UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

Network Support for AR/VR and Immersive Video Application: A Survey

Permalink

https://escholarship.org/uc/item/8zm6t24f

Journal

Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018, 1

Authors

Garcia-Luna-Aceves, J.J.
He, D.
Westphal, C.

Publication Date

2018-07-26

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Network Support for AR/VR and Immersive Video Application: A Survey

Dongbiao He[1], Cedric Westphal[2] and J. J. Garcia-Luna-Aceves[2]
[1]*Department of Computer Science and Technology, Tsinghua University, Bejing, China*
[2]*Department of Computer Engineering, University of California, Santa Cruz, Santa Cruz, CA, U.S.A.*
hdb13@mails.tsinghua.edu.cn. {cedric, jj}@soe.ucsc.edu

Keywords: AR/VR, 360 Degree Video, Field of View, QoE, Survey.

Abstract: Augmented Reality and Virtual Reality are rapidly gaining attention and are increasingly being deployed over the network. These technologies have large industrial potential to become next big platform with a wide range of applications. This experience will only be satisfying when the network infrastructure is able to support these applications. Current networks however, are still having a hard time streaming high quality videos. The advent of 5G Networks will improve the network performance, but it is unclear it will be sufficient to provide new applications delivering augmented reality and virtual reality services. There are few surveys on the topic of augmented reality systems, and their focus mostly stays on the actual displays and potential applications. We survey the literature on AR/VR networking, and we focus here on the potential underlying network issues.

1 INTRODUCTION

With the continuous development of Augmented Reality (AR) and Virtual Reality (VR) technologies, new challenges have beed arisen in network area for supporting these applications. Facebook and YouTube have already deployed support for some immersive videos, including 360 degree videos. These technologies are still in their infancy but many believe they have huge potential to shape the next experience for entertainment, education and commerce. Forecasters suggest around 30 million devices will be sold by 2020 generating revenue of around $21 billion (Newman, 2017).

Specialized head mounted display (HMD) devices such as the Oculus Rift and HTC Vive hit the market in 2016 and have been used mostly in gaming applications. A more mainstream experiences have come from much cheaper wraparound containers for smartphones such as Google Cardboard and the Galaxy Gear VR headset. While the very first commercial deployments, such as Google Glass for augmented reality, were not as successful as hoped, new products on the market keep trying to deliver an enhanced experience to users (Cass and Choi, 2015).

This experience will only be satisfying when the network infrastructure will be able to support these applications. Current networks however, are still having a hard time streaming high quality videos. The advent of 5G Networks will improve the network to better provide for new applications delivering augmented reality and virtual reality services. The 5G white paper (Alliance, 2015) specifically mentions augmented reality, 3D-video and pervasive video as use cases for dense urban networks. Yet, it is unclear that without architectural support in the network, such applications will receive the required resource to be succesful with consumers.

There are few surveys on the topic of augmented reality systems (say, (Van Krevelen and Poelman, 2010)), and their focus mostly stays on the actual displays and potential applications. (Westphal, 2017) listed some challenges for the network to support AR/VR and proposed Information-Centric Network as an architectural answer. We focus here on the potential underlying network issues. We present some of these issues in this survey.

We attempt to survey the literature to see how to better deliver an immersive experience at the network layer. There are different threads in this work. One key issue to efficiently deliver a satisfying immersive experience is to deliver what the user is viewing with high quality. Namely, the viewport of the user should be reliably provided in high resolution. However, delivering every possible view at high quality imposes a prohibitive cost in bandwidth. One answer is to *only* deliver the views that are actually seen at high quality.

In Section 2, we consider what can be done to *predict* the motion of the user's view, so as to deliver only the minimal amount of data that encompasses what

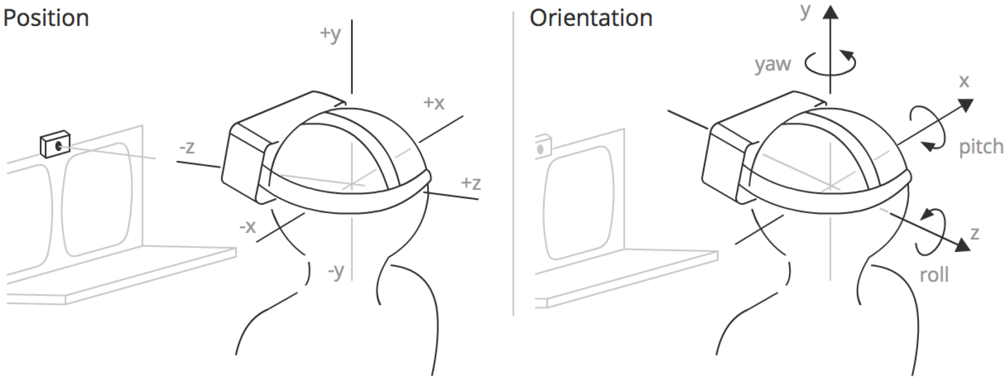

Figure 1: The head movement (web, 2018).

the user will be looking at in the immersive stream. In Section 3, we look at the *compression and coding* schemes to reduce the bandwidth footprint of the application.

In Section 4, we study how to deliver an application stream that is viewport dependent. Section 5 considers the potential caching strategies at the network layer to enhance immersive applications. Section 6 lists some of the empirical results and the datasets that can be used for testing new enhancements. Finally, Section 7 offers some concluding remarks.

2 PREDICTION OF THE USER'S FIELD OF VIEW

360 degree videos (Sometimes refers as omnidirectional videos), also known as immersive videos or spherical videos, are video recordings where a view in every direction is recorded at the same time, shot using an omnidirectional camera or a collection of cameras. During playback on normal flat display the viewer has control of the viewing direction like a panorama. It can also be played on a displays or projectors arranged in a cylinder or some part of a sphere.

Every day, these videos are becoming more and more ubiquitous due to the dramatic increase in the popularity of the Internet services such as social networking, e-commerce websites, e-learning websites etc. Determining the particular user's characteristics from its video stream is crucial to save bandwidth and improve the QoE for the users.

As displayed in figure 1, 360 degree videos use the position sensors to detect viewing information from the HMD. This allows the user to continually update a scene according to the head movement, rotation, etc. Prefetching strategies should carefully balance two contrasting objectives, namely maximizing quality and avoiding stalls in the played stream and prefetch new content from different channels to provide a seamless switch.

Zhang Hui (Zhang and Chen, 2016) predicts user service access information (such as occurrence time, occurrence location, service type and service content) in the next period by studying the user's behavior. They formulate a modified entropy weighted *Markov model* to accurately predict the user's service states with an adaptive feedback based weight correction method.

Andrew Kiruluta(Kiruluta et al., 1997) uses a *Kalman filtering model* to predict head movements. The key idea is to recognize that head movement trajectories can be assumed to be constrained on piecewise constant acceleration paths with an additional noise component to account for any perturbation from these paths. (Vintan et al., 2006) proposes *neural prediction techniques* to anticipate a person's next movement. They focus on neural predictors (multi-layer perceptron with back-propagation learning) with and without pre-training.

Mavlankar and Girod (Mavlankar and Girod, 2010) perform fixation prediction in videos using features like thumbnails, motion vectors, and navigation trajectories. With the advanced machine learning technologies, various supervised learning methods including *neural networks* are adopted for better feature extraction and prediction accuracy in fixation detection (Alshawi et al., 2016), (Chaabouni et al., 2016), (Nguyen et al., 2013). Chaabount et al. (Chaabouni et al., 2016) build a convolutional neural networkks (*CNN*) architecture and use residual motion as the features for predicting saliency in videos.

Alshawi et al. (Alshawi et al., 2016) observe the correlation between the eye-fixation maps and the spatial/temporal neighbors, which provides another way to quantify viewer attention on videos. Nguyen et al. (Nguyen et al., 2013) propose to adopt the information of static saliency (in images) and then take camera motions into considerations for dynamic saliency (in videos) prediction. (Fan et al., 2017) addresses the problem of fixation prediction for 360 degree video streaming to HMDs using two neural networks.

(Bao et al., 2016) are used for predicting head movement in 360 degree video delivery. They collect motion data for some subjects watching 360 degree videos. From the collected data, they observe a strong short-term auto-correlation in viewer motions, which indicates that viewer motion can be well predicted based on motion history.

In (Liu et al., 2017), they employ a Field-of-View (FoV) guided approach that fetches only the portions of a scene the users will see. They also use big data analytics to facilitate accurate head movement prediction (HMP), a key prerequisite for FoV-guided streaming. In addition, they propose enhancements that allow 360 degree videos to be efficiently streamed over multiple network paths such as WiFi and cellular.

3 VIDEO COMPRESSION AND CODING

Adaptive streaming of 360 degree video content in a virtual reality setting is a challenging task which requires smart encoding and streaming techniques to cope with today's and future application and service requirements. Video compression standards aim to minimize the spatiotemporal redundancies by exploiting the characteristics of human visual systems along with source coding techniques from information theory. Moreover, a client can periodically switch to neighboring captured views as the video is played back in time when watching 360 degree video. The technical challenge is to design coding structure to facilitate periodic view switching, while providing some level of error resiliency, so that error propagation due to differentially coded frames can be mitigated. Hence, most of coding strategies would be viewport-dependent for saving resource cost in transmission the video stream.

Begole (Begole, 2016) calculates that human can process 5.2Gbps of data based on the physical characteristics of human perception. This amount is beyond even 5G network support, and is calculated based upon the ability to distinguish 200 dots per degree within the typical human foveal field of view, with at least 150 degrees horizontally and 120 degrees vertically at a rate of 30 frames per second. However, this number assumes a compression ratio and some dedicated coding techniques.

To address this bandwidth scarcity problem, many 360 degree video streaming service providers have been actively working to address the concerns in encoding and transmitting 360 degree videos. Much of this effort has gone into encoding schemes that reduce the amount of information transmitted over the network during streaming. For example, (Rondao-Alface et al., 2017) proposed an end-to-end system that could achieve real-time streaming of 360 content at 16K resolution with very low latency over Wi-Fi or LTE towards untethered, smartphone-based HMDs. In (Yu et al., 2015), they put forward a framework for evaluating the coding efficiency in the context of viewing with a head-mounted display.

360 degree video requires increased video resolution (4K and beyond) to support the wider field of view (360 degree field of view vs regular 2D video that typically covers around 60-90 degrees field of view). Hence the bitrate requirements are also higher necessitating the need for efficient compression techniques for 360 degree video. The rest of this section is the literatures focusing on the video encoding in the system of multiview video streaming (IMVS) or 360 degree videos.

3.1 Encoding Overhead Optimization

(Graf et al., 2017) states that the state of the art video codecs (AVC/H.264, HEVC/H.265, VP8, VP9) and delivery formats (DASH/HLS) can be used to deploy a basic adaptive streaming service of omnidirectional video content. However, this is very inefficient as the typical Field of View of many VR devices is limited, and a lot of content is delivered, decoded, and rendered for nothing (e.g., what is happening outside of the users' FoV). Viewport adaptive streaming has been introduced to overcome this limitation but requires multiple versions of the same content for each view. That is, it adopts a similar strategy as in adaptive media streaming (DASH/HLS) but the number of versions of the same content significantly increases. This impacts (cloud) storage and (content delivery) network costs.

In (Cheung et al., 2011), they motivated the need for an interactive multi-view video streaming system, where an observer can periodically send feedback to the server requesting the desired views out of the many available ones. In response, the server will send the corresponding pre-encoded video data to the observer for decoding and display without interrupting the forward video playback. Observing that one can trade off expected transmission rate with a modest increase in storage when designing the pre-encoded frame structure, they formulated a combinatorial optimization problem, where the optimal structure contains the best mixture of I-frames (for random access), P-frames (for low transmission rate) and merge or M-frames (for low storage), trading off transmission rate with storage.

In (Liu et al., 2013), they proposed a new frame type called unified distributed source coding (uDSC) frame that can both facilitate view switching and halt error propagation. They then optimized transmission strategies for coding structures with uDSC frames for wireless IMVS multicast and wired IMVS unicast scenarios.

(Ozcinar et al., 2017a) targets both the provider's and client's perspectives and introduces a new content-aware encoding ladder estimation method for tiled 360 degree VR video in adaptive streaming systems. The proposed method first categories a given 360 degree video using its features of encoding complexity and estimates the visual distortion and resource cost of each bitrate level based on the proposed distortion and resource cost models. An optimal encoding ladder is then formed using the proposed integer linear programming (ILP) algorithm by considering practical constraints.

Adaptivity to the current user viewport is a promising approach but incurs significant encoding overhead when encoding per user or set of viewports. A more efficient way to achieve viewport adaptive streaming, presented in (Skupin et al., 2016), is to facilitate motion-constrained HEVC tiles. Original content resolution within the user viewport is preserved while content currently not presented to the user is delivered in lower resolution. A lightweight aggregation of varying resolution tiles into a single HEVC bitstream can be carried out on-the-fly and allows usage of a single decoder instance on the end device.

(Zhang et al., 2018) studied a navigation-aware adaptation logic optimization problem for interactive free viewpoint video systems that is able to minimize both the navigation distortion and the view-switching delay. They provide an optimal solution based on a dynamic programing (DP) algorithm with polynomial time complexity, and an approximate algorithm with effective performance to further reduce computational complexity in practice. The algorithm properly takes into account both video content characteristics and user interactivity level, which is efficient to find the proper tradeoff between view quality and number of reference views in constrained resource networks.

(Corbillon et al., 2017b) investigated some theoretical models for the preparation of 360 degree video for viewport adaptive streaming systems. They explored the interplay between the parameters that characterize the video area in which the quality should be better. They denote this special video area a Quality-Emphasized Region (QER) and address a theoretical model based on the fundamental trade-off between spatial size of the QERs and the aggregate video bitrate.

In (Sreedhar et al., 2016), they put forward a multi-resolution viewport adaptive projection schemes to measure the rate-distortion (RD) performance between different projection schemes. With the observation of their evaluations, the multi-resolution projections of Equirectangle and Cubemap outperform other projection schemes, significantly.

3.2 Resource Allocation

(De Abreu et al., 2015) proposed a novel adaptive transmission solution that jointly selects the optimal subsets of views streams and rate allocation per view for a hierarchical transmission in IMVS applications. They consider a system where the network is characterized by clients with heterogeneous bandwidth capabilities, and they aim to minimize their expected navigation distortion. To do so, clients are clustered according to their bandwidth capabilities and the different camera views are distributed in layers to be transmitted to the different groups of users in a progressive way. The clients with higher capabilities receive more layers (more views), hence benefiting of a better navigation quality. They have formulated an optimization problem to jointly determine the optimal arrangement of views in layers along with the coding rate of the views, such that the expected rendering quality is maximized in the navigation window, while the rate of each layer is constrained by network and clients capabilities. To solve this problem, they have proposed an optimal algorithm and a greedy algorithm with a reduced complexity, both based on dynamic programming.

(Chakareski et al., 2015) studied the scenario of multicasting to a collection of clients. To address the issue of heterogeneity amongst clients, they designed a scalable joint source and channel coding scheme for which they formulated a view-popularity-driven source-channel rate allocation and a view packing optimization problem that aims at maximizing the expected video quality over all clients, under transmission rate constraints and the clients' view selection preferences. Their system is superior to state-of-the-art reference systems based on H.264/SVC and the channel coding technique they designed, and H.264 and network coding. Finally, they developed a faster local-search-based method that still optimizes the source and channel coding components of their system jointly at lower complexity. It exhibits only a marginal loss relative to an exhaustive-search optimization.

(Rossi and Toni, 2017) proposed an optimal transmission strategy for virtual reality applications that is able to fulfill the bandwidth requirements, while optimizing the end-user quality experienced in the navigation. In further detail, they consider a tile-based coded content for adaptive streaming systems, and they propose a navigation-aware transmission strategy at the client side (i.e., adaptation logic), which is able to optimize the rate at which each tile is downloaded.

In (Chen et al., 2017), the problem of resource management is studied for a network of wireless virtual reality users communicating over small cell networks (SCNs). In order to capture the VR users' quality of service (QoS), a novel VR model, based on multi-attribute utility theory, is proposed. This model jointly accounts for VR metrics such as tracking accuracy, processing delay, and transmission delay.

4 VIEWPORT-DEPENDENT STREAMING

Virtual Reality devices are quickly becoming accessible to a large public. It is, therefore, expected that the demand for 360 degree immersive videos will grow consistently in the next years. In VR streaming, the user is immersed in a virtual environment and can dynamically and freely decide the preferred viewing position, called viewport. Unfortunately, VR streaming is often affected by low quality nowadays, due to the high bandwidth requirements of 360 degree videos. Viewport-dependent solutions have often been proposed for VR streaming, as they are able to reduce the bandwidth required to stream the VR video. Viewport-adaptive streaming has recently received a growing attention from both academic and industrial communities.

(Afzal et al., 2017) characterized 360 degree videos from the point of view of network streaming, and compared them to regular videos that have been the popular media format until now. Their comparison shows that 360 degree videos have substantially higher bit rates and a larger number of resolution formats; however, they also find that the bit rates for the 360 videos have less variability than regular videos, which can be highly beneficial for the network due to the network provisioning for the peak rates. To explain lower bit rate variability, they demonstrated that the average motion for the 360 video is less than that for a comparable regular video. They believe that this is because the motion described in a 360 degree video is that which is inherently in the scene, rather than the rotation or panning of the camera in space. This implies that the panning now occurs at the time of user viewing the video. Thus, the new requirement on the network is that it needs to be more responsive to the user changing field of view. They believe these aspects have deep implications on networked video streaming systems for both capacity and latency requirements.

A 360 multiview video streaming (IMVS) systemdegree video is captured in every direction from a unique point, so it is essentially a spherical video. Since current video encoders operate on a 2D rectangular image, a key step of the encoding chain is to project the spherical video onto a planar surface (showed in figure 2). The four projections that are most discussed for 360 degree video encoding are rectangular, cube map, pyramid and rhombic dodecahedron. From the images that are projected on these projections, it is possible to generate a viewport for any position and angle in the sphere without information loss. However, some pixels are over-sampled, in the case of rectangular mapping, resulting in degradation of performance of video encoders. On the contrary, the projection into a pyramid layout causes under-sampling. This results in distortion and information loss in some extracted viewports.

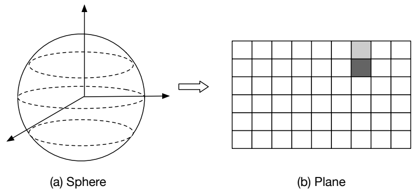

Figure 2: Tiles in 360 degree video.

(Nguyen et al., 2017a) analyzed and evaluated the impacts of the response delay to tiling-based viewport-adaptive streaming for 360 degree videos. Preliminary results indicate that viewport-adaptive streaming methods are effective under short response delay only. Specifically, for viewport adaptive methods to outperform EQUAL when the frame rate is 30fps, the buffer sizes in cases of the adaptation intervals of 4, 32, and 64 frames should be less than 1s, 0.5s, and 0.125s, respectively. When the buffer size exceeds 2s, EQUAL, which is a viewport-independent method, outperforms all considered viewport-adaptive methods. *So, viewport adaptive streaming seems to be ineffective when using existing HTTP Streaming solutions due to long response delay*.

He et al (He et al., 2018) looked at the joint adaptation of the field of view with the rate under variations of the network delay and the congestion. They proposed an algorithm to adapt the size of the field of view to be downloaded based upon the network and buffering delay to retrieve and view the stream segment so that it encompasses with high likelihood the user's viewport at the time of rendering.

4.1 QoE-driven Solution

Quality of Experience (QoE) is a measure of the delight or annoyance of a customer's experiences with a service such as web browsing, phone call and TV broadcast (wik, 2018). QoE is one of major factors for optimization 360 degree video streaming service. However, the users' QoE is far from trivial with the adaptive streaming strategies. In this paper, we list some papers with QoE guided solution for 360 degree video streaming transmission, and in particular, we investigate some literatures for bandwidth saving in section 4.2.

(Ghosh et al., 2017) proposed a novel adaptive streaming scheme for 360 degree videos. The basic idea is to fetch the unviewed portion of a video at the lowest quality based on users' head movement prediction, and to adaptively decide the video playback quality for the visible portion based on bandwidth prediction. Instead of using a robust manner requires overcome a series of challenges, they first define QoE metrics for adaptive 360 degree video streaming, formulating an optimization problem with a low complexity solution. The algorithm strategically leverages both future bandwidth and the distribution of users' head positions to determine the quality level of each tile. After that, they further provide theoretical proof showing that their algorithm achieves optimality under practical assumptions.

(Xie et al., 2017) leveraged a probabilistic approach to pre-fetch tiles so as to minimize viewport prediction error, and design a QoE-driven viewport adaptation system, 360ProbDASH. It treats user's head movement as probabilistic events, and constructs a probabilistic model to depict the distribution of viewport prediction error. A QoE-driven optimization framework is proposed to minimize total expected distortion of pre-fetched tiles. Besides, to smooth border effects of mixed-rate tiles, the spatial quality variance is also minimized. With the requirement of short-term viewport prediction under a small buffer, it applies a target buffer-based rate adaptation algorithm to ensure continuous playback.

(TaghaviNasrabadi et al., 2017) argued that playback is prone to video freezes which are very destructive for the Quality of Experience. They propose using layered encoding for 360 degree video to improve QoE by reducing the probability of video freezes and the latency of response to the user head movements. Moreover, this scheme reduces the storage requirements significantly and improves in-network cache performance.

(Corbillon et al., 2017c) targets at the problem of bandwidth waste: the Field of View (or viewport) is only a fraction of what is downloaded, which is an omnidirectional view of the scene. To prevent simulator sickness and to provide good Quality of Experience, the vendors of HMDs recommend that the enabling multimedia systems react to head movements as fast as the HMD refresh rate. Since the refresh rate of state-of-the-art HMDs is 120 Hz, the whole system should react in less than 10 ms. This delay constraint prevents the implementation of traditional delivery architectures where the client notifies a server about changes and awaits for the reception of content adjusted at the server. Instead, in the current Virtual Reality video delivery systems, the server sends the full 360 degree video stream, from which the HMD extracts the viewport in real time, according to the user head movements. Therefore, the majority of the delivered video stream data are not used.

4.2 Bandwidth Saving

(Graf et al., 2017) described the usage of tiles — as specified within modern video codecs such HEVC/H.265 and VP9 — enabling bandwidth efficient adaptive streaming of omnidirectional video over HTTP with various streaming strategies. Therefore, the parameters and characteristics of a dataset for omnidirectional video are proposed and exemplary instantiated to evaluate various aspects of such an ecosystem, namely bitrate overhead, bandwidth requirements, and quality aspects in terms of viewport PSNR.

In (Hosseini and Swaminathan, 2016b) and (Hosseini and Swaminathan, 2016a), they proposed a dynamic view-aware adaptation technique to tackle the huge bandwidth demands of 360 VR video streaming. They spatially divide the videos into multiple tiles while encoding and packaging, use MPEG-DASH SRD to describe the spatial relationship of tiles in the 360 degree space, and prioritize the tiles in the Field of View.

(Le Feuvre and Concolato, 2016) described how spatial access can be performed in an adaptive HTTP streaming context, using tiling of the source content, MPEG-DASH and its SRD extensions. They describe a configurable implementation of these technologies, within the GPAC open-source player, allowing experimentations of different adaptation policies for tiled video content.

(Ozcinar et al., 2017b) introduced a novel end-to-end streaming system from encoding to displaying, to transmit 8K resolution 360 degree video and to provide an enhanced VR experience using Head Mounted Displays (HMDs). The main contributions of the proposed system are about tiling, integration of the MPEGDynamic Adaptive Streaming over HTTP (DASH) standard, and viewport-aware bitrate level selection. Tiling and adaptive streaming enable the proposed system to deliver very high-resolution 360 degree video at good visual quality. Further, the proposed viewport-aware bitrate assignment selects an optimum DASH representation for each tile in a viewport aware manner.

Nguyen. et al (Nguyen et al., 2017b) also display a very similar solution with tiling in order to save the bandwidth, which tackles the aforementioned problems by proposing a novel adaptive tile-based streaming method over HTTP/2. In order to solve the bandwidth problem in streaming VR videos over HTTP/2, a dynamic adaptation method is crucial. In this paper, they propose a novel adaptive streaming method based on tiled streaming. By using H.265 standard, a video at the server is divided into spatial tiles, each of which is subdivided into multiple temporal segments. In order to support adaptive streaming method from client, each tile is also encoded into different versions. The priority of tiles is defined based on the user's viewport. Then, the priority feature of HTTP/2 is used to request the server to push the tiles of higher priority first.

In (Mangiante et al., 2017), they present a Field Of View rendering solution at the edge of a mobile network, designed to optimize the bandwidth and latency required by VR 360 degree video streaming.

(Qian et al., 2016) also argues that fetching the entire raw video frame wastes bandwidth. They consider the problem of optimizing 360 degree video delivery over cellular networks. They first conducted a measurement study on commercial 360 degree video platforms, then proposed a cellular-friendly streaming scheme that delivers only the 360 degree videos' visible portion based on head movement prediction.

In (Petrangeli et al., 2017b), they presented a novel framework for the efficient streaming of VR videos over the Internet, which aims to reduce the high bandwidth requirements and storage costs of current VR streaming solutions. The video is spatially divided in tiles using H.265, and only tiles belonging to the user viewport are streamed at the highest quality.

To save bandwidth, (Petrangeli et al., 2017a) spatially tiled the video using the H.265 standard and streamed only tiles in view at the highest quality. The video is also temporally segmented, so that each temporal segment is composed of several spatial tiles. In order to minimize quality transitions when the user moves, an algorithm is developed to predict where the user is likely going to watch in the near future. Consequently, predicted tiles are also streamed at the highest quality. Finally, the server push in HTTP/2 is used to deliver the tiled video. Only one request is sent from the client; all the tiles of a segment are automatically pushed from the server.

5 IN-NETWORK CACHING

Historically, the Internet has evolved in an adhoc manner where incremental patches were added to handle new requirements as they arose. This means that the underlying network model has not changed over the last decades, while the services using the Internet did so drastically. Information Centric Networking (ICN) is a network architecture that evolves from the traditional host-oriented communication model to a content centric model, which can be extremely beneficial in adaptive streaming (Westphal (Editor) et al, 2016). Particularly, ICN relies on location-independent naming schemes, in-network pervasive caching, and content-based routing to allow an efficient distribution of content over the network. Moreover, ICN nodes can seamlessly use all the available network interfaces to retrieve content. Content Centric Networking (CCN) and Named Data Networking (NDN) are typical instantiations of the ICN paradigm. (Zhang et al., 2017) proposed a VR video conferencing system over named data networks (NDN). (Westphal, 2017) made the case that ICN could answer some of the issues of AR/VR networking support. However, other caching solutions may help as well.

Edge computing is expected to be an effective solution to deliver 360 degree virtual reality videos over networks. Cache is one of underlying resources for enabling these applications. Mangiante et al. (Mangiante et al., 2017) present a Field Of View rendering solution at the edge of a mobile network, designed to optimize the bandwidth and latency required by VR 360 degree video streaming. Jacob Chakareski et al. (Chakareski, 2017) designed an optimization framework that allows the base stations to select cooperative caching/rendering/streaming strategies that maximize the aggregate reward they earn when serving the users, given specific caching and computational resources at each base station. Zhang Liyang et al. In (Matsuzono et al., 2017), they propose a low latency, low loss streaming mechanism, L4C2, convenient for for high-quality real-time delay-sensitive streaming. L4C2 is also built with in-network caching mechanism.

In particular, (Yeo et al., 2017) argues that client's increasing computation power and advancement in deep neural networks allow us to take advantage of *long-term redundancy* found in videos, which leads to quality improvement at lower bandwidth cost for Internet video delivery.

6 PUBLIC DATASETS

Recently, content and datasets for 360 degree video have been made public so as to promote reproducible research. Some researchers have built 360 degree video testbeds for collecting traces from real viewers watching 360 degree videos using HMDs. The collected datasets can be used in various 360 degree video applications with viewers using HMDs.

As for researchers, they can: (i) analyze the datasets for finding some key problems and get the insights for the research work, (ii) apply the datasets to validate and evaluate their systems and algorithms. The followings are some collected datasets.

(Corbillon et al., 2017a) presented a dataset including the head positions of 59 users recorded while they are watching five 70 s-long 360 degree videos using the Razer OSVR HDK2 HMD. They have published the dataset on the website alongside the used videos and the open-source software that they developed to collect the dataset. Finally they also introduced examples of statistics that can be extracted from the dataset to provide an overview of the users' behaviour and the videos characteristics, focusing on the viewport adaptive streaming scenario.

(Rai et al., 2017) presents a new dataset of 60 omnidirectional images with the associated head and eye movement data recorded for 63 viewers. A subjective experiment was conducted, in which they are asked to explore the images for 25 seconds, as naturally as possible. In addition, an image/observer agnostic analysis of the results from this experiment is also performed, which, in contrast to previous studies, considers both head and eye tracking data. Furthermore, they also provide guidelines and tools to the community to evaluate and compare saliency maps in such omnidirectional scenarios. They argue that the free availability of this dataset to the research community may help on the intensive research work that is being done nowadays regarding immersive media technologies to provide the best possible QoE to the end users.

(Lo et al., 2017) presented the dataset collected from ten YouTube 360 degree videos and 50 subjects. Their dataset is unique, because both content data, such as image saliency maps and motion maps, and sensor data, such as positions and orientations, are provided. Many 360 degree video streaming applications, both traditional ones (like R-D optimization) and novel ones (like crowd-driven camera movements) can benefit from their comprehensive dataset. In addition, the dataset can be extended in several ways, such as adding eye tracking data of the eyes movement as hints for future head movement.

The dataset presented by CIVIT (civ, 2018) provides four or eight different fisheye views generated from a Nokia OZO. This allows testing of multiple algorithms like depth estimation or panoramas stitching.

In (Wu et al., 2017), they present a head tracking dataset composed of 48 users (24 males and 24 females) watching 18 sphere videos from 5 categories. For better exploring user behaviors, they record how users watch the videos, how their heads move in each session, what directions they focus, and what content they can remember after each session.

7 CONCLUDING REMARKS

AR/VR and immersive video streaming push the boundaries of the network capability. The upcoming 5G roll-out will not alleviate all of the issues and therefore these applications need both in-network support as well as enhancements at the application layer to be successfully deployed. As such, it is a very active area of research. We attempted to depict this research landscape in this document.

We have seen how user's motion prediction could assist with reducing the bandwidth; how coding and compression schemes are being developed; how tiling and FoV can be adapted to the network conditions; how caching will assist with the deployment of such applications; and what datasets are currently available to researchers who would like to test new methods and algorithms.

Future research directions should include methods to improve performance in three directions:

- bandwidth consumption should be minimized by further improving prediction of the user's behavior, improved compression schemes, sharing and multicasting using efficient tiling, etc.
- delay responsiveness of the network should be improved; 5G will significantly reduce the network RTT, but reducing the segment length (together with the associated coded schemes), bringing the servers to the edge, providing QoS for immersive application, improving the hardware, etc, are all required.

- reducing the computational impact on the network; in order to support transcoding, or processing of the AR/VR uplink streams (for sensor and position data and for image and pattern recognition), the computing at the edge will significantly increase. Methods to minimize this impact, methods for better pattern recognition and for sharing this processing among users, etc, still need to be devised.

ACKNOWLEDGEMENTS

Dongbiao He is supported by the Tsinghua Scholarship for Overseas Graduate Studies.

REFERENCES

(2018). 360-degree stereo video test datasets - civit. http://www.tut.fi/civit/index.php/360-dataset/. Accessed: 2018-05-28.

(2018). Quality of experience. https://en.wikipedia.org/wiki/Quality_of_experience.

(2018). Webvr concepts. https://developer.mozilla.org/. Accessed: 2018-06-13.

Afzal, S., Chen, J., and Ramakrishnan, K. (2017). Characterization of 360-degree videos. In *Proceedings of the Workshop Viewing Dataset in Head-Mounted on Virtual Reality and Augmented Reality Network*, pages 1–6. ACM.

Alliance, N. (2015). NGMN 5G White Paper. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf.

Alshawi, T., Long, Z., and AlRegib, G. (2016). Understanding spatial correlation in eye-fixation maps for visual attention in videos. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6. IEEE.

Bao, Y., Wu, H., Zhang, T., Ramli, A. A., and Liu, X. (2016). Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 1161–1170. IEEE.

Begole, B. (2016). Why the internet pipe will burst if virtual reality takes off. http://www.forbes.com/sites/valleyvoices/2016/02/09/why-the-internet-pipes-will-burst-if-virtual-reality-takes-off/. Last accessed September 14th, 2016.

Cass, S. and Choi, C. (2015). Google Glass, HoloLens, and the Real Future of Augmented Reality. In *IEEE Spectrum*.

Chaabouni, S., Benois-Pineau, J., and Amar, C. B. (2016). Transfer learning with deep networks for saliency prediction in natural video. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1604–1608. IEEE.

Chakareski, J. (2017). Vr/ar immersive communication: Caching, edge computing, and transmission trade-offs. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 36–41. ACM.

Chakareski, J., Velisavljevic, V., and Stankovic, V. (2015). View-popularity-driven joint source and channel coding of view and rate scalable multi-view video. *IEEE Journal of Selected Topics in Signal Processing*, 9(3):474–486.

Chen, M., Saad, W., and Yin, C. (2017). Resource management for wireless virtual reality: Machine learning meets multi-attribute utility. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–7. IEEE.

Cheung, G., Ortega, A., and Cheung, N.-M. (2011). Interactive streaming of stored multiview video using redundant frame structures. *IEEE Transactions on Image Processing*, 20(3):744–761.

Corbillon, X., De Simone, F., and Simon, G. (2017a). 360-degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 199–204. ACM.

Corbillon, X., Devlic, A., Simon, G., and Chakareski, J. (2017b). Optimal set of 360-degree videos for viewport-adaptive streaming. *in Proc. of ACM Multimedia (MM)*.

Corbillon, X., Simon, G., Devlic, A., and Chakareski, J. (2017c). Viewport-adaptive navigable 360-degree video delivery. In *Communications (ICC), 2017 IEEE International Conference on*, pages 1–7. IEEE.

De Abreu, A., Toni, L., Thomos, N., Maugey, T., Pereira, F., and Frossard, P. (2015). Optimal layered representation for adaptive interactive multiview video streaming. *Journal of Visual Communication and Image Representation*, 33:255–264.

Fan, C.-L., Lee, J., Lo, W.-C., Huang, C.-Y., Chen, K.-T., and Hsu, C.-H. (2017). Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 67–72. ACM.

Ghosh, A., Aggarwal, V., and Qian, F. (2017). A rate adaptation algorithm for tile-based 360-degree video streaming. *arXiv preprint arXiv:1704.08215*.

Graf, M., Timmerer, C., and Mueller, C. (2017). Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 261–271. ACM.

He, D., Westphal, C., and Garcia-Luna-Aceves, J. (2018). Joint rate and fov adaptation in immersive video streaming. In *ACM Sigcomm workshop on AR/VR Networks*.

Hosseini, M. and Swaminathan, V. (2016a). Adaptive 360 VR video streaming based on MPEG-DASH SRD. In *ISM*, pages 407–408. IEEE Computer Society.

Hosseini, M. and Swaminathan, V. (2016b). Adaptive 360 vr video streaming: Divide and conquer. In *Multime-*

dia (ISM), 2016 IEEE International Symposium on, pages 107–110. IEEE.

Kiruluta, A., Eizenman, M., and Pasupathy, S. (1997). Predictive head movement tracking using a kalman filter. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(2):326–331.

Le Feuvre, J. and Concolato, C. (2016). Tiled-based adaptive streaming using mpeg-dash. In *Proceedings of the 7th International Conference on Multimedia Systems*, page 41. ACM.

Liu, X., Xiao, Q., Gopalakrishnan, V., Han, B., Qian, F., and Varvello, M. (2017). 360 innovations for panoramic video streaming.

Liu, Z., Cheung, G., and Ji, Y. (2013). Optimizing distributed source coding for interactive multiview video streaming over lossy networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1781–1794.

Lo, W.-C., Fan, C.-L., Lee, J., Huang, C.-Y., Chen, K.-T., and Hsu, C.-H. (2017). 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 211–216. ACM.

Mangiante, S., Klas, G., Navon, A., GuanHua, Z., Ran, J., and Silva, M. D. (2017). Vr is on the edge: How to deliver 360 videos in mobile networks. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 30–35. ACM.

Matsuzono, K., Asaeda, H., and Turletti, T. (2017). Low latency low loss streaming using in-network coding and caching. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pages 1–9. IEEE.

Mavlankar, A. and Girod, B. (2010). Video streaming with interactive pan/tilt/zoom. *High-Quality Visual Experience, Signals and Communication Technology*, pages 431–455.

Newman, N. (2017). Journalism, media, and technology trends and predictions 2017.

Nguyen, D. V., Tran, H. T., and Thang, T. C. (2017a). Impact of delays on 360-degree video communications. In *TRON Symposium (TRONSHOW), 2017*, pages 1–6. IEEE.

Nguyen, M., Nguyen, D. H., Pham, C. T., Ngoc, N. P., Nguyen, D. V., and Thang, T. C. (2017b). An adaptive streaming method of 360 videos over http/2 protocol. In *Information and Computer Science, 2017 4th NAFOSTED Conference on*, pages 302–307. IEEE.

Nguyen, T. V., Xu, M., Gao, G., Kankanhalli, M., Tian, Q., and Yan, S. (2013). Static saliency vs. dynamic saliency: a comparative study. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 987–996. ACM.

Ozcinar, C., De Abreu, A., Knorr, S., and Smolic, A. (2017a). Estimation of optimal encoding ladders for tiled 360° vr video in adaptive streaming systems. *arXiv preprint arXiv:1711.03362*.

Ozcinar, C., De Abreu, A., and Smolic, A. (2017b). Viewport-aware adaptive 360° video streaming using tiles for virtual reality. *arXiv preprint arXiv:1711.02386*.

Petrangeli, S., De Turck, F., Swaminathan, V., and Hosseini, M. (2017a). Improving virtual reality streaming using http/2. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 225–228. ACM.

Petrangeli, S., Swaminathan, V., Hosseini, M., and Turck, F. D. (2017b). An http/2-based adaptive streaming framework for 360°virtual reality videos. In *ACM Multimedia*, pages 306–314. ACM.

Qian, F., Ji, L., Han, B., and Gopalakrishnan, V. (2016). Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, pages 1–6. ACM.

Rai, Y., Gutiérrez, J., and Le Callet, P. (2017). A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 205–210. ACM.

Rondao-Alface, P., Aerts, M., Tytgat, D., Lievens, S., Stevens, C., Verzijp, N., and Macq, J. (2017). 16k cinematic VR streaming. In *ACM Multimedia*, pages 1105–1112. ACM.

Rossi, S. and Toni, L. (2017). Navigation-aware adaptive streaming strategies for omnidirectional video. In *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, pages 1–6. IEEE.

Skupin, R., Sanchez, Y., Hellge, C., and Schierl, T. (2016). Tile based hevc video for head mounted displays. In *Multimedia (ISM), 2016 IEEE International Symposium on*, pages 399–400. IEEE.

Sreedhar, K. K., Aminlou, A., Hannuksela, M. M., and Gabbouj, M. (2016). Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications. In *ISM*, pages 583–586. IEEE Computer Society.

TaghaviNasrabadi, A., Mahzari, A., Beshay, J. D., and Prakash, R. (2017). Adaptive 360-degree video streaming using layered video coding. In *Virtual Reality (VR), 2017 IEEE*, pages 347–348. IEEE.

Van Krevelen, D. and Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality*.

Vintan, L., Gellert, A., Petzold, J., and Ungerer, T. (2006). Person movement prediction using neural networks.

Westphal, C. (2017). Challenges in networking to support augmented reality and virtual reality. In *IEEE ICNC*.

Westphal (Editor) et al, C. (2016). Adaptive Video Streaming in Information-Centric Networking (ICN). IRTF RFC7933, ICN Research Group.

Wu, C., Tan, Z., Wang, Z., and Yang, S. (2017). A dataset for exploring user behaviors in VR spherical video streaming. In *MMSys*, pages 193–198. ACM.

Xie, L., Xu, Z., Ban, Y., Zhang, X., and Guo, Z. (2017). 360probdash: Improving qoe of 360 video streaming using tile-based http adaptive streaming. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 315–323. ACM.

Yeo, H., Do, S., and Han, D. (2017). How will deep learning change internet video delivery? In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, pages 57–64. ACM.

Yu, M. C., Lakshman, H., and Girod, B. (2015). A framework to evaluate omnidirectional video coding schemes. In *ISMAR*, pages 31–36. IEEE Computer Society.

Zhang, H. and Chen, J. (2016). A novel user behavior prediction and optimization algorithm for single-user multi-terminal scenario. In *Computational Intelligence and Design (ISCID), 2016 9th International Symposium on*, volume 2, pages 144–147. IEEE.

Zhang, L., Amin, S. O., and Westphal, C. (2017). Vr video conferencing over named data networks. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 7–12. ACM.

Zhang, X., Toni, L., Frossard, P., Zhao, Y., and Lin, C. (2018). Adaptive streaming in interactive multiview video systems. *IEEE Transactions on Circuits and Systems for Video Technology*.