Snap-changes: a Dynamic Editing Strategy for Directing Viewer's Attention in Streaming Virtual Reality Videos

Lucile Sassatelli, Anne-Marie Pinna-Déry, Marco Winckler, Savino Dambra, Giuseppe Samela, Romaric Pighetti, Ramon Aparicio-Pardo*
Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

ABSTRACT

Cinematic Virtual Reality (VR) has the potential of touching the masses with new exciting experiences, but faces two main hurdles: one is the ability to stream these videos, another is their design and creation. Indeed, rates are much higher and in addition to discomfort and sickness that might arise in fully immersive experience with a headset, users might get lost when exploring a 360° videos and miss main elements required to understand the underlying plot. We take an innovative approach by addressing jointly the creation and streaming problems. We introduce a technique called snap-changes, aimed at directing viewers to points of interest pre-defined by the content producer. We design a VR editing tool and a custom 360° video player, to provide the content creator with the ability to drive the user's attention, and report results from two sets of user experiments that indicate that snap-changes indeed help reduce user's head motion.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **Empirical studies in HCI**;

KEYWORDS

360° video, video editing, user studies, streaming, attention driving

1 INTRODUCTION

Owing to its immersive capacity and the unprecedented feeling of presence it provides, Virtual Reality (VR) is becoming a powerful instrument for buidling 360° videos in a broad range of goals and genres (documentaries, storytelling, journalism, etc.). By the end of 2017, 24 million households worldwide have purchased a VR headset/Head-Mounted Display (HMD) [15]. However, there are two main hurdles to the large adoption of VR videos: one is the ability to stream these videos, another is the design and creation of these videos for best user's experience.

On one hand, traditional filmmaking is built around a scenario for which the creator has to guide viewers' attention. A major tool to do so is editing, to create a sequence of scenes whose logical composition unfolds a plot. Cinematic VR however enables users to have a more active role by exploring the virtual space while the scenario plays. VR filmmaking is still a much unexplored domain, wide open for innovation as presented in [3]. On the other hand, 360° videos are much heavier to stream, and an increasingly considered strategy to stream them is to split the sphere in space and time, into tiles and segments, and send in High Quality (HQ) only the part in the user's Field of View (FoV), the rest in Low Quality (LQ). This has been standardized in the Spatial Relationship Description (SRD) amendment to the MPEG-DASH standard [14]. Anticipating the user's head position several seconds ahead is therefore key, and if done wrongly, leads to a substantial bandwidth overhead to replace the tiles wrongly downloaded in LQ. If the current network bandwidth does not permit replacing without risk of stalling the playback, then LQ is displayed to the user, which is detrimental to the Quality of Experience (QoE). Anticipating better the head position by lowering its velocity is therefore key in limiting the bandwidth required to ensure that HQ be displayed in the FoV.

In this article, we propose a VR editing tool, named snap-changes, that aims both at giving the creator a partial control on the user's motion, hence attention, while at the same time benefiting the streaming process. The innovation of our approach stems from the observation that these two are no longer decoupled as in legacy video, where the users' gaze scanning the images does not immediately impact the network bandwidth consumption. This technique was developed in a research project with a filmmaking company as client.

Contributions:

• We design a VR editing tool and a custom 360° video player, to provide the content creator the ability to drive the user's attention, and possibly ease the streaming process at the same time, depending on the desired trade-off between user's freedom and attention/bandwidth consumption narrowing.

• We illustrate with two sets of user experiments that snap-changes can help reduce the user motion according to the scene complexity (diversity, spreading and motion of the points of interests, angle of view, etc.).

The concept of snap-changes is introduced in Sec. 2. Sec. 3.1 describes how the new concept is incorporated into a VR video player to make the streaming decision benefit from the creator's input. Experimental results are presented in Sec. 3, while Sec. 4 discusses the multi-disciplinary research directions opened by this work.

*S. Dambra, G. Samela and R. Pighetti are now with Eurecom, Politecnico di Bari and Francelabs, respectively. Corresponding author: sassatelli@i3s.unice.fr

AVI '18, May 29-June 1, 2018, Castiglione della Pescaia, Italy

ACM ISBN 978-1-4503-5616-9/18/05…$15.00
https://doi.org/10.1145/3206505.3206553

2 SNAP-CHANGES FOR VR EDITING

First the necessary background on editing is reviewed, then the concept of snap-change is introduced.

2.1 Region of interest and related works

An image is made of different zones which can be weak or strong attractors for human attention. The latter are called Regions of Interest (RoI), or salient points (see, e.g., Fig. 2). High-saliency regions are characterized by both (i) low-level features (related to the primary human visual system) and (ii) high-level features (related to the semantics) [4]. The user's motion being heavily driven by the so-defined salient regions, it is therefore crucial to take them into account in our editing design so as to limit the head motion. Other works aimed at driving the user in 360° environments, yet not to lower head velocity and bandwidth. Some inlaid more or less subtle artifacts to catch users' attention [1, 10], while others completely drove the FoV, possibly yielding harsh vestibular contradictions [13]. None of them relied on fast cutting to feel natural and go mostly unnoticed, as our proposal does.

2.2 The concept of editing

Film editing consists in selecting shots from the raw footage and assembling them (see Fig. 1). Editing is by construction a main tool to control the user's attention. Several types of editing exist [2]. From circa 2006, editing has entered the so-called "post-classical" phase, where the average shot length (ASL) has dropped from 8-11s in the 1930's to 3-4s today [5]. It is in particular aimed at changing scene fast (for example from a room to a car) without showing the intermediate steps and letting the brain fill in. In this article, this is what we leverage to control motion.

The concept of VR editing was introduced in May 2016 by the Google principal filmmaker Jessica Brillhart [3]. In VR, a 360° sequence-shot can be represented as a circular color stripe; the radial axis is the time. Editing consists not only in sorting the stripes (shots) in time, but also in rotating them around each other to control where the viewer arrives in the next scene, depending on her point of view in the previous scene (see Fig. 2). Controlling such transitions between the scenes means to control how the user feels teleported from one world to the other, and is hence central for QoE and motion limitation.

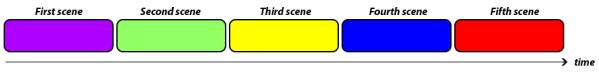

Figure 1: Editing legacy videos: arranging scenes over time.

2.3 Dynamic editing

Moving the 360° camera is tricky, and a priori constrains techniques such as traveling [18]. HCI research has provided valuable guidelines for creating VR applications[16]. It has been shown that linear and constant-speed motion is tolerated by the ear-vision system, as well as so-called snap-changes [9], which are fast-cuts in a 360° scene to allow moving while skipping any non-linear motion that would create sickness when the user does not move, and letting the brain just "fill in the blanks" without the vestibular system being involved. We leverage this concept to drive the user's attention (e.g., making user's focus on RoIs decided by the filmmaker), reduce movement, and increase feeling of immersion by not missing main events.

Figure 2: Left: Region of Interest (RoI). Right: 360° scenes over time, black (white) dot is RoI at the beginning (end) of the scene.

Definition. A snap-change consists in repositioning the user (at runtime and when needed) in front of a RoI, by rotating the sphere in a snap (as shown in Fig. 3). On one hand, the user will undergo these re-positioning, but we posit that, if done based on the scene content, they can go mostly unnoticed. On the other hand, by taking some freedom off of the user, we remove the gaze uncertainty by the same amount, the decision of which quality to fetch based on future gaze position are hence made exact, thereby lowering the amount of required replacements. The streaming application therefore needs to be upgraded to consider the presence of the forthcoming snap-changes and hence exploit the advantages of editing.

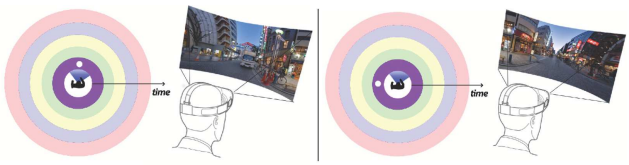

Figure 3: A snap-change re-positions the user in front of another sector of the same scene, in a snap.

Implementation. The angular position of the desired video sector is found by superimposing a *millimetric paper* onto the projection of the 360° video in Adobe Premiere. Fig. 4 illustrate this in a sector centered around the white truck whose angular position is −90°. The angular value specifies by how many degrees the sphere should be rotated in order to have the desired part of the video in the FoV.

The last information needed for a snap-change are the indexes of the tiles overlapping the desired video sector, so as to use this knowledge in the fetching process (described in Sec. 3.1). These three pieces of information are gathered into an XML tag:

```
<?xml version="1.0"?>
<snapchange>
        <milliseconds>25000</milliseconds>
        <roiDegrees>-90</roiDegrees>
        <foVTile>1,2,4,5</foVTile>
</snapchange>
```

All the tags are chronologically sorted into an XML file which is used by the client application to dynamically edit the content.

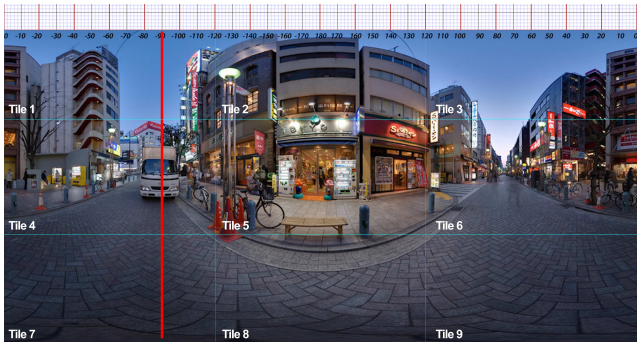

Figure 4: Identification of the targeted RoI angular position for snap-change construction.

Following the "30 Degree Rule" which states that the shift in camera angle between two shots of the same subject must exceed 30° [11], we check 100ms before the time of the snap-change if the difference between the current position and the snap-change angle is lower than 30°. If so, the snap-change is not triggered. Otherwise, it is and the sphere is rotated by the angular difference in a snap, to make the user face the desired video sector.

3 EMPIRICAL EVALUATION

This article's focus is on the impact of snap-changes on user's behavior, in particular head velocity and position w.r.t. elements intended by the creator to be noticed. For the impact on bandwidth savings, we refer the reader to [6].

3.1 Testbed

We make an Android application in charge of streaming and playing the video, and logging metrics. Our code and data are freely available online [7, 8]. It builds mainly on Android, and the Samsung Gear VR framework [17]. We use a Samsung Galaxy S7 Edge coupled with a Samsung Gear VR headset. The snap-changes description is leveraged at the player to pre-fetch the tiles in future FoV in HQ, regardless of the current head position. The prediction error at the time of snap-changes is therefore zero, hence no replacement, incurring bandwidth waste, is needed to ensure HQ in FoV. More details on the network algorithm can be found in [6].

3.2 Hypotheses and experimental plan

The core idea of introducing snap-changes is to be able to predict the future head position by (re-)capturing the user's attention. By getting the user to follow the important element(s) in the scene, we expect:

- H1: the head motion (i.e., velocity) to be lowered according to the complexity of the scene;
- H2: the user to notice the elements intended by the creator.

We run two sets of user experiments to get insights on the above hypotheses, the first (resp. second) set named SET1 (resp. SET2) meant to investigate H1 (resp.H2). SET1 has been designed to show the impact of snap-changes (referred to as *Dynamic* in figures) on user's head motion, compared to no snap-changes (referred to as *Static*). Owing to the high variability of motion between users and contents, we set each user to watch the same content with and without snap-changes. SET2 uses another protocol aimed at showing that snap-changes indeed make users see and remember elements they are more likely to miss without snap-changes. To prove so, a user goes through Static and Dynamic, but on two different parts of the content to have two disjoint sets of elements to notice in each case. The video used in the experiments is a 360°-video thriller series known as "Invisible" [12]. From episodes 3 and 6, we create a content with a variety of scene transitions, which lasts 5min 20 s, and whose story is intelligible. SET1 is made of 4 comparisons with 4 different users. From small sample size statistics, we know that a low number of users only allows to observe so-called big effects to be statistically significant. SET2 is made of 10 users. Randomization has been used to prevent bias from viewing order (Static or Dynamic first).

3.3 Results

While SET1 enabled to show that snap-changes lower head motion with a p-value of 0.10, we set to investigate H1 by analyzing time series of users, illustrated by one specific user in Fig. 5. First, it is visible that the time average of head motion is lowered with snap-changes. The top dots show triggered replacements, which vary with the velocity. Indeed, the slower the head the better the prediction at the time of requesting the video chunk (a few seconds ahead of its playback). As well, the bottom graphs show the variations of the quality in the FoV (fraction of tiles in HQ). It decreases when the prediction was wrong and replacements cannot be triggered (because the playback buffers are not filled enough), that is when the motion is higher.

The snap-changes are depicted by magenta triangles: a triangle at the top (resp. middle) means the snap has not been (resp. has been) triggered because the users was looking within (resp. outside) a sector of 30° from the snap-targeted angle. Non-triggered snaps therefore mean the user was looking at the region intended by the creator. In such cases (until about 100s for this user), we observe that looking at the intended Point of Interest (PoI) lowers the need to explore and move fast. In turn, when the snap-changes are triggered, the re-positioning of the user in front of the PoI seem to entice him not to move away from the PoI (no two snap-changes are triggered in a row).

Second, the colored bar in the bottom of the figure depicts the complexity of different periods of video, that we assume dependent on factors liable to impact user's motion: spherical spreading of interesting sounds, of PoIs, motion of PoIs, duration of the scene (allowing exploration outside the PoIs), changes of scenes/environments (depicted by red vertical lines). Blue, magenta and red grades the increasing complexity. For instance, period B is made of different environments where the PoIs move through a large sector of the sphere (often crossing it over 180°), while C features a fistfight close to the camera but in the corner of a room. We observe that while the motion is reduced when the user's attention is captured in the case of Dynamic, its absolute level remains dependent on the scenes' complexity.

Fig. 6 shows that the introduction of snap-changes has a direct impact on the elements (objects or actions) noticed by the user: the task score is significantly higher with snap-changes. Interestingly,

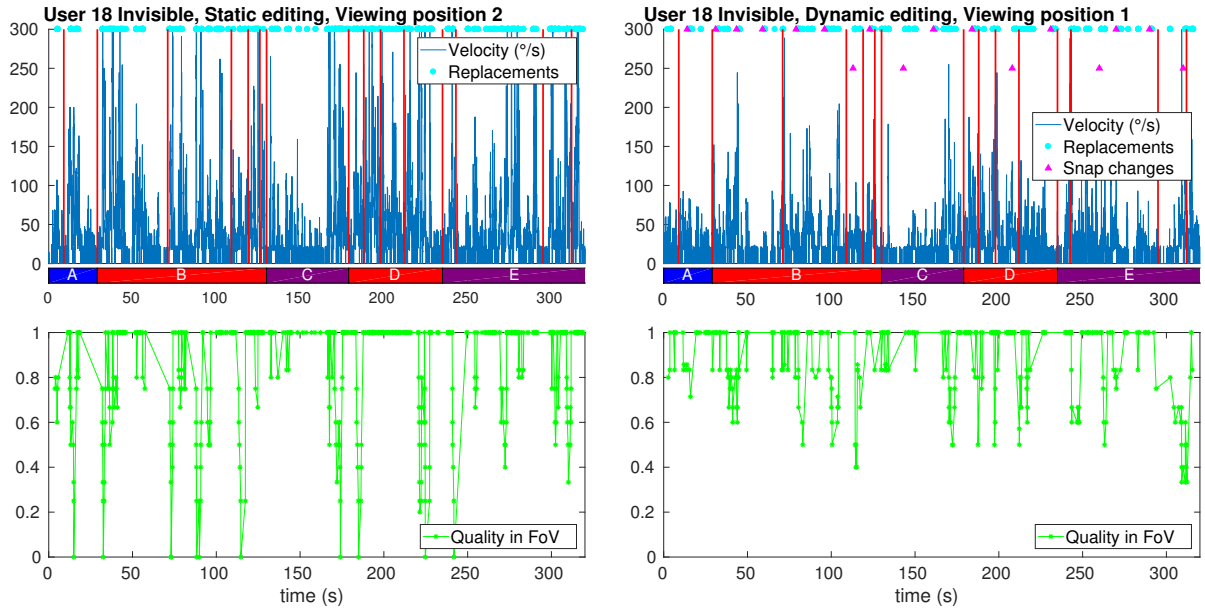

Figure 5: A user from SET1. Left: without snap-changes. Right: with snap-changes. Top: y-axis in degree/s for the velocity. The time average of head motion is lowered with snap-changes. The top dots show triggered replacements, which vary with the velocity. The colored horizontal bar depicts the complexity of different periods of the video.
Bottom: y-axis as a fraction of tiles of HQ in all the tiles in the FoV.

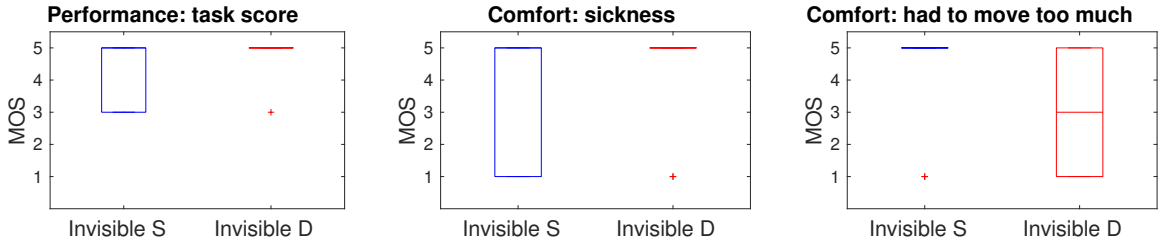

Figure 6: Subjective metrics as boxplots from SET2, without (Static - S) and with (Dynamic - D) snap-changes.

the users seem to feel less sick with a dynamically edited video, but feel like they need to move more, which seems contradictory with the objective data in Fig. 5. However, these last two metrics are not significant owing to the too low number of persons having felt that way. This is nonetheless a very interesting impact of snap-changes, that will be the subject of future work.

4 DISCUSSION AND FUTURE WORK

This paper contributes with a technique for dynamically editing VR videos: snap-changes allow to gently direct user attention to RoIs pre-defined by the filmmaker, while slowing down the head and enabling to predict its position, which in turn saves bandwidth (essential for VR streaming distribution). To the best of our knowledge, this technique is original in both scientific and industrial communities. It is worth noting that the use of snap-changes is part of the task of edition, so that it is up to filmmakers to decide into what extent users are allowed to freely explore the scene when they should focus on a PoI. We demonstrate the feasibility of the concept with a testbed application. Preliminary results show that snap-changes are easy-to-use and effective tools to maintain or re-capture user's attention in VR, and can help lower the quantity of motion required to follow dynamic scenes. Nonetheless, further studies with a larger user population are necessary to provide definite guidelines. In particular we want to cross-check our hypotheses and investigate how snap-changes can be tunned to lower sickness. Also, the triggering of snap-changes can inform the filmmaker on whether and when users disengage from the plot to explore the scenario. This point might suggest that snap-changes can be used as a systematic strategy for learning how to build more immersive and engaging 360° experiences.

ACKNOWLEDGMENTS

This work has been supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with reference number ANR-15-IDEX-01.

REFERENCES

[1] E. Barth, M. Dorr, M. Böhme, K. Gegenfurtner, and T. Martinetz. 2006. Guiding the mind's eye: improving communication and vision by external control of the scanpath. In *Human Vision and Electronic Imaging XI (Proc. of the SPIE)*, B. E. Rogowitz, T. N. Pappas, and S. J. Daly (Eds.), Vol. 6057. 116–123.

[2] D. Bordwell and K. Thompson. 2006. *Film Art: An Introduction.* McGraw-Hill.

[3] J. Brillhart. 2016. VR and cinema. https://medium.com/@brillhart. (May. 2016). Google I/O conf.

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

[5] J.-E. Cutting, K.-L. Brunick, J.-E. DeLong, C. Iricinschi, and A. Candan. 2011. Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception* 2, 6 (2011), 569–576.

[6] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry. 2018. Film Editing: New Levers to Improve Virtual Reality Streaming. In *ACM International Conference on Multimedia Systems (MMSys)*.

[7] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry. 2018. TOUCAN-VR. *Software* (DOI: 10.5281/zenodo.1204442 2018). https://github.com/UCA4SVR/TOUCAN-VR

[8] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry. 2018. TOUCAN-VR-data. *Software* (DOI: 10.5281/zenodo.1204392 2018). https://github.com/UCA4SVR/TOUCAN_VR_data

[9] Facebook. 2017. VR201: Lessons from the Frontlines. (Apr. 2017). Facebook Developers Conference.

[10] S. Grogorick, d E. Eisemann M. Stengel a, and M. Magnor. 2017. Subtle Gaze Guidance for Immersive Environments. In *Proceedings of the ACM Symposium on Applied Perception (SAP '17)*. Article 4, 7 pages.

[11] Hollywood lexicon. 2000. 30 Degree Rule. http://www.hollywoodlexicon.com/thirtydegree.html. (2000).

[12] D. Liman. 2016. Invisible. http://www.imdb.com/title/tt6178894/. (2016). VR series, Samsung.

[13] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell Me Where to Look: Investigating Ways for Assisting Focus in 360 degree Video. In *ACM Conf. on Human Factors in Computing Systems (CHI)*. 2535–2545.

[14] O. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual, and S.-Y. Lim. 2016. MPEG DASH SRD: Spatial Relationship Description. In *ACM MMSyS*.

[15] Parks and Associates. 2017. Virtual Reality: Disrupting the Entertainment Experience. (Jun. 2017). Industry report.

[16] S. Saarinen, V. Makela, P. Kallioniemi, J. Hakulinen, and M. Turune. 2017. Guidelines for Designing Interactive Omnidirectional Video Applications. In *IFIP INTERACT*.

[17] Samsung. 2017. Samsung Gear VR Framework. (2017).

[18] C. Milk (Within). 2016. The birth of virtual reality as an art form. (Jun. 2016). TED talk.