User-Adaptive Editing for 360 degree Video Streaming with Deep Reinforcement Learning

Lucile Sassatelli, Marco Winckler, Thomas Fisichella, Ramon Aparicio-Pardo

User-Adaptive Editing for 360° Video Streaming with Deep Reinforcement Learning

Lucile Sassatelli, Marco Winckler, Thomas Fisichella, Ramon Aparicio*
Université Côte d'Azur, CNRS, I3S
Sophia Antipolis, France

ABSTRACT

The development through streaming of 360° videos is persistently hindered by how much bandwidth they require. Adapting spatially the quality of the sphere to the user's Field of View (FoV) lowers the data rate but requires to keep the playback buffer small, to predict the user's motion or to make replacements to keep the buffered qualities up to date with the moving FoV, all three being uncertain and risky. We have previously shown that opportunistically regaining control on the FoV with active attention-driving techniques makes for additional levers to ease streaming and improve Quality of Experience (QoE). Deep neural networks have been recently shown to achieve best performance for video streaming adaptation and head motion prediction. This demo presents a step ahead in the important investigation of deep neural network approaches to obtain user-adaptive and network-adaptive 360° video streaming systems. In this demo, we show how snap-changes, an attention-driving technique, can be automatically modulated by the user's motion to improve the streaming QoE. The control of snap-changes is made with a deep neural network trained on head motion traces with the Deep Reinforcement Learning strategy A3C.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Networks** → *Network simulations*; • **Computing methodologies** → *Neural networks*.

KEYWORDS

360° video streaming; deep reinforcement learning; film editing; motion prediction; user attention; recurrent neural networks

ACM Reference Format:
Lucile Sassatelli, Marco Winckler, Thomas Fisichella, Ramon Aparicio. 2019. User-Adaptive Editing for 360° Video Streaming with Deep Reinforcement Learning. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3343031.3350601

*Corresponding author: sassatelli@i3s.unice.fr

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350601

1 INTRODUCTION

Virtual Reality (VR) equipment and contents have been developing in the last couple of years, both from a technological and commercial point of view [3, 9]. Despite exciting prospects, the development of immersive applications is however persistently hindered by the difficulty to access them through Internet streaming owing to their bandwidth requirements [2]. A simple principle to lower the data rate is to send the non-visible part of the sphere with lower quality, and the Field of View (FoV) in high quality (e.g., the SRD extension to the MPEG-DASH standard splits an equirectangular projection into several tiles [11]). Doing so requires however to predict, at the time of the transmission, where the user is going to look at. Such prediction is only partly possible over very short time horizons (less than 2 seconds [10]). It therefore requires to keep the buffer size small. However, buffers are one of the main components of modern streaming to absorb network variations. To let the buffers build up to several seconds yet keeping the buffered qualities most up-to-date with user's motion, a wrong prediction must, if network permits, be corrected by downloading again in High Quality (HQ) segments initially fetched in Low Quality (LQ). The quality in the FoV and the consumed bandwidth therefore get highly dependent on the user's motion in 360° video streaming.

In [4, 12], we have shown how, complementarily to predicting the user's motion, we can control it to ease streaming and improve QoE. We have introduced so-called snap-changes, a film editing technique aimed at periodically re-gaining control on the FoV by repositioning, in a snap, that is from one frame to the next, the user in front of a pre-defined FoV (see Fig. 1). When bandwidth conditions do not allow to perform replacements (to keep the buffered qualities up to date with the moving FoV), the lever of snap-changes has been shown in [4] to enable to get back to a satisfactory level of QoE. The question is then how to modulate this new lever for 360° video streaming. This is subject of this demo.

The goal is to reach optimal QoE while the tile qualities are requested based on the current FoV at the time of download request, which can be several seconds ahead of playback in order to enable buffering (set to 4 sec. in the demo). We can hence formally define the general 360° video streaming control problem as deciding what to send (which part in HQ or LQ) and how to display it to the user (with a snap-change or not). The input parameters are multi-modal (network's state, human user's state, video content) and dynamic. The solutions for regular video streaming can resort to legacy dynamic optimization techniques such as Model Predictive Control [16] or Lypunov-based optimization [14]. However, it has been recently shown by Mao et al. in [7] that Deep Reinforcement Learning (DRL) approaches yield state-of-the-art performance. It has been also shown in several recent works (e.g., [10, 15]) that deep architectures provide best results for head motion prediction in

VR (in particular based on recurrent networks such as Long Short Term Memories, aka LSTMs). That is why we believe an important research avenue is investigating approaches based on deep neural networks to obtain user-adaptive and network-adaptive 360° video streaming systems.

Contribution: To progress towards this goal, we propose to demo our system incorporating the control of snap-changes by a deep neural network trained with the DRL strategy A3C [8]. The frequency of snap-change triggering is adapted online to the user's motion. The instantaneous reward (QoE obtained from each video segment playback) is made of the quality in the FoV penalized with the frequency of snap-changes. Indeed, snap-changes can be seen as restrictions of the user's freedom, which we intend to best modulate to maximize streaming QoE.

To the best of our knowledge, our recent works [4, 12, 13] are the first to make the connection between active attention-driving techniques and streaming algorithms design. Also, the automated control of editing cuts has never been proposed for 360° videos.

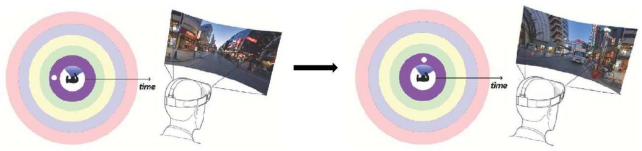

Figure 1: A snap-change is a repositioning of the user in front of a pre-defined FoV in a snap, i.e., from one frame to the next. It can be seen as an editing cut.

2 BUILDING BLOCKS

The 360° video streaming system we demonstrate comprises three main components: a (customizable) Deep Neural Network (DNN) making snap-change triggering decisions, an Android player incorporating the DNN and an external display showing the currently playing FoV and online streaming analytics.

Android player: Our 360° video player implements a streaming strategy based on MPEG-DASH SRD [11] tiled videos with buffering, replacements and FoV-based quality adaptation for the Samsung Gear VR on Android (details are provided in [4] and the code is available online [5]). There are as many buffers as the number of tiles (set to 6×4 in the demo). The maximum time interval between segment download and playback is hence the buffer size in seconds (set to 4 sec. in the demo). An XML file describing the possible snap-changes is parsed. It contains the triggering time of the snap-changes, the angular coordinates of the FoV to reposition the user to, and the list of tiles in this FoV. Upon the download request of the first tile of each new segment (of duration 1 sec.), if its playback time is after the next undecided snap-change, then the decision module is invoked to decide whether this snap will be triggered. If so, the tiles' qualities to download are decided based on the snap-change's FoV. Otherwise, they keep being decided based on the user's FoV at the time of the download.

Deep Neural Network: Our contribution is to enable the incorporation of a customizable DNN to control the triggering of snap-changes, and to propose an architecture and a training framework to demo the concept. The input is a simple representation of the environment with the past tiles in the FoV (over the last 2 sec.) and the future snap-changes already decided for the buffered segments. The output is the triggering probability, going out of a softmax layer. We use Asynchronous Actor Critic (A3C) [8] as the DRL approach, building on the implementation of [7]. The DNN is trained offline with a simulator of our exact 360° streaming player. It is fed with the 1121 head motion traces (with a sample each 0.2s) from the open dataset [6]. The instantaneous reward function $r(t)$, t describing the segment playback time, is set to $r(t) = Qual_{FoV}(t) - \gamma^{t-T(t)}$, where $T(t)$ is the time of the last snap triggered before t. Hyper-parameters are: buffer size is 4s, $\gamma = 0.3$, number of training epoch is 1000, number of processes (agents) is 8, fraction of data for training, validation and test is 70%, 10% and 20%, respectively, the number of units in each layer is 128. After the training on a multi-CPU machine with TensorFlow [1], the best model in validation is exported into a TensorFlow Lite format for inference on the Samsung phone.

3 DURING THE DEMO

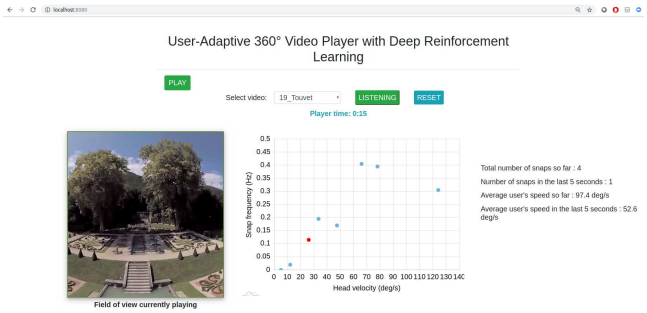

Figure 2: The external display.

A Samsung 7 Edge phone will play a video watched in a Samsung Gear VR headset by the visitor. The video is streamed from a laptop over WiFi. The visitor will first choose which video they prefer to watch, available from the open dataset in [6]. By changing deliberately their head motion speed, they will observe they get repositioned in front of new FoVs with different frequencies. At the end of the viewing experience, the user will be able to see on the externally displayed analytics how the snap-change frequency has varied depending on their head speed.

External display: During the demo, an external display will show the current FoV of the headset user (enabling the audience to see the snap-changes), as well as online analytics tracking the snap-change decisions made so far by the DNN, as depicted in Fig. 2. In particular, it will show that the frequency of snap-changes adapts to the head motion speed: more frequent snap-changes are required when the head's speed is higher, as the downloaded qualities are less likely to correspond to the user's FoV a few seconds later.

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] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.

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

[3] International Data Corporation. 2018. Demand for Augmented Reality/Virtual Reality Headsets Expected to Rebound in 2018. Industry report.

[4] S. Dambra, G. Samela, L. Sassatelli, R. Pighetti, R. Aparicio-Pardo, and A.-M. Pinna-Déry. 2018. Film Editing: New Levers to Improve VR Streaming. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 27–39. https://doi.org/10.1145/3204949.3204962

[5] 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

[6] E. J. David, J. Gutiérrez, A. Coutrot, M. Da Silva, and P. Le Callet. 2018. A Dataset of Head and Eye Movements for 360&Deg; Videos. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 432–437. https://doi.org/10.1145/3204949.3208139

[7] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. ACM, New York, NY, USA, 197–210. https://doi.org/10.1145/3098822.3098843

[8] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.), Vol. 48. PMLR, New York, New York, USA, 1928–1937. http://proceedings.mlr.press/v48/mniha16.html

[9] MPEG. 2018. Omnidirectional Media Application Format.

[10] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. 2018. Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction. In *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 1190–1198.

[11] 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*.

[12] Lucile Sassatelli, Anne-Marie Pinna-Déry, Marco Winckler, Savino Dambra, Giuseppe Samela, Romaric Pighetti, and Ramon Aparicio-Pardo. 2018. Snapchanges: 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 (AVI '18)*. ACM, New York, NY, USA, Article 46, 5 pages. https://doi.org/10.1145/3206505.3206553

[13] Lucile Sassatelli, Marco Winckler, Thomas Fisichella, Ramon Aparicio-Pardo, and Anne-Marie Dery-Pinna. 2019. A New Adaptation Lever in 360° Video Streaming. In *29th Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'19)*. Amherst, MA, United States.

[14] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman. 2016. BOLA: Near-optimal bitrate adaptation for online videos. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*. 1–9. https://doi.org/10.1109/INFOCOM.2016.7524428

[15] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. 2018. Gaze Prediction in Dynamic 360°Immersive Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5333–5342.

[16] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. *SIGCOMM Comput. Commun. Rev.* 45, 4 (Aug. 2015), 325–338. https://doi.org/10.1145/2829988.2787486