Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2017

Axe Sophi@Stic :	SystèmesRéseaux|
Titre du sujet :	Streaming Virtual Reality: Learning for Attentional Models and Network Optimization
Mention de thèse :	Informatique
HDR Directeur de thèse inscrit à l'ED STIC :	Frédéric Precioso

Co-encadrant de thèse éventuel :

Nom :	Sassatelli
Prénom :	Lucile
Email :	sassatelli@i3s.unice.fr
Téléphone :	0492942772

Email de contact pour ce sujet :	sassatelli@i3s.unice.fr
Laboratoire d'accueil :	I3S

Description du sujet :

Virtual Reality (VR) has taken off in the last two years thanks to the democratization of affordable head-mounted displays (HMDs), giving rise to a new market segment along with sizable research and industrial challenges. The technology is indeed still nascent and immature, entailing poor to downright sickening experience. The reasons are multiple, and can be cast into two categories.

The 360° videos can be stereoscopic to create a 3D effect, but those on main distribution platforms (e.g., YouTube 360) are mostly monoscopic to date. While the absence of a 3D effect limits the sense of immersion, it also prevents a major hurdle to the proper rendering of stereoscopic views in near-eye displays, which lies in the vergence/accommodation conflict.
Possible solutions rely on so-called foveated rendering, where regions not in the gaze's target are

blurred away to reproduce the natural focusing process in a real scene, thereby lowering the visual discomfort and the cognitive load. This requires eye-tracking, planned to be integrated in the next generation of HMDs [1].
On the other hand, the streaming delivery of VR over the Internet is highly challenging owing to required rates much higher than for conventional videos (about 28Mbps [2,3,4] to 5.2 Gbps for an artifact-free VR experience with sight only [5]).
A simple principle is to send the non-visible part of the sphere with lower quality. That is enabled by MPEG DASH-SRD, which splits the sphere into pre-defined tiles and encodes them into pre-defined qualities. The question is then how to allocate the bandwidth resource by choosing which quality to possibly send for tile. Existing works mostly adapt the strategies that were devised in the field of panoramic and zoomable videos [6].

The challenge is to reconcile the goals of allowing the client to buffer several seconds of video to absorb the network instability, while at the same time adapting to the moving user's Field of View (FoV). This is all the more challenging because head motions are faster than fingers zooming or panning a video, and gaze is faster than head motions.
In [6], the attempt to predict the FoV based on crowd-sourcing and ARMA was shown inefficient. The question of chunk replacement after download to keep the buffer content up to date with the user's motion is also poorly explored [7].

The question we want to answer in this PhD thesis is therefore: How to structure the VR content in the cloud and optimally decide what to transmit to maximize the streaming quality?
Our strategy is to devise a comprehensive processing chain tailored to VR and targeted at streaming.

The first contribution is the first component of the chain: predicting the user's gaze with the design of attentional models taking account of the specific and multiple modalities of the VR experience.

The probability of where a user is going to look in an image or a video is provided by saliency maps, obtained with supervised learning or a prioris on the Human Visual System [8].
In the last three years, the number of new models of visual and sound attention using Deep Learning (DL) techniques has soared [9,10] and their accuracy revolutionized the attentional modeling field for static images [11], with only few recent preliminary extensions to video [12,13,14]. The user experience of watching a 360° video in a headset with hearphones is however a wholly different experience than a legacy video on a TV screen, and the sound is a major way to draw the user attention in VR. Existing attentional models for the sound are however much fewer mainly oriented at speech recognition [15,16]. Last April Facebook disclosed its new VR streaming strategy (not yet deployed) based on the DL for prediction, but restricted to crowd-sourced data on watched tiles statistics (not necessarily watched in a HMD) and not gaze positions [17].

The second contribution is the second component of the chain: structure the VR content to design a cloud transcoder and decide what to send and when, given the network state.

First, the storing overhead with DASH (due to the different qualities available) and difficulty to choose the quality has led to a sequence of works since 2013 showing the benefit of continuous instead of discrete encoding rate adaptation [18-20].
Second, splitting the sphere into pre-defined tiles with the same quality over a given tile entails bandwidth waste when only a part of a tile is rendered in the FoV, and the loss of compression ration increases with the number of tiles (as the tiles are made independent and inter-tile correlation cannot be exploited by the coder in H.264 or H.265).
Departing from approaches discretizing the quality and spatial domains, we instead devise a smooth transcoder, which will encode on the fly the content tailored to a selected delivery. Inspired from [18], a main idea to define the structure of the content is to leverage the solutions of foveated rendering, such as [21], which allow to decrease the required processing power at the PC tethered to the headset, and turn these computational gains into bandwidth gains. From there, we will design online decision algorithms fed by the gaze prediction and the network feedback to control the transcoder and the transmission process to the client. This type of online control based on feedback and stochastic prediction can write as Markov Decision Problems (MDPs) for which a manifold of resolution frameworks can be leveraged, from drift-plus-penalty to reinforcement learning (such as Q-learning or Time Difference Learning combining Monte Carlo and Dynamic Programming). We will build on these models to properly formalize the optimization problem of transcoder control and transmission scheduling. We will thereby derive adaptive buffering and replacement algorithms to allow the client's buffer to build while keeping most up-to-date with respect to the FoV.

References
https://mycore.core-cloud.net/public.php?service=files&t=2c64ca9d769ad195f422040cfb863e8c

URL : https://mycore.core-cloud.net/public.php?service=files&t=2c64ca9d769ad195f422040cfb863e8c

English version:

Virtual Reality (VR) has taken off in the last two years thanks to the democratization of affordable head-mounted displays (HMDs), giving rise to a new market segment along with sizable research and industrial challenges. The technology is indeed still nascent and immature, entailing poor to downright sickening experience. The reasons are multiple, and can be cast into two categories.

The 360° videos can be stereoscopic to create a 3D effect, but those on main distribution platforms (e.g., YouTube 360) are mostly monoscopic to date. While the absence of a 3D effect limits the sense of immersion, it also prevents a major hurdle to the proper rendering of stereoscopic views in near-eye displays, which lies in the vergence/accommodation conflict.
Possible solutions rely on so-called foveated rendering, where regions not in the gaze's target are blurred away to reproduce the natural focusing process in a real scene, thereby lowering the visual discomfort and the cognitive load. This requires eye-tracking, planned to be integrated in the next generation of HMDs [1].
On the other hand, the streaming delivery of VR over the Internet is highly challenging owing to required rates much higher than for conventional videos (about 28Mbps [2,3,4] to 5.2 Gbps for an artifact-free VR experience with sight only [5]).

A simple principle is to send the non-visible part of the sphere with lower quality. That is enabled by MPEG DASH-SRD, which splits the sphere into pre-defined tiles and encodes them into pre-defined qualities. The question is then how to allocate the bandwidth resource by choosing which quality to possibly send for tile. Existing works mostly adapt the strategies that were devised in the field of panoramic and zoomable videos [6].

The challenge is to reconcile the goals of allowing the client to buffer several seconds of video to absorb the network instability, while at the same time adapting to the moving user's Field of View (FoV). This is all the more challenging because head motions are faster than fingers zooming or panning a video, and gaze is faster than head motions.
In [6], the attempt to predict the FoV based on crowd-sourcing and ARMA was shown inefficient. The question of chunk replacement after download to keep the buffer content up to date with the user's motion is also poorly explored [7].

The question we want to answer in this PhD thesis is therefore: How to structure the VR content in the cloud and optimally decide what to transmit to maximize the streaming quality?
Our strategy is to devise a comprehensive processing chain tailored to VR and targeted at streaming.

The first contribution is the first component of the chain: predicting the user's gaze with the design of attentional models taking account of the specific and multiple modalities of the VR experience.

The probability of where a user is going to look in an image or a video is provided by saliency maps, obtained with supervised learning or a prioris on the Human Visual System [8].
In the last three years, the number of new models of visual and sound attention using Deep Learning (DL) techniques has soared [9,10] and their accuracy revolutionized the attentional modeling field for static images [11], with only few recent preliminary extensions to video [12,13,14]. The user experience of watching a 360° video in a headset with hearphones is however a wholly different experience than a legacy video on a TV screen, and the sound is a major way to draw the user attention in VR. Existing attentional models for the sound are however much fewer mainly oriented at speech recognition [15,16]. Last April Facebook disclosed its new VR streaming strategy (not yet deployed) based on the DL for prediction, but restricted to crowd-sourced data on watched tiles statistics (not necessarily watched in a HMD) and not gaze positions [17].

The second contribution is the second component of the chain: structure the VR content to design a cloud transcoder and decide what to send and when, given the network state.

First, the storing overhead with DASH (due to the different qualities available) and difficulty to choose the quality has led to a sequence of works since 2013 showing the benefit of continuous instead of discrete encoding rate adaptation [18-20].
Second, splitting the sphere into pre-defined tiles with the same quality over a given tile entails bandwidth waste when only a part of a tile is rendered in the FoV, and the loss of compression

ration increases with the number of tiles (as the tiles are made independent and inter-tile correlation cannot be exploited by the coder in H.264 or H.265).
Departing from approaches discretizing the quality and spatial domains, we instead devise a smooth transcoder, which will encode on the fly the content tailored to a selected delivery. Inspired from [18], a main idea to define the structure of the content is to leverage the solutions of foveated rendering, such as [21], which allow to decrease the required processing power at the PC tethered to the headset, and turn these computational gains into bandwidth gains. From there, we will design online decision algorithms fed by the gaze prediction and the network feedback to control the transcoder and the transmission process to the client. This type of online control based on feedback and stochastic prediction can write as Markov Decision Problems (MDPs) for which a manifold of resolution frameworks can be leveraged, from drift-plus-penalty to reinforcement learning (such as Q-learning or Time Difference Learning combining Monte Carlo and Dynamic Programming). We will build on these models to properly formalize the optimization problem of transcoder control and transmission scheduling. We will thereby derive adaptive buffering and replacement algorithms to allow the client's buffer to build while keeping most up-to-date with respect to the FoV.

References
https://mycore.core-cloud.net/public.php?service=files&t=2c64ca9d769ad195f422040cfb863e8c

URL : https://mycore.core-cloud.net/public.php?service=files&t=2c64ca9d769ad195f422040cfb863e8c