $j > j_{i,min}$ **then**
8 decrease quality or cancel download if quality already minimum;
9 update *data*;
10 **if** $data \leq C_t\Delta$ **then**
11 break;
12
13
14 **else if** $j_i(t) \leq j \leq j_{i,min}$ **then**
15 decrease quality if not yet minimum;
16 update *data*;
17 **if** $data \leq C_t\Delta$ **then**
18 break;
19
20
21
22 $j = j - 1$;
23 **if** $j < j(t)$ *AND* $data > C_t\Delta$ **then**
24 break constraint of satisfying $buf_i(t+\Delta) \geq B_{min}, \forall i$, and defer the download of as many segments as needed verifying $j > j(t)$ (at least the next is kept scheduled), in descending order of playback position and distance to FoV(t)
25

6.3. Simulation results

To assess how much gain can VW and SD bring to a responsive reference FoV-based adaptation, we set $\Delta = 1$, $K = 2$, $B_{min} = 3s$ and $B_{max} = 10s$ for all adaptation strategies *Adaptation-VW*, *Adaptation-unaware*, *Adaptation-aware-VW* and *Adaptation-aware-SD*. There are $L = 2$ possible quality levels. The results are shown for the head motion traces of User 4 of the experiments, on the Boxing video for VW and Bar video for SD. The results are qualitatively equivalent for all users or videos. The x-axes represent the user time, which may be dilated compared to video playback time, as it accounts for startup delay, stalls and SD.

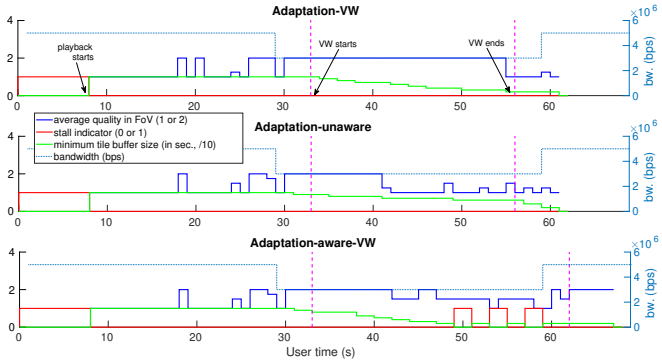

Fig. 16: Time series for *Adaptation-VW*, *Adaptation-unaware* and *Adaptation-aware-VW*. Purple dotted lines mark the VW period (set in video time, shown in user time).

The assessment of streaming with VW is shown in Fig. 16. As VW is envisioned to be a complementary lever the adaptation logic may trigger, we consider the typical network scenario for which VW has been designed: upon sensing a bandwidth drop, the adaptation logic decides to trigger a VW alternatively to dropping the quality, as we can see *Adaptation-unaware* does, or undergoing stalls, as *Adaptation-aware-VW* does. *Adaptation-VW* however is allowed to maintain high quality in the FoV with avoiding playback stalls.

The assessment of streaming with SD is shown in Fig. 17. The bandwidth is considered this time limited to $5Mbps$ (sufficient to play traditional non-360° in Full HD) and constant. We mention it is not uncommon to evaluate systems where the client has constant but insufficient bandwidth, see, e.g., [29, Sec. 6]. First, we observe that slowing down the video in the first seconds allows to fetch a higher quality without inflating the playback startup delay: the startup delay for *Adaptation-SD* is the same as for *Adaptation-unaware*, while fetching the high quality hurts the startup delay for *Adaptation-aware-SD*. To illustrate the interest of SD more generally, we have placed a second SD period from 13s to 17s (in video time) to emulate a new scene starting at 13s. We can see that *Adaptation-SD* is able to maintain maximum quality most of the playback time, while *Adaptation-unaware* cannot maintain this level of quality in the FoV to ensure the same level of buffering. *Adaptation-aware-SD* on the other hand, strives to deliver the same quality as *Adaptation-SD* in the FoV, but this comes at the expense of playback stalls.

7. Discussion

VW and SD will be particularly useful when there is a significant discrepancy between the available bandwidth and the bitrate of the highest quality of the sphere: the higher this discrepancy, the narrower the area where the quality can be maximum. This discrepancy will worsen with future headsets with significantly increased resolution (such as the newly

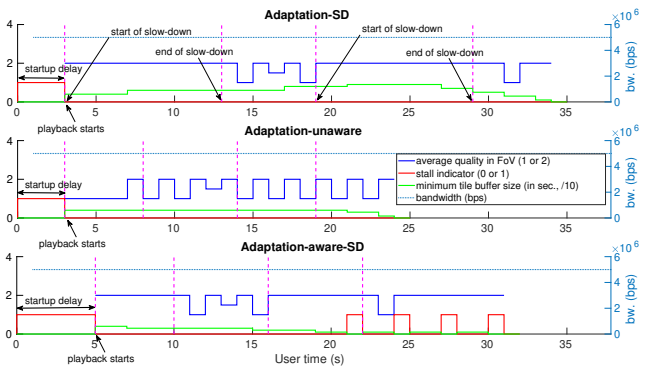

Fig. 17: Time series for *Adaptation-SD*, *Adaptation-unaware*, *Adaptation-aware-SD*. Purple dotted lines mark the SD period (set in video time, shown in user time).

released Varjo with 50 megapixels per eye). Resorting to SD and VW will enable to increase this area. This has been echoed by the findings of two very recent articles, [29, 52] which investigate the performance of current commercial live 360° video streaming services. Interestingly, in [29], Liu et al. show that the user-perceived resolution is often very low, between 240p and 360p, and that streaming the viewport only can increase the visible quality and alleviate stalls. This is particularly striking for Facebook 360 Live, which encodes the video into 2 quality levels only. Indeed, live 360° video streaming suffers from very important one-way delay, which is mainly due to transcoding operations in the cloud, between the producer and the client [52]. By having the producer only uploading the visible FoV to transcode, Liu et al. in [29] show that this delay can be cut. Similarly, VWs have the potential to diminish the computational requirements of transcoding to high qualities by reducing the size of the frame to transcode.

The very idea of our approach is to (i) identify and (ii) exploit trade-offs between different aspects of the VR user's experience in degraded bandwdith environment. This article has focused on (i) and uncovered dimensions of the user's experience in 360° (VW and SD) that can be advantageously modulated. The question is then how to trade between the different types of impairments to reach an optimal level of experience under limited communication resources. To achieve (i), we have considered hand-picked video periods. Automatizing the triggering of these impairments depending on the scene's content, the available bandwidth and the user's behavior is a major perspective, which should leverage the manifold of computer vision and new machine learning tools. Beyond addressing the problem from a system point of view, a major perspective in Human Computer Interactions, is to identify how to include the user in these choices, allowing her to find her preferences in the control of the reception of content in a degraded network environment.

8. Conclusions and Future works

This article has identified two new types of impairments to help streaming VR videos under limited bandwidth. We have built on the recent characterization of human attention in VR to introduce Virtual Walls and Slow Down, which we show to be well-accepted and useful to improve the level of experience compared with quality adaptation alone. The SD and VW impairments are complementary in that they are meant to apply to different types of scenes (exploration and concentrated focus, respectively). User experiments with a double-stimulus approach show that both VW and SD impairments are generally preferred by the users over the compression-only reference. A thorough analysis of quantitative subjective assessments, as well as objective metrics (from logs) enabled to understand the important factors involved in the user's preference. We have also confirmed the usefulness of these impairments from a system perspective, illustrating with network simulations how much gain in streaming performance can VW and SD bring to reference FoV-based adaptations.

One important perspective is then to design network-adaptive and user-adaptive strategies that would decide to trigger the best suited type of impairment at any point in time. Such a system may be thought of in an interactive manner where the user could choose which type of trade-off between VW/SD and compression would she like, depending on the available network bandwidth. This type of adaptation should leverage machine learning-based tools to build on relevant features of the user's motion and network bandwidth profile.

Acknowledgments

This work has been supported by the French government, through the UCA JEDI and EUR DS4H Investments in the Future projects ANR-15-IDEX-0001 and ANR-17-EURE-0004.

References

[1] Almquist, M., Almquist, V., Krishnamoorthi, V., Carlsson, N., Eager, D.: The prefetch aggressiveness tradeoff in 360 degree video streaming. In: Proceedings of the 9th ACM Multimedia Systems Conference. pp. 258–269. MMSys '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3204949.3204970, http://doi.acm.org/10.1145/3204949.3204970

[2] Ballard, T., Griwodz, C., Steinmetz, R., Rizk, A.: Rats: Adaptive 360-degree live streaming. In: Proceedings of the 10th ACM Multimedia Systems Conference. pp. 308–311. MMSys '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3304109.3323837, http://doi.acm.org/10.1145/3304109.3323837

[3] Bangor, A., Kortum, P., Miller, J.: Determining what individual sus scores mean: Adding an adjective rating scale. Journal of Usability Studies **4**, 114–123 (May 2009)

[4] Bastug, E., Bennis, M., Medard, M., Debbah, M.: Toward Interconnected Virtual Reality: Opportunities, Challenges, and Enablers. IEEE Communications Magazine (2017)

[5] Beerends, J.G., De Caluwe, F.E.: The influence of video quality on perceived audio quality and vice versa. J. Audio Eng. Soc **47**(5), 355–362 (1999), http://www.aes.org/e-lib/browse.cfm?elib=12105

[6] Brooke, J.: Sus-a quick and dirty usability scale. Usability evaluation in industry pp. 189–194 (1996)

[7] Carlsson, N., Eager, D., Krishnamoorthi, V., Polishchuk, T.: Optimized adaptive streaming of multi-video stream bundles. IEEE Transactions on Multimedia **19**(7), 1637–1653 (July 2017). https://doi.org/10.1109/TMM.2017.2673412

[8] Caruso, E.M., Burns, Z.C., Converse, B.: Slow motion increases perceived intent. Proceedings of the National Academy of Sciences, USA (113), 9250–9255 (2106)

[9] Corporation, I.D.: Demand for Augmented Reality/Virtual Reality Headsets Expected to Rebound in 2018 (Mar 2018), industry report

[10] Coutrot, A., Guyader, N.: Learning a time-dependent master saliency map from eye-tracking data in videos. CoRR **abs/1702.00714** (2017)

[11] Dambra, S., Samela, G., Sassatelli, L., Pighetti, R., Aparicio-Pardo, R., Pinna-Déry, A.M.: TOUCAN-VR. Software (DOI: 105281/zenodo1204442 2018), https://github.com/UCA4SVR/TOUCAN-VR

[12] Dambra, S., Samela, G., Sassatelli, L., Pighetti, R., Aparicio-Pardo, R., Pinna-Déry, A.M.: Film editing: New levers to improve vr streaming. In: Proceedings of the 9th ACM Multimedia Systems Conference. pp. 27–39. MMSys '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3204949.3204962, http://doi.acm.org/10.1145/3204949.3204962

[13] David, E.J., Gutiérrez, J., Coutrot, A., Da Silva, M.P., Callet, P.L.: A dataset of head and eye movements for 360 degree videos. In: Proceedings of the 9th ACM Multimedia Systems Conference. pp. 432–437. MMSys '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3204949.3208139, http://doi.acm.org/10.1145/3204949.3208139

[14] Example: Manifest file. http://yt-dash-mse-test.commondatastorage.googleapis.com/media/car-20120827-manifest.mpd (2012)

[15] Farmani, Y., Teather, R.: Viewpoint snapping to reduce cybersickness in virtual reality. Graphic interfaces (05 2018). https://doi.org/10.20380/GI2018.21

[16] Fearghail, C.O., Ozcinar, C., Knorr, S., Smolic, A.: Director's cut - analysis of aspects of interactive storytelling for vr films. In: Rouse, R., Koenitz, H., Haahr, M. (eds.) Interactive Storytelling. pp. 308–322. Springer International Publishing, Cham (2018)

[17] Gaddam, V.R., Riegler, M., Eg, R., Griwodz, C., Halvorsen, P.: Tiling in interactive panoramic video: Approaches and evaluation. IEEE Transactions on Multimedia **18**(9), 1819–1831 (Sep 2016). https://doi.org/10.1109/TMM.2016.2586304

[18] Girod, B., Färber, N., Steinbach, E.G.: Adaptive playout for low latency video streaming. In: IEEE International Conference on Image Processing (ICIP). vol. 1, pp. 962–965 (2001)

[19] Google: VR180 cameras (2019), https://vr.google.com/vr180/

[20] Graf, M., Timmerer, C., Mueller, C.: Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 261–271. MMSys'17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3083187.3084016, http://doi.acm.org/10.1145/3083187.3084016

[21] Grogorick, S., Stengel, M., Eisemann, E., Magnor, M.: Subtle gaze guidance for immersive environments. In: Proceedings of the ACM Symposium on Applied Perception. pp. 4:1–4:7. SAP '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3119881.3119890, http://doi.acm.org/10.1145/3119881.3119890

[22] Hassenzahl, M., Tractinsky, N.: User experience - a research agenda. Behavior and Information Technology **25**, 91–97 (2 2006), http://attrakdiff.de

[23] Hu, H., Lin, Y., Liu, M., Cheng, H., Chang, Y., Sun, M.: Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1396–1405 (July 2017). https://doi.org/10.1109/CVPR.2017.153

[24] Jeppsson, M., Espeland, H., Griwodz, C., Kupka, T., Langseth, R., Petlund, A., Qiaoqiao, P., Xue, C., Pogorelov, K., Riegler, M., Johansen, D., Halvorsen, P.: Efficient live and on-demand tiled hevc 360 vr video streaming. In: 2018 IEEE International Symposium on Multimedia (ISM). pp. 81–88 (Dec 2018). https://doi.org/10.1109/ISM.2018.00022

[25] de Jesus Oliveira, V.A., Brayda, L., Nedel, L., Maciel, A.: Designing a vibrotactile head-mounted display for spatial awareness in 3d spaces. IEEE Trans. on Visualization and Computer Graphics **23**(4), 1409–1417 (Apr 2017)

[26] Kim, J., Hahn, M.: Voice activity detection using an adaptive context attention model. IEEE Signal Processing Letters **25**(8), 1181–1185 (Aug 2018). https://doi.org/10.1109/LSP.2018.2811740

[27] Lewis, J.R., Sauro, J.: The factor structure of the system usability scale. In: Kurosu, M. (ed.) Human Centered Design. pp. 94–103. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)

[28] Li, Y., Markopoulou, A., Apostolopoulos, J., Bambos, N.: Content-aware playout and packet scheduling for video streaming over wireless links. IEEE Transactions on Multimedia **10**(5), 885–895 (Aug 2008). https://doi.org/10.1109/TMM.2008.922860

[29] Liu, X., Han, B., Qian, F., Varvello, M.: LIME: Understanding Commercial 360 degree Live Video Streaming Services. In: Proceedings of the 10th ACM Multimedia Systems Conference. pp. 154–164. MMSys '19, ACM, New York, NY, USA (2019)

[30] Magliano, J., Zacks, J.M.: The Impact of Continuity Editing in Narrative Film on Event Segmentation. Cognitive Science **35**(8), 1489–1517 (2011)

[31] Mather, G., Sharman, R.J., Parsons, T.: Visual adaptation alters the apparent speed of real-world actions. Scientific Reports **7**(1) (2017)

[32] MPEG: Omnidirectional Media Application Format (Jan 2018)

[33] Nguyen, A., Yan, Z., Nahrstedt, K.: Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In: ACM Multimedia Conference. pp. 1190–1198. ACM (2018)

[34] Niamut, O., Thomas, E., D'Acunto, L., Concolato, C., Denoual, F., Lim, S.Y.: Mpeg dash srd: Spatial relationship description. In: ACM MMSyS (May 2016)

[35] Oculus: Oculus Best Practices (2017), version 310-30000-02

[36] Pavel, A., Hartmann, B., Agrawala, M.: Shot orientation controls for interactive cinematography with 360 video. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. pp. 289–297. UIST '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3126594.3126636, http://doi.acm.org/10.1145/3126594.3126636

[37] Petrangeli, S., Swaminathan, V., Hosseini, M., Turck, F.D.: An HTTP/2-based adaptive streaming framework for 360 virtual reality videos. In: ACM Multimedia Conf. (Oct 2017)

[38] Rondao Alface, P., Macq, J.F., Verzijp, N.: Interactive omnidirectional video delivery: A bandwidth-effective approach. Bell Lab. Tech. J. **16**(4), 135–147 (Mar 2012). https://doi.org/10.1002/bltj.20538, http://dx.doi.org/10.1002/bltj.20538

[39] Sassatelli, L., Winckler, M., Fisichella, T., Dezarnaud, A., Lemaire, J., Aparicio-Pardo, R., Trevisan, D.: TOUCAN-VR Data 2019. Data (2019), https://github.com/UCA4SVR/TOUCAN-VR_data_2019

[40] Sassatelli, L., Winckler, M., Fisichella, T., Dezarnaud, A., Lemaire, J., Aparicio-Pardo, R., Trevisan, D.: TOUCAN-VR Generalized Operations. Software (2019), https://github.com/UCA4SVR/TOUCAN-VR/tree/generalized_ops

[41] Sassatelli, L., Pinna-Déry, A.M., Winckler, M., Dambra, S., Samela, G., Pighetti, R., Aparicio-Pardo, R.: Snap-changes: A dynamic editing strategy for directing viewer's attention in streaming virtual reality videos. In: Proceedings of the 2018 International Conference on Advanced Visual Interfaces. pp. 46:1–46:5. AVI '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3206505.3206553, http://doi.acm.org/10.1145/3206505.3206553

[42] Serrano, A., Sitzmann, V., Ruiz-Borau, J., Wetzstein, G., Gutierrez, D., Masia, B.: Movie Editing and Cognitive Event Segmentation in Virtual Reality Video. ACM Trans. on Graphics (2017)

[43] Singla, A., Göring, S., Raake, A., Meixner, B., Koenen, R., Buchholz, T.: Subjective quality evaluation of tile-based streaming for omnidirectional videos. In: Proceedings of the 10th ACM Multimedia Systems Conference. pp. 232–242. MMSys '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3304109.3306218, http://doi.acm.org/10.1145/3304109.3306218

[44] Sitzmann, V., Serrano, A., , Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in VR: How Do People Explore Virtual Environments? IEEE Trans. on Visualization and Computer Graphics (2018)

[45] Slavkovic, A.: Loglikelihood and confidence intervals. http://personal.psu.edu/abs12/stat504/Lecture/lec3_4up.pdf (2005)

[46] Union, I.T.: Subjective video quality assessment methods for multimedia applications (Apr 2008), iTU-T P.910

[47] Union, I.T.: Methodology for the subjective assessment of the quality of television pictures (Jan 2012), recommendation ITU-R BT.500-13

[48] Verhelst, W., Roelands, M.: An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 2, pp. 554–557 vol.2 (April 1993).

https://doi.org/10.1109/ICASSP.1993.319366
[49] Vimeo: Video compression guidelines. https://vimeo.com/help/compression (2019)
[50] Waldin, N., Waldner, M., Viola, I.: Flicker Observer Effect: Guiding Attention Through High Frequency Flicker in Images. Comput. Graph. Forum **36**(2), 467–476 (May 2017)
[51] Xiao, M., Zhou, C., Swaminathan, V., Liu, Y., Chen, S.: Bas-360: Exploring spatial and temporal adaptability in 360-degree videos over http/2. In: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications. pp. 953–961 (April 2018). https://doi.org/10.1109/INFOCOM.2018.8486390
[52] Yi, J., Luo, S., Yan, Z.: A Measurement Study of YouTube 360 degree Live Video Streaming. In: Proceedings of the 29th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. pp. 49–54. NOSSDAV '19, ACM, New York, NY, USA (2019)
[53] Zhang, X.L., Wang, D.: Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans. Audio, Speech and Lang. Proc. **24**(2), 252–264 (Feb 2016). https://doi.org/10.1109/TASLP.2015.2505415, http://dx.doi.org/10.1109/TASLP.2015.2505415