Anticipate the users' behavior for a deeper immersion

Laura Toni and Thomas Maugey

Introduction

Immersive media technology aims at endowing any final user with an unprecendent sense of full immersion in virtual (or real-world) environments. This is possible by projecting the user at the center of the 3D scene, which dynamically changes with the user interaction. This interactivity is driven by the headset mounted devices (HMD) in Virtual Reality, or by the user smartphone in Augmented Reality, or by tablet or remote control in Free Viewpoint Television, as depicted in Figure 1.

This user's interaction with the scene has created novel challenges from a coding and transmission perspective [1-3]. While in classical video streaming, the entire scene is encoded, delivered and displayed at the user side, in interactive/immersive systems only a portion of the full 3D scene is actually displayed. For example, in omnidirectional videos, the scene displayed in the HMD (viewport) is only a portion of the acquired spherical scene. Similarly, in multiview video coding, while a great number of views might be acquired,

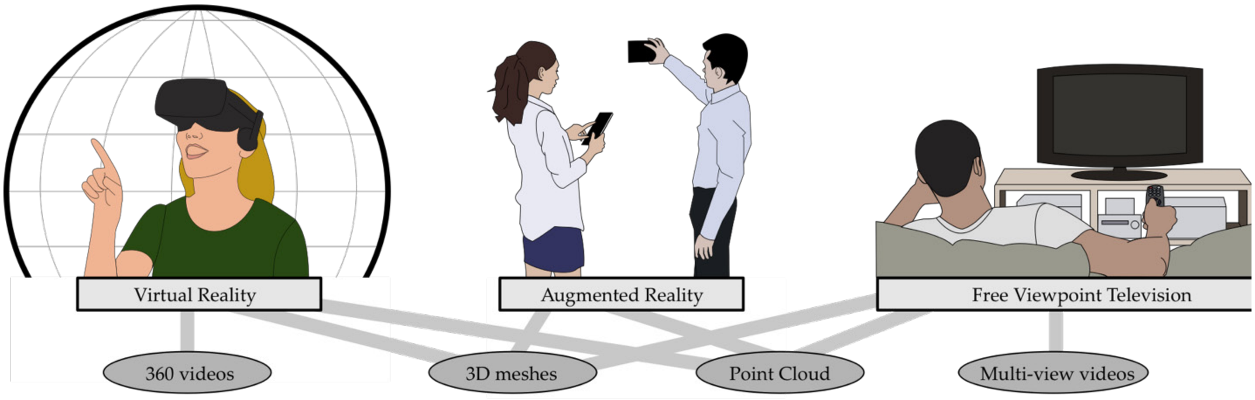

Figure 1. Immersive media technologies and their related data formats.

only a limited subset of them can be displayed at the same time. However, in practice the piece of data displayed by the user is not known *a priori* when both coding and streaming are performed. Therefore, when there is no prediction of the user behavior, the entire multimedia content needs to be coded and prefetched to users. This might lead to a reduction of the overall quality of the content in case of limited network resources. *It is therefore essential to properly predict users behavior to efficiently code and stream interactive content.* In this letter, we show how the user behavior is exploited in both bit allocation and streaming optimization strategies and we highlight the different interactive models that the two optimization problems require.

The Importance of Content Popularity in Bit Allocation Strategies

In data compression techniques, a maximum bit budget is usually available for compressing the data under consideration. The general criteria for an optimal compression is usually to describe with higher bitrates the more important data (and conversely), leading to an unequal bit allocation. In multi-view (MV) systems, for example, different cameras might be encoded at different quality levels [4-6], while in VR settings, different portions of the spherical content can be encoded with different quantization steps [7-9]. It is therefore essential to have proper metrics to reflect the « importance » of the data, *i.e.*, the content popularity. *In interactive services, this popularity reflects the probability for a piece of data to be displayed at the user's side*. In the following, we provide an overview on the optimization that an encoder needs to solve for carrying out the proper bit allocation strategy and we describe the associated challenges for the visual attention community.

Coding Problem formulation

Let us consider an interactive service, in which the video content acquired over time needs to be encoded. The overall goal for the encoder is to optimize the bit allocation strategy such that, on average, users consume high-quality media content while navigating. Decomposing a video content acquired by a camera into multiple portions, we denote by $x_{i,t}$ the i-th portion of the content acquired at time t. This can represent the i-th camera view in multi-view setting, or the i-th tile or portion of spherical content in VR settings. We then identify the full content acquired at time t with $\boldsymbol{x_t} = [x_{1,t}, \dots, x_{i,t}, \dots, x_{N,t}]$.

In this framework, the encoder seeks the best bit allocation strategy for each portion of the content. Denoting by $b(x_{i,t})$ the allocation for $x_{i,t}$ (*e.g.*, the QP for each $x_{i,t}$ content [9]), b($\mathbf{X}$) is the allocation strategy for the whole video content acquired in T successive acquisition time, with $\boldsymbol{X} = [\boldsymbol{x_1}, \dots, \boldsymbol{x_t}, \dots, \boldsymbol{x_T}]$. Therefore, the general problem formulation becomes

$$b^*: \arg\min_{b(\boldsymbol{X})} \sum_{t=1}^{T} \sum_{i=1}^{N} p(x_{i,t})\ D[b(x_{i,t})] + \lambda\, R[b(\boldsymbol{X})]$$

where $D[b(x_{i,t})]$ is the distortion of the i-th portion of the content acquired at time t when encoded with $b(x_{i,t})$ allocation strategy and $R[b(\boldsymbol{X})]$ is the total coding cost associated to the allocation strategy b($\mathbf{X}$).

Popularity estimation

The above problem formulation requires the *a priori* knowledge of *(i)* the video content characteristics (to evaluate D and R), *(ii)* the probability $p(\mathbf{X})$ of a data $\mathbf{X}$ to be requested by a final user (*content popularity*, *cf.* Figure 2 left). The latter is a new metric needed for interactive services and how to predict this content popularity is still an open question. Therefore, a compelling question that we pose to the visual attention community is: « *how can we predict the content popularity?*» Or analogously, « *how can we estimate the probability $p(\mathbf{X})$ of a data $\mathbf{X}$ to be requested by user?*».

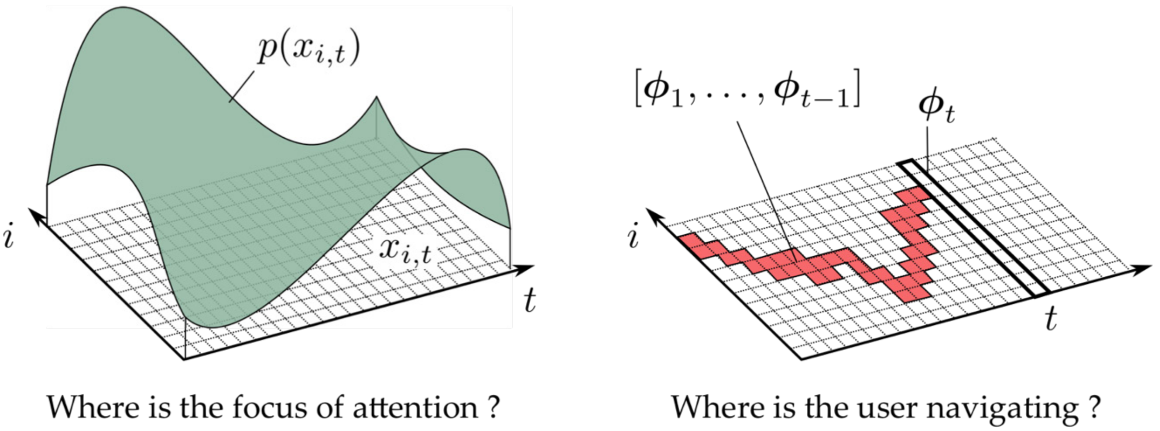
$p(x_{i,t})$
i
$x_{i,t}$
t
Where is the focus of attention ?
$[\phi_1, \ldots, \phi_{t-1}]$
ϕ_t
i
t
Where is the user navigating ?

Figure 2. A priori popularity vs navigation modeling

The Importance of Navigation Paths Prediction in Personalized Streaming

While in the previous problem the goal was to seek the best compression strategy to maximize the quality for a multitude of users, here we rather focus on personalized streaming strategies, properly designed for a specific user or class of users. A possible application of this personalized strategy is the adaptive streaming system, where a video content is encoded in multiple representations (multiple coding rates and resolutions) and stored at the server, and the client selects the representation to download [10]. The intelligence on which representation best fits the need of each client is therefore located at the client side, where the user behavior is known. In the context of interactive strategies, this personalized strategies have been optimized for MV systems [11-13] as well as for omnidirectional content [14-16]. In both scenarios, the personalized strategy optimization is performed knowing the user's past displayed data and predicting the future user's navigation. In the following, we first provide a general overview on the optimization problem to be solved in personalized streaming strategies, and then we describe the challenges on user behavior prediction.

Streaming Problem formulation

Similarly to the bit allocation problem formulation, we consider the whole video content acquired in T successive acquisition times $\boldsymbol{X}$. We denote by $\pi_{i,t} = \pi(x_{i,t})$ the streaming strategy for the content $x_{i,t}$ and by $\mathbf{\Pi} = [\boldsymbol{\pi_1}, \dots, \boldsymbol{\pi_t}, \dots, \boldsymbol{\pi_T}]$, $\pi_t = [\pi_{1,t}, \dots, \pi_{i,t}, \dots, \pi_{N,t}]$ the strategy for the whole video $\boldsymbol{X}$. For example, $\pi_{i,t}$ can be a binary variable denoting whether $x_{i,t}$ is scheduled or not [11]. Differently, $\pi_{i,t}$ can specify which representation is sent to the user for $x_{i,t}$. At the client side, the final user displays only a portion of the overall acquired content. We therefore introduce a displaying variable $\phi_{i,t}$ such that $\phi_{i,t} = 1$ if the user displays $x_{i,t}$, $\phi_{i,t} = 0$, otherwise, and we generalize the display vector as $\mathbf{\Phi} = [\boldsymbol{\phi_1}, \dots, \boldsymbol{\phi_t}, \dots, \boldsymbol{\phi_T}]$, $\phi_t = [\phi_{1,t}, \dots, \phi_{i,t}, \dots, \phi_{N,t}]$.

Equipped with the above notation, the streaming optimization can be formulated as follows

$$\mathbf{\Pi}^*\text{: arg}\min_{\mathbf{\Pi}} \sum_{t=1}^{T} \mathcal{D}[p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \mathbf{\Pi}] + \lambda R[\mathbf{\Pi}]$$

where $\mathcal{D}[p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \mathbf{\Pi}]$ is the objective function reflecting the quality experienced during the navigation or interaction (QoE). In interactive systems, this QoE does not take into account only the distortion of the displayed video, but also other factors such as the smoothness of the quality while navigating. A frequently-adopted metric for the QoE is, for example, the combination of both quality and quality variation over time [17]:

$$\begin{aligned}\mathcal{D}[p(\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \mathbf{\Pi}] &\\ &= \sum_{i=1}^{N} p(\phi_{i,t})\, D(\pi_{i,t}) \\ &+ \mu \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} p(\phi_{i,t}|\phi_{j,t-1}) \Delta D(\pi_{i,t}, \pi_{j,t-1})\end{aligned}$$

where μ is the multiplier that allows to assign the appropriate weight to quality variations in the objective metric, and

$\Delta D(\pi_{i,t}, \pi_{j,t-1})$ is the distortion variation experienced over time. This variation is weighted by the probability $p(\phi_{i,t}|\phi_{j,t-1})$ of displaying the *i*-th portion at time *t*, given that the *j*-th portion was previously displayed. This probability reflects the *navigation path* of the user (*cf*. Figure 2 right).

User navigation modeling

Most of the works focusing on personalized streaming strategies assume to know (or accurately predict) the navigation path, while we actually know that estimating user interactivity is an open challenge. It is worth noting that in this personalized strategies, knowing a global content popularity $p(\boldsymbol{X})$ is not enough. It is additionally required to estimate the behavior of each user *over time*. Solving this problem must take into account both the visual content (as in the popularity estimation) and the user behavior modeling (*e.g.*, highly dynamic vs. static navigation). In other words, the open questions posed to the visual attention community are: «How do we model and categorize users behavior over time?» and «Knowing both the content displayed by one user in the past, and his behavior modeling, can we anticipate the future navigation path?».

References

[1] T. El-Ganainy, and M. Hefeeda. “Streaming Virtual Reality Content”, arXiv:1612.08350 (2016).

[2] M Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes”, *Proc. IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, Fukuoka, Japan, Oct. 2015.

[3] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, “Free-viewpoint TV”, *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67-76, Jan. 2011.

[4] G. Cheung, V. Vladan, and A. Ortega, “On dependent bit allocation for multiview image coding with depth-image-based rendering”, *IEEE Transactions on Image Processing*, vol.20, no. 11, pp. 3179-3194, May. 2011.

Laura Toni *received the M.S. and Ph.D. degrees, both in electrical engineering, from the University of Bologna, Italy, in 2005 and 2009, respectively. Between 2009 and 2011, she worked at the Tele-Robotics and Application (TERA) department at the Italian Institute of Technology (IIT), investigating wireless sensor networks for robotics applications.*

In 2012, she was a Post-doctoral fellow in the Electrical and Computer Engineering Department at the UCSD. Between 2013 and 2016, she was a Post-doctoral fellow in the Signal Processing Laboratory (LTS4) led by Prof. Pascal Frossard at the Swiss Federal Institute of Technology (EPFL), Switzerland. In 2016, she has been appointed as Lecturer in the Electronic and Electrical Engineering Institute of University College London (UCL).

Thomas Maugey *graduated from Supélec, France in 2007. He received the M.Sc. degree from Supélec and Université Paul Verlaine, Metz, France, in 2007.*

He received his Ph.D. degree in Image and Signal Processing at TELECOM ParisTech, Paris, France in 2010. His supervisors were Béatrice Pesquet-Poposecu and Marco Cagnazzo.

From October 2010 to October 2014, he was a postdoctoral researcher at LTS4 of EPFL, headed by Pascal Frossard. Since November 2014, he is a Research Scientist at Inria Rennes-Bretagne-Atlantique. He works in the team SIROCCO headed by Christine Guillemot.

[5] J. Chakareski, V. Velisavljevic, and V. Stankovic, "View-popularity-driven joint source and channel coding of view and rate scalable multi-view video", *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no.3, pp. 474-486, Apr. 2015.

[6] A. De Abreu, G. Cheung, P. Frossard, and F. Pereira, "Optimal Lagrange multipliers for dependent rate allocation in video coding", *arXiv:1603.06123*, Mar. 2016.

[7] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo, and J. Wen, "Novel tile segmentation scheme for omnidirectional video", *Proc. of IEEE International Conference on Image Processing (ICIP)*, Phoenix, USA, Sep. 2016.

[8] A. Ghosh, V. Aggarwal, and F. Qian, "A rate adaptation algorithm for tile-based 360-degree video streaming", *arXiv:1704.08215*, Apr. 2017.

[9] F. De Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding", *Proc. of IEEE Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016.

[10] T. Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles", *Proc. ACM Conf. on Multimedia systems (MMSys)*, pp. 133-144, San José, USA, Feb. 2011.

[11] L. Toni, T. Maugey, and P. Frossard, "Optimized packet scheduling in multiview video navigation systems", *IEEE Transactions on Multimedia*, vol. 17, no.9, pp. 1604-1616, Sep. 2015.

[12] A. Hamza, and M. Hefeeda, "A DASH-based free viewpoint video streaming system", *Proc. of ACM Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV)*, pp. 55, Singapore, Singapore, Mar. 2014.

[13] M. Zhao, X. Gong, J. Liang, J. Guo, W. Wang, X. Que, and S. Cheng, "A cloud-assisted DASH-based scalable interactive multiview video streaming framework", *Proc. of Picture Coding Symposium (PCS)*, Cairns, Australia, Jun. 2015.

[14] M. Hosseini, and V. Swaminathan, "Adaptive 360 VR video streaming based on MPEG-DASH SRD", *Proc. IEEE International Symposium on Multimedia (ISM)*, San José, USA, Dec. 2016.

[15] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-Adaptive Navigable 360-Degree Video Delivery", *Proc. of IEEE Conference on Communications*, Paris, France, May 2017.

[16] L. Wang, D. Dai, J. Jiang, T. Yang, X. Jiang, Z. Cai, Y. Li, and X. Li "FISF: Better User Experience using Smaller Bandwidth for Panoramic Virtual Reality Video", *arXiv:1704.06444*, Apr. 2017.

[17] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, "Modeling the time—Varying subjective quality of HTTP video streams with rate adaptations", *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2206-2221, May 2014.