CHAPTER

1 Streaming and User Behaviour in Omnidirectional Videos

Silvia Rossi[a], Alan Guedes[a], and Laura Toni[a]

[a]*Department of Electrical & Electrical Engineering, UCL, London (UK),*
{s.rossi,a.guedes,l.toni}@ucl.ac.uk

ABSTRACT

Omnidirectional videos (ODVs) have gone beyond the passive paradigm of traditional video, offering higher degrees of immersion and interaction. The revolutionary novelty of this technology is the possibility for users to interact with the surrounding environment, and to feel a sense of engagement and presence in a virtual space. Users are clearly the main driving force of immersive applications and consequentially the services need to be properly tailored to them. In this context, this chapter highlights the importance of the new role of users in ODV streaming applications, and thus the need for understanding their behaviour while navigating within ODVs. A comprehensive overview of the research efforts aimed at advancing ODV streaming systems is also presented. In particular, the state-of-the-art solutions under examination in this chapter are distinguished in terms of system-centric and user-centric streaming approaches: the former approach comes from a quite straightforward extension of well-established solutions for the 2D video pipeline while the latter one takes the benefit of understanding users' behaviour and enable more personalised ODV streaming.

KEYWORDS

omnidirectional video, user behaviour, ODV streaming

1.1 Introduction

Over the past few years, the synergistic development of new mobile communication services (*i.e.*, 5G mobile networks) and new cutting-edge portable devices (*i.e.*, smartphones) have helped for a breakthrough in video streaming services. In this context, the concept of *immersive and interactive communication* is spreading, identifying a completely novel way of communicating with other people and displaying multimedia content. Traditional remote communications (*e.g.*, television, radio, video calling) are no more sufficient tools for our society: humans are inherently social, in need of realistic experiences, and traditional remote communications do not offer such full sense of immersion and a natural experience/interactions [1]. The impact of realistic experience in remote communications would interest society on wide levels as it addresses a compelling need in reducing environmental impact, in enabling remote working, answering also natural emergencies needs (*e.g.*, reduced travel in pandemic, tornadoes, etc.). In this context, Virtual Reality (VR) is an example of immersive technology which has already landed in our everyday life , with an impact ($21.83 Billion in 2021, with a projection of growth to $69.6 Billion by 2028 [2]) across major economic sectors beyond entertainment, *e.g.*, , e-healthcare, e-education, and cultural heritage [3].

VR technology refers to a fully digital environment that replaces the real world and in which the user is **immersed**, allowing the user to experience a completely new reality. The revolutionary novelty introduced by VR and immersive technology at large is indeed to provide viewers with the possibility to interact with the digital surrounding environment, and with feelings of engagement and presence in a virtual space, even if they are not physically there [4,5]. Specifically, **presence** refers to the illusory feeling experienced by the user of being in a virtual environment different from the physical one where they are actually located [4]. A condition necessary for presence is the immersion which refers more to technical properties of the system that are needed to simulate a realistic virtual environment [6]. **Interactivity** is instead the possibility for users to change the virtual environment with their movements [7]. Thus, the role of user interaction is crucial to “be present” in the virtual world: being able to move naturally helps the illusion of being in a different place. Novel types of multimedia content are therefore needed to ensure a sufficient level of immersion, presence, and interactivity, which are the three crucial factors to guarantee high Quality of Experience (QoE) in a VR system [8,9]. In particular, *omnidirectional videos* (ODVs) (also named 360° or spherical videos), which captures a 360° scene instantaneously, are emerged as a new type of media content that promises an immersive and interactive experience. The viewer is placed at the centre of the virtual space (*i.e.*, viewing sphere) and provided with a VR device – typically a Head-Mounted Display (HMD) – he/she experiences a 3-Degree of Freedom (DoF) interaction with the content, by looking up/down (pitch) or left/right (view) or by tilting their head from side to side (roll), as shown in Figure 1.1 a. These head rotations enable free navigation within the immersive scene as the user displays *only* a restricted Field of View (FoV) of the environment around themselves, named the *viewport*, identified by the viewing direction at any given time (Figure 1.1 b). Hence, the sequence of the user viewing direction over time can be approximated by projecting the viewport centre on the viewing sphere (Figure 1.1 c) and it can be used to identify the user behaviour in an immersive experience. The new type of navigation within the video content shows a clear evolution of the user’s role from merely passive in traditional video applications into interactive consumers in VR systems.

The new format as well as the new way of consuming the content open the gate to many promising VR applications, but also pose completely new challenges. One key challenge raised from the interactivity level is the high resolution and low-latency required to ensure a full sense of presence. The user needs to have ultra-low switching delays when changing the displayed viewport to avoid discomfort. This can be ensured

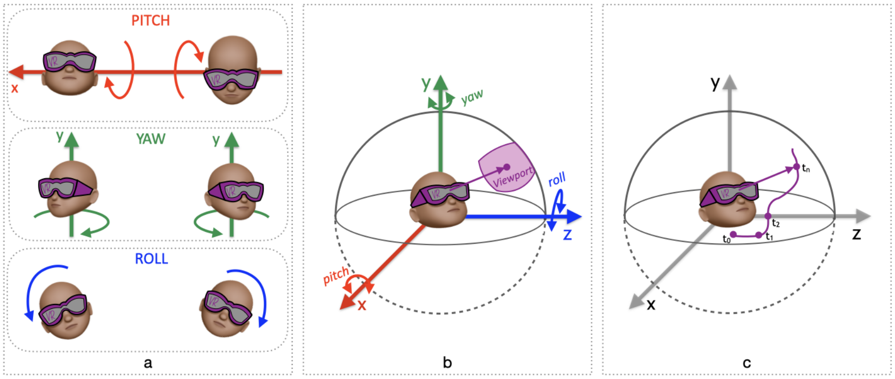


FIGURE 1.1

Navigation in immersive content: a) rotational head movements (pitch, yaw and roll); b) navigation system on viewing sphere at generic time instant; c) navigation trajectory over time.

by sending to all users the entire content at high-quality, assuming that the desired viewport will be then exported during rendering. This solution is the first one that has been proposed, extending the well-established and optimized methods for 2D videos to spherical content. These methodologies are usually user-agnostic and aimed at improving the overall performance through enhancements in the system. The main drawback of these solutions is that they are extremely bandwidth-consuming (up to 43 MB/s for 8K spherical video [10]) since the entire ODV content is delivered to final clients, pushing the available bandwidth to the limit, with a negative impact on the final quality. In practice, only a small portion (typically around 15% of the entire video [11]) of the overall content is displayed by the user, making these solutions extremely inefficient. Recently, more attention has been put on the final users, leading to personalized systems, which put the user at the center of the system and tailor every aspect of the coding–delivery–rendering chain to the viewer interaction. For example, only the predicted content of interest for the final user is pushed into the delivery network [11]. The main aim of these personalized systems is to optimise the user QoE but also to overcome ODV streaming limitations, such as reducing bandwidth and storage usage. However, this comes at the price of requiring the knowledge of user interactivity patterns in advance. Therefore, inspired by recommender system and 5G network communities [12–14], we propose to name the first line of improvements as *system-centric streaming* while the second one as *user-centric streaming*. Figure 1.2 shows this proposed taxonomy for ODV streaming solutions and all subcategories that will be under examination in the following.

In this chapter, we provide an overview of the research efforts that have been dedicated to advance ODV streaming strategies, with a specific attention to the more recent user-centric systems. Due to this popularity, many surveys papers [15–24] have been published already to summarise the main contributions to ODV streaming

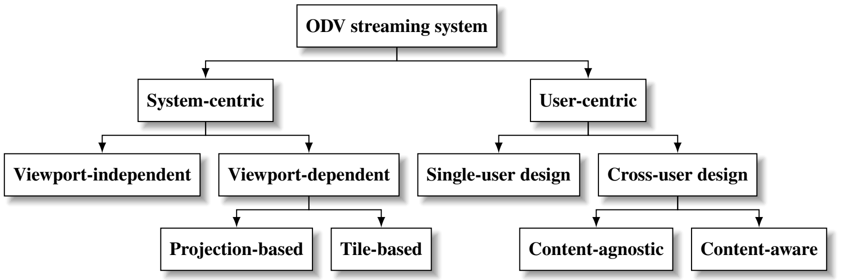


FIGURE 1.2

Proposed taxonomy of researches related with ODV streaming systems.

systems. Table 1.1 depicts these works visualizing their main topics of interest, highlighting also the level of investigation across the end-to-end pipeline.

As it is evident from Table 1.1, the majority of existing surveys are deeply focused on compression, delivery and quality assessment aspects. For instance, Zink *et al.* [18] have provided a general overview of the main challenges and the first attempts of solution per each step of the ODV streaming pipeline, from acquisition to the final user rendering experience. Chen *et al.* [15] have mainly explored the most recent projections methods aimed at improving video coding and transmissions, and reducing video quality distortions. More insights on system design and implementations have been described in [17,20,23]. In particular, both Fan *et al.* [17] and Yaqoob *et al.* [20] have examined existing protocols and standards together with optimal ODV streaming solutions. Ruan *et al.* [23] have instead investigated solutions for VR systems but mainly from a network services perspective. Visual quality artefacts have been deeply investigated by Azevedo *et al.* [19] describing their sources and features at each step of the system; authors have also presented an overview of existing tools for quality assessment (objective and subjective). Similarly, a deep focus on visual quality assessment, together with attention models and compression, is given by Xu *et al.* [22]. Authors have highlighted the importance of predicting where viewers mainly put their attention during immersive navigation (*i.e.*, saliency maps) to benefit the entire system since users are the final consumers. They have also partially addressed the need of understanding behavioural features to help in modelling user attention presenting the main outcomes from existing navigation dataset analysis. Following this direction, the recent work presented by Chiarotti [24] has showed the importance of estimating navigation path also for quality evaluation, neglecting however the behavioural analysis. To the best of our knowledge, these are the only existing surveys which explicitly brings out the importance of the new role of users in ODV streaming applications, and thus the need of understanding their behaviour. As shown in Table 1.1, behavioural analysis has been highly overlooked. One of the main contributions of this chapter is to fill in this gap by discussing in-depth the role of the user in ODV streaming strategies.

The remaining of this chapter is structured as follows: Section 1.2 describes recent advances in video streaming towards supporting ODV, from services popularisation to

Table 1.1 Surveys related with ODV streaming systems. Level of investigation per each topic: ■ mentioned; ■■ sufficient; ■■■ deep.

Survey	Acquisition	Compression	Delivery	Rendering	Quality Assessment	Prediction	Behavioural Analysis
Chen *et al.* [15]		■■■	■■		■■■		
He *et al.* [16]		■■	■■			■■	
Fan *et al.* [17]	■■	■■■	■■■		■■		
Zink *et al.* [18]	■	■	■	■	■		
Azevedo *et al.* [19]	■■■	■■	■	■■■	■■		
Yaqoob *et al.* [20]		■■	■■■		■■	■■	
Shafi *et al.* [21]		■■■	■■■	■■	■■■		
Xu *et al.* [22]		■■■	■■		■■■	■■	■
Ruan *et al.* [23]	■	■	■	■			
Chiarotti [24]		■■	■■■		■■■	■■■	
Our chapter		■■	■■■			■■■	■■■

standardising. Section 1.3 provides an overview of coding and delivery strategies for the system-centric streaming, especially introducing differences between viewport-independent and viewport-dependent methods. Section 1.4 highlights the role of the user behaviour in ODV streaming and lists datasets for ODV user analyses. Section 1.5 describes how such novel interactivity can drastically improve the status quo using user-centric streaming. To conclude, we present final remarks and highlight new directions in Section 1.6.

1.2 Streaming Pipeline: Evolution Towards ODV

In this section, we provide an overview of the ODV streaming pipeline. We start with an historical overview of ODV to contextualize the first steps that opened the gate to ODV streaming research, which is the main focus of this current chapter, we then conclude by explaining how key components of the streaming pipeline, from acquisition to rendering, have evolved and have been standardized to enable ODV services.

Table 1.2 depicts the historical evolution that led to current technology used on ODV systems. This evolution has been characterised by three key components: 1) large-scale utilization of ODV applications; 2) ODV displaying technology; 3) technological advances in the streaming pipeline. The first service that appeared in 2007 based on omnidirectional content was the Google Maps Street View, which allows users to virtually navigate on a street using a sequence of omnidirectional images [25]. After this, the ODV market has grown significantly mainly when YouTube and Facebook (and Vimeo) allowed the upload and share of 360-degree content on their platforms in 2015 (in 2017) [29,30,36]. The interest in ODV systems then has been grown exponentially: for example, BBC and the French cultural network ARTE used 360-video for immersive documentaries. Now, 360-degree content is widely used across multiple sectors (*e.g.*, e-culture, entertainment, retail, live sports) amplified even further from recent attention to metaverse applications. This widespread

Table 1.2 ODV streaming historical timeline

2007	Google launches Street View [25]
2011	*ISO publishes MPEG-DASH* [26]
2014	Google launches Cardboard and Facebook acquired Oculus. [27,28]
2015	YouTube and Facebook social platforms allow ODV upload [29,30]
	MPEG standardised MPEG-DASH SRD to support tiled streaming [31]
2016	BBC and ARTE begin sharing ODV content [32,33]
	Facebook proposes Pyramid projection [34]
	MPEG standardised HEVC encoding with Motion Constrained Tile Set [35]
	First viewport prediction algorithms
2017	Vimeo platform allows ODV upload [36]
	YouTube proposes equiangular cubemap projection [37]
	First user navigation datasets for User Behavioural Analyses
2019	*MPEG standardised OMAF* [38]
2021	Facebook promotes VR towards metaverse applications [39]

of ODV services was further pushed by the advances on HMDs: in 2014 Google proposed a affordable mobile-based HMD called Cardboard, while Facebook made a two-billion-dollar acquisition of the HMD company Oculus. This has led to an ever-growing desire for the users to experience ODV systems, highlighting the compelling need for research advances and even standardising steps on ODV streaming pipeline. The well known MPEG-dynamic adaptive streaming over HTTP (DASH)–de-facto streaming solutions standardised in 2011 [26]– has been improved to enable ODV systems (italic and blue text in Table 1.2). Moreover, new sphere-to-plane projections were proposed from Facebook and YouTube, namely pyramid (2016) and equiangular cubemap (2017), to map the spherical content into 2D domain.

In parallel, DASH streaming was extended to the tile-based encoding that has played a key role in viewport dependent streaming. The video content is spatially cropped in different bitstreams named tiles, each of those independently coded from the other tiles, allowing for unequal quality levels [40]. This is possible thanks to the *HEVC Motion-Constrained Tile Set* (MCTS) technology [35], which eliminates dependencies between tiles, restricting the encoding of visual objects motion, also called motion vectors (MV), at tile boundaries. Tiles from different encoding quality can therefore be combined in a single HEVC bistream and the reconstructed bitstream is HEVC compliant and requires only a single decoder for the playback. The other key aspect of tile-based streaming is the DASH *Spatial Relationship Description* (SRD) [31], which enables the transmission of only a portion of the video to display devices. This, in combination with multi-quality tile-based coding allows us to send at high quality only the portion of interest to the VR user. This will be a key advance in viewport-dependent streaming technologies (discussed in Section 1.3.2).

These above DASH extensions were then consolidated on *Omnidirectional Media Format (OMAF)* [38], the first international standard for storage and distribution of

ODVs; a result of the efforts of MPEG, 3GPP and VR Industry Forum (VRIF)[41]. OMAF specifies tile-based streaming, ensuring that the OMAF player rewrites the encoded tiles syntax structures to merge them and decode them as one single bitstream [38,42]. Several tools have been proposed to experiment with OMAF: as encoding tools, the Kavazaar [43] that easily supports tile sets; as players, the Fraunhofer OMAF.js[1] and Nokia[2]. OMAF allowed the inclusion of other media types beyond ODVs such as still images, spatial audio, associated timed text, multiple viewpoints, and even a 2D video overlayed on the ODV [44]. These technological advances and standardization pushed research efforts to improve even further the ODV pipeline to achieve better services in terms of bandwidth, storage, networking caching, and perceived user quality. In the following, we describe the entire pipeline from acquisition to rendering to show how this has been adapted from classical 2D video to ODV streaming. Then, in the following sections, we provide an overview of the main technological advances mainly from the coding and streaming perspective. Initial efforts were mainly focused on system-centric streaming, see Section 1.3. At the same time, some researchers focused their studies on understanding how users navigate within VR content (bold and green text in Table 1.2). By analysing users' trajectories, these studies were able to develop viewport prediction algorithms (2016) and make them available to the community with datasets of user navigation patterns (2017). These datasets are intensely described in Section 1.4. These initial works opened the gate to user-centric system research, described in depth in Section 1.5.

1.2.1 Adaptive ODV Streaming Pipeline

We now describe each processing step of the ODV streaming pipeline, focusing mainly on MPEG-DASH [26][3], and highlighting the main novelties from classical video streaming pipeline to ODV. Figure 1.3 depicts the adaptive ODV streaming pipeline, with dash dotted green boxes characterizing the system-centric streaming (on what we are going to focus on at the moment) and solid line blue boxes identifying further evolution of ODV streaming toward user-centric ones – described in Section 1.5.

As already discussed in Chapter 3 of this book [46], the first step is the *Acquisition*, which is different from the 2D video counterpart due to the spherical nature of the media format. Most ODV cameras currently capture an entire 360-degree field of view using multi-sensor systems, in which the final picture is generated from multiple inputs signals. These signals are then processed by a stitching algorithm where the overlapping regions between the camera views are aligned and then warped in sphere surface. To be processed by existing 2D media processing tools, the spherical content is projected into a planar representation, which is usually called *Panorama*. This is

[1] `https://github.com/fraunhoferhhi/omaf.js`

[2] `https://github.com/nokiatech/omaf`

[3] It is worth mentioning that other ODV streaming protocols (not purely DASH based) have been proposed [45], however we mainly focus on DASH advances as this conceptually covers the majority of the works.

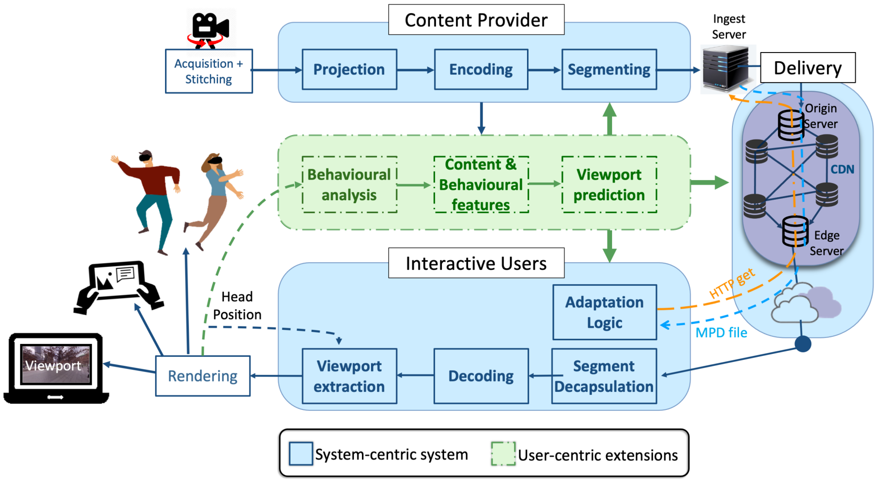
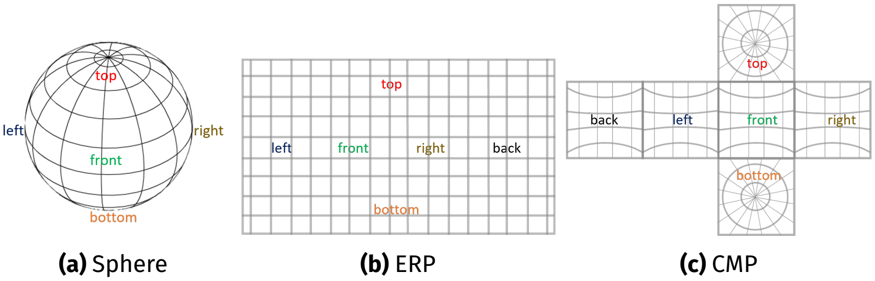


FIGURE 1.3

System-centric and User-centric ODV streaming pipeline.

the *Projection* step. The most commonly employed sphere-to-plane projections are the equirectangular projection (ERP) (Figure 1.4b) and cubemap projection (CMP) (Figure 1.4c) [47], which are supported by the OMAF standard. The ERP is the simplest and the most popular projection format , which maps the viewing sphere onto the panorama through the longitude and the latitude values. However, it is well known that this project suffers from an unequal sampling of points introduced in the pole areas, leading to more artifacts and less efficient compression and rendering tasks at the poles [19]. The other well known projection is the Cubemap, which has been introduced by the industrial sector initially for the gaming community. In this projection, the sphere is first mapped to a cube, which is then unfolded and each face is arranged into the panorama frame. While a lower distortion is achieved with CMP than with ERP, the project still suffer from unequal distribution of pixels (higher distribution towards the corners of the cube), which still affects the overall quality during rendering phase.

Once the content is projected into a 2D plane, it can be processed by the *Encoding* step, using the state-of-the-art codec from classical 2D media compression, such as High-Efficiency Video Coding (HEVC/H.265) [48]. Typically, in DASH, clients *pull* the content from the server, instead of being the server *pushing* it to clients. Specifically, each content is encoded into multiple resolutions and quality levels (*representations*), and each client dynamically selects the best representations when fetching video segments using HTTP requests. Hence, the encoding step produces multiple quality levels (representations). Each encoded representation is then segmented: the *Segmenting* step breaks the video into *temporal* chunks (usually 2s long) and stored at the server side. The different available representations of these



(a) Sphere (b) ERP (c) CMP

FIGURE 1.4

Widely adopted projections from a) spherical domain to plane: b) ERP and c) CMP.

chunks are described at the server, which in the case of DASH, is done on the DASH Media Presentation Description (DASH-MPD). The client then selects the most appropriate chunk representation to request, as explained later. DASH streaming was extended to the tile-based encoding that has played a key role in viewport dependent streaming. The video content is *spatially* cropped into different bitstreams named tiles, each of those encoded at a different coding rates and resolutions independently from the other tiles. This enables per-tile representations that are stored at the server [40], providing the client with the freedom to select unequal quality in the ODV.

The chunks created by the encoder are then ingested in a HTTP origin server that will process clients requests, *Delivery* step. This origin server is usually inside a content delivery network (CDN), which is an optimised infrastructure that redistributes chunks on multiple servers, ultimately reaching clients locations. At the client site, the chunks are processed by *Segment Decapsulation* to extract HEVC bitstream, which is then decoded in the *Decoding* step and fed into the playout buffer. In parallel, still at the decoder side, the *Adaptation Logic* dynamically decides the best representations to request for the upcoming chunk. This selection is based on the client connection and buffer condition as well as the device capabilities. The pipeline ends with *Rendering*, which back-projects the decoded planar representation in the spherical geometry and displays the content of interest to the final user. This content of interest (viewport) is evaluated in the prior step *Viewport extraction*, in which the current user viewing direction is translated into the displayed viewport.

1.3 System-centric streaming

Advances in the streaming pipeline were initially aimed at improving the overall system performance in terms of consumed bandwidth, storage cost, and networking reliability metrics. These streaming solutions are defined in this chapter as system-centric ones as they usually optimize systems design, being usually user-agnostic (neglecting user behaviour analyses and prediction). In the following, we further categorize system-centric solutions in viewport-independent and viewport-dependent streaming. *Viewport-independent* solutions are the most similar to traditional adaptive

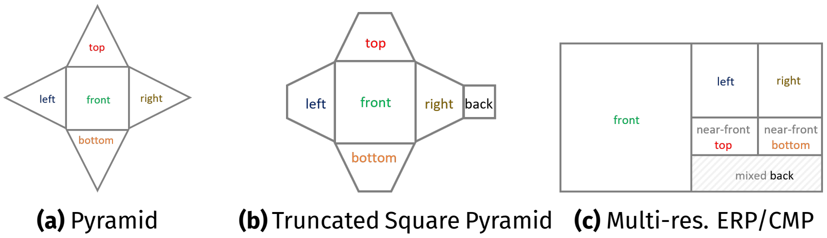
2D video streaming, in which the entire panorama is encoded and treated equally. Specifically, each representation available at the server side is encoded with a uniform quality and resolution across the entire panorama. As a consequence, for any final client to be able to display VR content at high-quality, the entire panorama will be downloaded at high quality. This ensures low latency for viewport-switching, but it is extremely costly in terms of storage and bandwidth usage [49]. Moreover, it assumes that the whole panorama is equally important, which is clearly not the case in VR systems. To strike the optimal balance between bandwidth waste and switching latency, ODV user interactivity needs to be taken into account, leading to streaming strategies adapting not only to the content but also to users, *viewport-dependent* streaming. The key assumption is that not the entire panorama is equally important since during the navigation users focus more on some areas than others. In 2012, Alface *et al.* [50] were the first ones to show that an ideal knowledge of the user interaction could lead to saving bandwidth, when transmitting only the viewport (region of interest to the user). In the following, we present in details advances on both viewport-dependent and viewport-independent streaming solutions.

1.3.1 Viewport-independent streaming

Beyond initial works that applied classical DASH streaming to ODVs [29], recent works that fall under the viewport-independent framework are mainly focused on improving the encoding step, overcoming issues related to sphere-to-plane projection. One goal is to encode the ODV in the spherical domain with no need to project the content into a bi-dimensional space. For instance, Vishwanath *et al.* [51] propose to encode the inter-frame motion vector (MV) directly on the sphere instead of the planar domain, avoiding distortions during the sphere-to-plane conversion. The MV is expressed as rotation on the sphere (rotational motion vector) and results demonstrate good gains compared to standard HEVC 2D video encoding. Also working on the sphere, Bidgoli *et al.* [52] propose a representation learning approach for spherical data. To avoid distortion from planar projection and yet exploit the benefit of deep learning for features extraction, the authors propose a new sphere convolution operation to keep the high expressiveness of 2D convolution while having the advantage of uniform sampling on the sphere. This can lead to better coding efficiency, as shown in the paper. These aforementioned approaches show promising use of an on-the-sphere encoding strategy for omnidirectional images.

1.3.2 Viewport-dependent streaming

Viewport-dependent solutions are based on the assumption that the content is not equally important (*spatially*), since users require to display only a portion of it. Therefore, it is clear that a non-uniform streaming should be considered to ensure higher quality to the areas more likely to be displayed. The adaptation to the viewport has been implemented in both the projection and the (tile-based) encoding step of the streaming pipeline. The former (*projection-based approach*) adapts the bit rate



(a) Pyramid **(b)** Truncated Square Pyramid **(c)** Multi-res. ERP/CMP

FIGURE 1.5

Viewport-dependent projections.

allocation during projection, making this an unequal allocation that prioritizes areas of interest for the final users. The latter (*tile-based approach*) has emerged when tiled based coding was proposed by HEVC encoding MCTS and most of the ODV works have focused afterwards on tile-based streaming for ODV. We detail these two approaches in the following.

Projection-based approach

Projection-based approaches (also called viewport-dependent projections [20]) aim at projecting the spherical content in such a way that the most important areas (most likely to be displayed) are the least distorted during the projection step of the pipeline. Selecting unequal levels of encoding rates across the panorama offers better QoE, but at a price of generating multiple versions of the same panorama (each version with a different area at high-quality). This impacts the storage cost (higher than the viewport-independent case) and the caching hit ratio (lower since users might prefer different versions depending on their focus of attention). Examples of viewport-dependent projections are the Pyramid projection (Figure 1.5a) and [34], the truncated square pyramid projection (TSP) (Figure 1.5b) [53]. The first one projects the sphere onto a pyramid, with an intentional unequal allocation of points – more data points to the pyramid base. The pyramid is then unwrapped and resized into 2D space (see Figure) and encoded with classical 2D codec tools. The novelty is that this projection encodes at the higher quality the area corresponding to the bottom face of the pyramid, which will be match with the viewing direction of the users. The projection however suffers from a rapid quality degradation when moving from the bottom face to the lateral ones. This is obviated in the TSP, a truncated version of the pyramid projection. Compared to a viewport-independent ERP, Facebook estimated that CMP has 25% improvement compression efficiency while Pyramid has a 80% gain [34]. With the ultimate goal of leaning toward more viewport-dependent projections, variations of the classical ERP and CMP have been proposed. Multi-resolution variants of ERP and CMP have been presented in [53], in which the content is first projected into the plane (or the cube) and then re-sized and scaled to fit the rectangular plane, as shown in the Figure 1.5c. In the final arrangement, some areas have a much larger scale and sampling, and this should cover ideally the viewport of interest to final users.

Tiled-based approach

Corbillon *et al.* [54] proposed an unequal encoding bitrates for ODVs, such that within the panorama one region of attention is encoded at higher quality. Each user would then select the version with the high-quality region overlapping with their viewport. This pillar work has introduced the concept that the content is not spatially equally important, opening the gate to tile-based approach. Today, it has been improved by current standardised tiled streaming using the HEVC MCTS, which encodes tiles at different bitrates and resolutions, creating per-tile representations that are stored at the server. The key novelty of a tile-based system is that tiles can be independently fetched by the client at the desired quality level. This creates a more flexible transmission strategy able to compose the user viewport with higher quality tiles. This higher level of flexibility however comes at the price of server processing and storage costs (a large number of representations encoded and made available at the main server), and larger computational costs for the final user to select the representation to request (larger search space in the adaptation logic). To overcome these limitations and profit the most from tile-based strategies, researchers have focused their studies on 1) optimal design of the tiles at the encoder side; or 2) optimised tiles distribution in the network delivery (*e.g.*, client adaptation, multi-path and caching).

The tile-based encoding strategy directly impacts the server-side storage costs, hence the compelling need to optimize it for omnidirectional content. Graf *et al.* [55] published one of the first works in this direction, showing that different tiling patterns have a different impact on the server-side costs, with final result that a 6×4 pattern shows most improvements. Researchers evolved then the studies toward viewport-dependent tile-based design [56,57]. Ozcinar *et al.* [56] design an optimal tiling scheme and the required encoding bitrate levels based on a probabilistic model that reflects the users interactivity, showing the gain in terms of users' QoE. Still optimizing the tiles at the server side but with more focus on the storage costs, Rossi *et al.* [57] propose an optimization problem at the server side aimed at selecting the optimal bitrate per tile to be stored at the main server to minimise the quality distortion at the user side while respecting the storage cost constraint. This optimization has been done considering non-rectangular and 6-tiles configuration and including users viewing direction probability in the optimization model.

Instead of optimizing the tile design at the server side, many other works have been focused on optimizing designs at the user side, studying optimal adaptation logic in viewport-dependent tile-based solutions. Ashan *et al.* [58] propose a buffer-based algorithm to optimize the bitrate of area in the panorama overlapping the users' viewport. The main challenge is that this does not simply imply selecting the optimal representation from the MPD, but it rather needs a mapping from tile's bitrate to viewport quality. Other efforts instead focused on adapting the strategy to users' interaction, exploiting either the knowledge of the viewing direction [59] or probabilistic/ average models such as the heatmap [60]. Fu *et al.* [59] propose an adaptation logic based on sequential reinforcement learning (RL), with the agent reward being the quality experienced by the users. The RL agent implicitly learns the interaction model and refines the adaptation logic based on the users' behaviour, showing a gain of 12% QoE improvement with respect to non viewport-dependent strategies. Rossi *et al.* [60] propose a navigation-aware adaptation logic using a probabilistic heatmap model. Other parallel researches have improved network aspects such as multi-path, pre-fetching, decoding offloading, and caching strategies. Duanmu *et al.* [61] developed a multi-path video streaming in heterogeneous WiFi and 5G networks. To efficiently utilize the high bandwidth available in 5G and maximize the rendered video quality, their client fetches videos from two networks. The 5G network was used to pre-fetch high-quality video chunks and re-transmit chunks that cannot be delivered in time through WiFi. Nguyen *et al.* [62] also address pre-fetching but using the new package prioritization feature from HTTP/2 to delivery high-quality tiles. Then, in a new direction, Nguyen *et al.* [63] investigated multi-path streaming but for decoding offloading. They consider a mobile VR device and a PC with access to the stream server and connected using a millimetre-wave (mmWave) network in the 60 GHz band. At the same time, tiles are fetched and decoded by devices, and the PC shares the decoded frames via mmWave to the mobile. Such collaborative decoding offloaded processing costs by 25%, speeding up of the decoding of all tiles. Finally, Maniotis *et al.* [64] improves the server cache hit ratio by using a RL framework based on tiles heatmap, showing an improvement of the proposed viewport-dependent caching mechanism against other caching schemes such as Least Frequently Used (LFU) and Least Recently Used (LRU).

The works mentioned so far have shown the gain of proposing viewport-dependent tile-based streaming and its impact across different steps of the pipeline. However, limitations still remain in the intrinsic definition of tile-based streaming. For example, the MCTS restricts the inter-frame prediction of visual objects between tiles, reducing the encoding efficiency. To obviate this limitation, Son *et al.* [65] propose new HEVC extensions to enable inter-frame prediction, when an encoded object has dependency from outside the tile using the HEVC scalability extension and extracting the object from an up-sampled version of the base layer. Another limitation of tile-based coding is the intra-frame redundancy, which is not minimized across tiles. Bidgoli *et al.* [66] address this limitation by allowing the encoder to detect reference regions outside the current fetched tiles. Then, they propose an encoding in which the intra-frame reference region is at the centre of a requested tile to improve client intra-frame decoding.

Despite these last limitations, viewport-dependent tile-based streaming has been widely adopted for ODV streaming. A key aspect across all these works is the user's information, expressed for example either via viewport trajectory or heatmap. The advances made so far have motivated researchers to dig deeper in the study of user's behaviour, as discussed in the following section.

1.4 The role of the user in ODV

In the following, we provide an overview on the recent ODV multimedia datasets made recently available to our community and the tool developed to perform behavioural data analysis from these datasets.

1.4.1 ODV navigation datasets

We provide an overview of the datasets collecting user's navigation data during immersive experiences. We summarize these datasets in Table 1.3, and highlight that they are limited to *i) publicly* available dataset, with data related to *ii)* ODV content (no images), and *iii)* navigation trajectories (*i.e.*, head and/or eye movements). In order to mimic a real-life condition, VR users cannot display the entire environment around themself but only a restricted portion (i.e., viewport). Specifically, the sequence of spatio-temporal points representing the user's viewing direction over time identifies users' attention within an immersive experience. Based on VR device technology, the user's viewing direction can be represented either by head or eye movements. The head movement determines the Field of View (FoV) as the pixel area of ODV, which is displayed by a given user over time, while the eye movement datasets contain the specific area within the FoV that captures the user attention and can be classified as salient.

The navigation trajectories are collected via a three-steps collection procedure: *1)* Test Preparation, *2)* Subjective Test, and *3)* Data Formatting and Storage. The selection of video content to be used during subjective experiments is one of the first steps. During *test preparation*, ODVs are selected based on several criteria: video length, number of video, content category, features and attributes. Considering the video sequence length, Ozcinar *et al.* present [71] the shortest group of video (*i.e.*, 10 sec.) while Duanm *et al.* [83] have the longest one with sequence in the range of 60-120 seconds. The widest range of ODVs is instead proposed by Xu *et al.* [76] with videos of variable length, from 10 up to 80 seconds. Also, in terms of number of video contents there are various choices: only 5 videos (with 2 more ODVs used during the training phase) in [67] as opposed to 208 ODVs in [76]. Interestingly, two of the most recent datasets presented [78,79] integrate previous databases such as [67–69,73,74,76] and [67,68,77,80,81], respectively. Therefore, they become among the largest and most heterogeneous ODV datasets currently publicly available. Another criteria to select ODVs is based on three main categories: *Content Genres*, *Content Features*, and *Camera Motion*. For instance, authors in [75] select their video only based on the content Genres, offering mainly sport activities related videos. A wider range of genres (e.g., music shows, documentaries, short movies, computer animation and gaming) can be found in most of the publicly available datasets [57,69,70,73]. Other authors choose their content based on attributes such as camera motion

[77], outdoor/indoor scene [74] or a mix of all the aforementioned video categories (*i.e.*, indoor/outdoor scene, fixed/moving camera, and different content genres) [76]. It is worth to mentioning that the most recent dataset presented by Chakareski *et al.* [82] is the only one which presents full UHD ODVs.

The second step of the data collection campaign is the *subjective test*, which represents the core data collection step. Most of the datasets collect navigation data during free-task experiments, which means that viewers could move inside the video

Table 1.3 ODV navigation datasets publicly available. Link to each dataset can be found in the Bibliography.

Y	Reference	Test Preparation			Subj.	Available Data	
		ODVs	Len.	Category		Format	Others
2017	Corbillon [67]	5	70s.	Content Genres	59	Quaternion	Open source software.
	Lo [68]	10	60s.		50	Euler angles	Saliency and motion maps.
	Wu [69]	18	164-655s.	Content Genres	48	Quaternion	Free-task and Task experiments.
	Xu [70]	48	20-60s.	Content Genres	40	Spherical coord.	VQA task.
2018	Ozcinar [71]	6	10s.	Content Genres	17	Spherical coord.	Open source software.
	Fremerey [72]	20	30s.	-	48	Euler angles	Open source software.
	Xu [73]	58	10-80s.	Content Genres	76	Spherical coord.	HM and EM data.
	David [74]	19	20s.	Content Features	57	Spherical coord.	HM and EM data, saliency maps.
	Zhang [75]	104	20-60s.	Content Genres	20★	Spherical coord.	HM and EM data and heatmaps.
	Xu [76]	208	20-60s.	Content Genres and Feat., Camera Motion	31★	Spherical coord.	HM and EM data.
2019	Nasrabadi [77]	28	60s.	Camera Motion	60	Quaternion	Questionnaire on attention.
2020	Rossi [57]	15	20s.	Content Genres	31▲	Spherical coord.	Data also from Laptop and Tablet, code ODV storage optimisation.
	Rondón [78]	306	20s-655s	-	∼42★	Spherical coord.	Aggregation datasets [67–69,73,74,76].
2021	Dharmasiri [79]	88	30-655s.	-	∼45★	Euler angles ♦	Aggregation datasets [67–69,77,80,81], code video segment categorization.
	Chakareski [82]	15	36s.	Content Genres	5-12	Euler angles	RD characteristics of full UHD ODVs.

★ per video.
▲ per video and device.
♦ roll angle was ignored.

content as they wished. There are, however, a few examples where users were asked to take some specific actions. For instance, the work presented in [69] proposes two different experiments: a first set of ODVs are used to identify the natural behaviour in a free-navigation experiment, while a second one is more specific to VR live streaming applications. In fact, in the second experiment live recorded ODV have been used to mimic the case of live-streaming data-tracking, and beyond objective head-movements trajectories, also subjective perception of the video content was captured. Similarly, authors [77] study participant's attention, presence and discomfort levels by a questionnaire at the end of the vision session. Auxiliary subjective quality scores were collected also in [70], in which the dataset is used for Visual Quality Assessment (VQA) tasks. Looking more specifically at the capturing of the objective data (*i.e.*, , trajectories), there are different types of VR devices that can be used, such as laptop, tablet and smartphone. Under the assumption that it provides to most immersive experience, the most widely adopted device is the the head-mounted display (HMD): a helmet with a display and movement sensor able to adapt the rendered image to user's viewing direction. Other datasets however do exist with data collected by other devices. Duanmu *et al.* [83] propose navigation trajectories experienced only on laptop while Rossi *et al.* [57] collect users trajectories across multiple devices (tablet and laptop, in addition to HMD) with the main purpose of studying the effect of the displaying device in the user's navigation. Moreover, most of the presented ODV datasets provide the viewers navigation trajectories as a sequence of head movements over time [67–73]. Even if the head position is a valuable proxy of the user viewing direction, people can still move their eyes and focus on a specific area of the displayed viewport keeping the head fixed. Thus for specific applications, such as visual attention modelling or quality assessment, recording eye gaze movements during the navigation is equally valuable. Hence, there are also dataset containing both information [74–76].

The last part of the creation of an ODV dataset is the *data formatting and storage* of the collected navigation trajectories in an immersive scenario. Since the navigation within ODV is restricted to 3-DoF movements (Figure 1.1), only rotational movements are captured neglecting potential translational movements. These rotational movements can be represented based on several conventions within a spherical system: Euler angles (*i.e.*, yaw, pitch and roll), spherical coordinates (*i.e.*, latitude and longitude), and quaternion. The first two formats are the most common, as shown in Table 1.3, while quaternion is employed only by [67,77], highlighting the higher accuracy and robustness of the quaternion in representing rotational movements. Finally, some of the current publicly available dataset provide also some other data: software that have been used to record users' attention during subjective experiments in order to encourage the community to extend their collected data [67,71,72]; saliency [74] and motion maps [68]; other algorithms such server storage optimisation and video segment categorization in [57] and [79], respectively; Rate-Distortion (RD) characteristics of UHD ODV to correlate with user navigation in [82].

1.4.2 Behavioural Analysis within ODVs

Most of the above papers present, along with the dataset, a general statistical characterization of users' behaviour. This has opened the gate to a new research area aimed at capturing key features that characterize users' interaction while experiencing ODV. In the following, we depict the main findings of this prominent area of research on behavioural analysis. In particular, we distinguish two strands of investigation: a more traditional one aimed at identifying general behavioural features of users while navigating; and a second one focused on identifying more specific and representative users features of the users' behaviour such as users' similarity based on trajectory-based data analysis.

Traditional Data Analysis

The way in which users typically interact with ODVs has been analysed mainly in terms of general metrics such as angular velocity, frequency of fixation, and mean exploration angles. Other than these quantitative metrics, a visual (and qualitative) tool used to study user's behaviour in VR is the heatmap, which identifies areas of the content mostly attended by viewers within a time interval. The investigations based on these aforementioned metrics have given intuitions to answer some key-questions. Finding an answer for these issues is indeed a first step towards the design of user-centric solutions for ODV system. We now summarise the most relevant questions (from a more generic to a more specific behavioural perspective) and the works that aimed at answering them.

Where do users usually look at (on average)? Understanding the areas on which users focus the most within the spherical content is key to optimise the streaming pipeline via viewport-dependent coding techniques or adaptation logic as described in Section 1.3.2, and thus ensuring a good final quality. For this reason, many different researchers at first focused on showing through statistical analysis that users prefer to look at the equatorial area of ODVs rather than at the poles [67,71,74–76]. For instance, some [67,71] use heatmaps averaged across users per each content to show that the users head is densely distributed over time in the equatorial region and in particular, above 100° in terms of latitude. A similar behaviour has been confirmed by Xu *et al.* [76] analysing eye movements: the distribution of gaze fixations show indeed an equatorial tendency, and moreover users move more frequently on the left and right directions than up and down. Deeper investigations on the equatorial bias have highlighted that viewers spend more time towards the front center area of ODVs where typical the main relevant scene is located [70,72,73,83]. Finally, Fremerey *et al.* [72] illustrate in their behavioural analysis that users change their viewing direction only for a short period of time and in general prefer to display the video from a more central and comfortable position.

How do users actually move over time? The average spatial distribution of users' attention might not be enough to characterise their behaviour. For example, a deeper understanding on how viewers actually navigate over time within the video content would allow us to distinguish behaviours that can be predicted or not. In detail, erratic and random navigation within the spherical content is usually challenging to predict. This is the case of the beginning of the navigation within new ODV, in which a predominance of exploratory movements has been pointed out highlighting their randomness, and therefore their difficulty to be anticipated [67,69,79]. However, after a period in the range of 10-20 seconds, viewers tend to converge to common (and more predictable) directions, which typically correspond to the main Focus of Attentions (FoAs) in the scene. Once users find their main point of interest, they tend to not move too much [79]. Corbillon *et al.* [67] show that in a window of 2 seconds, the 95% of analysed subjects stay within a ray of $\pi/2$ from the initial position. This more static and understandable (FoA driven) behaviour make the users' interaction much more predictable than the initial explorative phase.

How is the user behaviour affected by VR devices? Since ODVs can be experienced by different apparatus, a consequent rising question is on the influence of the selected VR devices on the users interactivity. Focusing on the spatial average distribution, the device seems to have a small impact on the overall visitation density: a central and equatorial bias is preserved when navigating via a desktop [83]. Looking at the users' attention over time, however, a dependency on the device has been observed. After a highly-exploratory behaviour at the beginning of the immersive experience (which remains despite the adopted viewing device [84]), a more dynamic navigation has been observed for users displaying ODVs with a laptop than with HMD or tablet [57]. For example, this can be related to a lower sense of immersion and engagement that leads to more explorative movements. Thus, taking into account different VR devices is relevant and fruitful to understand the difference in terms of user interactivity. Further analysis has been presented by Broeck *et al.* [84], focusing on user experience with heterogeneous VR devices such as HMD, tablet and smartphones. In particular, two kinds of omnidirectional video sequences were analysed: video with a static scene (*e.g.*, recorded with a fixed camera) and moving scene. The results show that users explore more in the latter type of content. As expected, the immersion sensation is higher with HMD than other devices while the tablet offers less immersion. Beyond traditional objective (implicit) metrics such as statistical tool, VR experience can be analysed through subjective (explicit) metrics such as users' feedback, and interesting observations can be deduced by comparing implicit and explicit feedback. In [84], authors interviewed participants querying about their sense of immersion during the experience. This explicit data matches the more implicit ones based on users movements since subjects felt more immerse with HMD, despite being less comfortable than with smartphones.

How is the user behaviour affected by content genres? Similarly to the previous question, the correlation between users movements and video content has been analysed from different perspectives. Xu M. *et al.* [70] show that when the main objects in the scene are not located in the central area of the video, users are not focused in the central area of the content, but they instead follow the FoA. Ozcinar *et al.* [71] show a direct correlation between the distribution of fixation points and the video complexity, in terms of Spatial Information (SI) and Temporal Information (TI). In particular, the lower is the TI, the greater is the number of fixations located in the same region. Moreover, the way in which users navigate inside ODVs change for different video categories [57,85]. For instance, Almquist *et al.* [85] highlight that viewers tend to be more uniformly distributed for videos without moving objects. While Rossi *et al.* observe that ODVs without a main object in the scene bring users to have highly exploratory interaction, especially with HMD. On the contrary, the level of interactivity is not correlated with the viewing device if there is a main focus of attention in the video which capture user attention [57]. Finally, the correlation between navigation patterns and the video content is so relevant such that Dharmasiri *et al.* [79] use the fixation distribution as a proxy for their video categorization algorithm.

Are users consistent in their navigation? A final and yet important question is if users tend to navigate in a consistent way. In other words, researchers have studied the average behaviour of users but also their variance and deviation. Small variance means that heatmaps can be representative enough of the users' behaviour, leading to reliable prediction of the user interaction. A general high consistency among users in terms of spatial distribution has been identified by general statistical analysis for head movements. For instance, Xu *et al.* [70] evaluate a high linear correlation between heatmaps generated by two random sets of users. A similar outcome has been shown in [73] where in particular, it has been highlighted that almost 50% of users focus on the same viewing area among 8 quantized regions of the ODV. These works mainly focused on summary statistics such as mean and variance of users' behaviour averaged across users. Looking at deeper analysis focused on pairwise comparison, authors in [76] have evaluated the average intersection angle of eye-gaze direction per each pair of participants across each content. This analysis highlights heterogeneity in users' behaviour, in contrast with the observations carried out from the other studies. This inconsistency suggests the need to go beyond traditional statistical analysis to better understand the user behaviour within immersive content. Moreover, beyond the inconsistency just highlighted, it is worth mention that most of the above studies are focused on behaviours averaged over time – for example heatmaps and eye fixation. These metrics are highly informative about the spatial behaviour (where do users tend to look at) but only partially informative about at the temporal behaviour (we can deduce how much erratic users tend to be, but not really if two users are interacting similarly). However, deeper temporal analysis is essential to develop reliable users prediction algorithm, deeper behavioural analysis and enable user-centric systems. In the following, we then review this second strand of research focused on trajectory-based data analysis and highlight the key outcomes and novelty that emerge and how this can lead to user-centric systems.

Trajectory-based Data Analysis

As emerged in the previous discussion, behavioural analysis based on statistical tools or heatmap provides a general understanding of user's behaviour in ODV content failing in detecting deep insights into users' attention dynamics, such as how much viewers interact in harmony among themselves. Specifically, there are still crucial questions that need to be addressed: "*Are users interacting in a similar way? Can human behaviour be predicted?*" Answering to these questions is indeed essential for many and not trivial tasks. In this context, detecting viewers who are navigating in a similar way would help to improve the accuracy and robustness of predictive algorithm and thus, to personalise the delivery strategy. This information could also be exploited for identifying key navigation trajectories which can be used either to optimise video coding or QoE assessment. Beyond immersive video streaming system applications, being able to detect users who are interacting in a similar way from those who are not, might be essential for medical purposes such as studying psychiatric disorders [86,87]. Equipped with this motivation, a new direction of behavioural analysis has started aimed at identifying behaviour similarities among users, across video content and/or devices. This has led to the development of new metrics and tools built in the space-time trajectory domain.

Clustering is one of the most popular and robust techniques to infer data structure and it has been therefore employed in the context of VR applications. Based on intuitions from vehicle trajectory prediction, Petrangeli *et al.* [88] model each user navigation as independent trajectories in terms of roll, pitch, and yaw angles, and apply a spectral clustering [89,90] to identify trajectories with similar behaviour over time. The dominant trajectories, identified by the main clusters, are eventually used to predict new viewers. While this method is efficient in discovering general trends of users' attention, it is not focused on identifying clusters that are consistent in terms of displayed content; meaning that users in the same group do not necessarily consume the same portion of ODV. Clustering to perform long-term trajectory predictions is presented in [91], where authors first adopt a well-known spectral clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), to identify key clusters and trajectories from a set of training samples. Then, when new (test) users start an ODV streaming session, they are assigned to a specific cluster and future trajectories are predicted accordingly. This association step is based on viewing direction comparison measured in the equirectangular plane, neglecting the actual spherical geometry. Therefore, the clusters identified by both algorithms suffer of two major shortcomings: 1) not identifying group of users necessarily consuming the same portion of the ODV; 2) not necessarily considering the spherical geometry into account. To overcome these limitations, a novel spherical clustering tools specific for ODV users has been proposed in [92]. This is a graph-based algorithm that is able to identify groups of viewers based on their consistency in the navigation. In practice, authors first define a metric to quantify the common displayed portion of the content (*i.e.*, overlapped viewport) among users; based on this metric, they build a graph whose nodes are associated to different viewers, and finally, a clustering method based on Bron-Kerbosh algorithm is applied to build clusters as cliques (*i.e.*, subgraph of inter-connected nodes). Thus, the algorithm detects and groups users *only if* they consistently display similar viewports over time while consuming the same ODV content, and this viewport comparison is done taking into account also the spherical geometry of the content. Due to its robustness and specificity for VR content, this tool has been used to analyse navigation patterns in some publicly available datasets such as [57,77]. The number of total clusters and users per cluster is a proxy of users' similarity in exploring ODVs: the fewer the clusters, the more users are focused on a similar area of the video content highlighting a similar behaviour. Both [57,77] have showed that video characterised by a dominant focus of attention (*i.e.*, moving objects) are explored in a similar way by users, except few exceptions due to camera motion [77]. For instance, videos which have a vertical camera motion bring viewers not to be focused on a specific area but to be distributed on the landscape.

With the idea of developing an objective metric able to assess users' similarity in a spatio-temporal domain, Rossi *et al.* [57] propose a novel *affinity metric* to quantify the detected similarities among viewers via the spherical clustering. This is evaluated as the normalised weighted average of cluster popularity (i.e., how many users per cluster): 0 when users are not clustered together because highly dispersed in the navigation; 1 when users share a strong similarity in their trajectories. Equipped with this tool, they noticed that the affinity between users is highly correlated to the selected VR device: HMD leads to higher *affinity metric* than other devices. A similar outcome has been also identified by authors [93], which directly compare different prediction models for navigation trajectories collected by different devices such as HMD, PC and smartphone. Other than showing that exploration done via desktop is more static than the other devices, it has also been verified a higher prediction accuracy for HMD users, especially in the near future. Following this research direction, a novel viewport clustering algorithm as a tool for behavioural analysis and video categorization has been proposed recently [79]. The main novelty is to consider multimodal clustering analysis, in which the spherical location of each viewer is augmented by other modalities as input such as head movement speed and the percentage of the sphere explored in a given time window. In this way, users in the same clusters are similar not only in terms of displayed viewport but also for the style of exploration.

Beyond clustering, other tools have been proposed to understand similarities and study user predictability. In [94], authors propose to base their behavioural analysis on tools from information theory such as actual entropy, transfer entropy and mutual information, which quantify randomness and uncertainty in the trajectories of users. For example, the actual entropy is low for users that experience “repetitive" behaviour (trajectories) over time, leading to highly predictable users. From a first intra-behaviour analysis, aimed at studying the behaviour of one single user across diverse contents, profiling behaviours have been identified: some viewers conserve similar navigation independently by the video content. Moreover, they also quantify a more randomness in navigation trajectories within ODVs lacking specific FoA. As second line of investigation, an inter-user behaviour analysis has also been conducted, measuring how informative other users' behaviour is for a current user. In other words, the study quantifies how much information about the predictability within a specific content can be extracted by the navigation of a given group of users. This analysis is similar to the previous one based on clusters and users' similarity. However, the information theory tools capture a more meaningful behaviour and better quantify viewers similarity during their navigation than metrics based only on the spatial location of users, such us the spherical clustering proposed in [92].

1.5 User-centric ODV streaming

The tools and metrics discussed in Section 1.4.2 enable a deeper understanding and prediction level of the users' interaction, opening the gate to personalised and user-centric systems, Figure 1.3 (light blue and green boxes), where the different steps of the pipeline are tuned based on single user's behaviour. In the following, we review the main contributions that have been proposed toward user-centric systems, distinguishing the work based on the type of behavioural information: i) extracted from a single user (*single-user design*), ii) extracted from multiple users (*cross-user design*).

1.5.1 Single-User Designs

A key step in user-centric systems is the prediction of the users' head movement. The first and among the simplest technique (and most widely adopted) is based on the past and current trajectory of a single user, neglecting other viewers and video content information. Qian *et al.* [95] have experimentally shown how three simple logistic regression models, such as average, Linear Regression (LR) and Weighted LR (WLR) with a moving window of 1 second are able to successfully anticipate the user behaviour in the next short time window (*i.e.*, 0.5, 1, and 2 seconds). For the first time, authors are able to prove the potentiality of the prediction. Specifically, giving higher fetching priority to tiles that most likely will be displayed can reduce the bandwidth usage up to 80%. Multiple works have followed all adopting such simple predictive algorithms for user-centric adaptation logic, showing the gain in terms of bandwidth, re-buffering, and final quality experience by the user. Experimental validation has been proposed in [96], in which the user-centric adaptation logic has been tested on real-world 4G bandwidth. Results have shown that the proposed strategy maintains a good displayed quality (the same achieved when sending the entire panorama) but with a reduction of bandwidth overload up to 35%. Logistic regression based on historical data has been used also in [97], in which the adaptation logic for ODV streaming optimises the optimal representation per tile to request according to network bandwidth and predicted users' head movements. Their analysis emphasised the need of accurately predicting future viewport position for ODV streaming. A WLR algorithm is considered in a similar framework but in the case of scalable video coding [98]. Specifically, high-quality representations of tiles within the predicted viewport are prefetched shortly before being visualised to ensure high quality in the displayed content and at the same time to reduce storage costs at the server-side, which has been an open challenge especially for ODV that is highly data intensive.

Most of these works have shown the potentiality of user-centric systems, but suffer from poor prediction accuracy in the long-term (mainly due to the lack of other users and content information). At the same time, the behavioural studies highlighted in Section 1.4.2 have shown a strong consistency and similarity in the way in which users navigate ODVs, motivating cross-users designs described in the following subsection.

1.5.2 Cross-User Designs

To breakthrough the limitation of single-user frameworks, a new research direction has been carried out aimed at exploiting behavioural information from *multiple* users to identify and predict the most popular trends in navigating ODVs and develop user-centric systems accordingly. In the following, we review these efforts first describing the work that are *content-agnostic*, mainly based on user cross-users information. Then, we introduce the *content-based* works, in which users' information is augmented by content information.

Content-agnostic Designs

At first, linear model (LR) and classical clustering have been widely used to infer single user behaviour from cross-user information [88,91,99,100]. For example, with the main intent to improve the long-term users' viewport prediction, Ban *et al.* [99] propose a viewport prediction approach based on K-Nearest-Neighbors (KNN) algorithm and aimed at combining both behavioural characteristics of the single-user and those extracted by benchmark viewers. Specifically, the algorithm is composed of two main steps: *i)* the user head position is predicted via LR model based on the historical movements of only the single user under investigation; *ii)* this prediction is then used to form the K nearest users set, as the users with the closest viewport centres previously collected. The K-NN set is used to compute the viewing probability per tiles. This is one of the proposed methods by Xie *et al.* [91], aimed at using cross-user information to identify main clusters of users and then predict new users mimicking the behaviour of the closest cluster. Per each detected group, the viewing probability of tiles is then computed and applied to support the viewport prediction during video playback of new viewers. Even if the prediction is more accurate in a time window of 3 seconds and longer, viewers are clustered based on Euclidean distance, neglecting the actual spherical geometry, resulting in not fully representative clusters. With the idea of exploiting the spherical geometry, other white-box models have been proposed. For instance, Hu *et al.* in [101] use a graph-based approach to improve the accuracy of viewport prediction in a QoE-optimised ODV streaming system. Authors first predict tiles in the FoV by a tile-view graph learned from historical users' navigation trajectories: a weighted graph is constructed in which each vertex corresponds to a tile while the weight is given by users' behaviour. From the constructed graph, a tile view probability is finally evaluated and used to optimise the downloading bitrate per tile in a limited bandwidth system but maximising the users' quality of experience. Authors deeply compare their proposed system with other algorithms (both navigation predictive and streaming) reaching 20% improvement in terms of users' QoE.

Beyond the models discussed above, cross-users data analysis have opened the gate to deep learning frameworks aimed at inferring non linear interactivity models from a training dataset of collected trajectories based. These models have been augmented also by auxiliary losses [80] or by probabilistic models [102,103] or state-of-the-art transformers [104] aimed at inferring the prediction error. For example, a Gaussian distribution based on previous immersive navigation experiences is used to model the distribution of short-term prediction error for new viewers in the system. This prediction approach is used to improve a viewport-dependent streaming system following a tile-based streaming approach in [102] and an improved coding technique (*i.e.*, pyramid projection) to adapt quality distribution based on users' behaviour in [103]. Instead, Chao *et al.* [104] propose a viewport prediction transformer method for ODV, named 360° Viewport Prediction Transformer (VPT360), taking advantage of transformer architecture [105].

Content-aware Designs

In the previous sections we have described the advances made toward user-centric systems, in the case in which user behavioural analysis was carried out by looking only at users data. However, the studies in Section 1.4.2 have shown that users' attention is steered by the content as well. Therefore, in the following we review user-centric designs in which researchers have used both users trajectories and content features to infer user behaviour.

A very well known metric that maps content features into user attention is the *saliency map*, which estimates the eye fixation for a given omnidirectional content. Since a correlation between saliency map and user trajectory has been empirically proved in [106], many efforts have been dedicated to study, infer, and exploit saliency in ODV streaming. Specifically, deep learning frameworks aimed at predicting users trajectories where augmented by using saliency maps as further input [73,107,108]. Different learning architectures and paradigms were considered in these studies: Reinforcement Learning (RL) based approach looking at the user's behaviour as sequential actions taken over time [73]; and recurrent learning approach exploiting the temporal correlation of users trajectories [107,108]. Xu *et al.* [73] proposed an RL based workflow that first estimates the saliency map for each frame and then predicts the viewport direction based on historical data and predicted saliency. This prediction is cast as a RL agent that aims at minimizing the prediction loss (dissimilarity between the predicted and ground-truth trajectories). Its viewport prediction is however short-term being limited to the next frame only (*i.e.*, about 30ms prediction ahead). In the case of recurrent neural networks, Nguyen *et al.* [107] feed a long short-term memory (LSTM) network with both saliency (inferred by a CNN model) and historical head orientation from users. The learning framework was able to overcome main limitations such as central saliency bias and single object focus (*i.e.*, ODV users quickly scan through all objects in a single viewport). Interestingly, Rondón *et al.* [108] show that the historical data points (in terms of past trajectories) and content features may influence the future trajectories differently based on the prediction horizon. They observe that users trajectory is affected by the content mainly if toward the end of the trajectory. This is explained by the fact that at the initial phase of the trajectory users tend to have more erratic (and less content driven) behaviour. As a consequence, they propose a prediction model that initially prioritises user trajectory inertia, counting more the visual content at a later stage. As a consequence, they consider both positional features and saliency as two time series to feed to the LSTM, as opposed to other works that only consider the former ones. In this way, Rondón *et al.* sequence-to-sequence method showed better results compared to other LSTM ones such as Nguyen *et al.* [107] which do not consider time-dependency in the saliency maps. These deep learning frameworks have strong potentiality but it is well known that they are data-hungry, with a tendency of poor training accuracy or lack of generalisation in the case of the limited datasets. Hence, works have been also presented in a parallel direction of "shallow" learning frameworks, such as the one from Zhang *et al.* [109]. Authors designed their trajectory prediction as a sparse directed graph problem inferred by past users' positions, saliency map data, and the biological human head model (which defines transition constraints on the graph given the physiological constraints such as impossible head movements).

Beyond saliency maps, other content features have been considered. In the case of dynamic scenes, for example, saliency might not be representative enough and *content motion* can be preferred as a feature. Such motion can be captured from either optical video flow [110–112] or from individually detected objects' movements [113,114]. The optimal flow can be extracted by the well-known differential method Lucas-Kanade algorithm, as well as by Gaussian mixture models [112]. In [110], the optical flow as well as the saliency and the past trajectory are input to a LSTM-based prediction model. User-centric systems that exploit the proposed fixation prediction network achieve a reduction of both bandwidth consumption and initial buffering time. Deep learning frameworks have also been considered to extract content features and favor the temporal prediction of users viewing trajectories. Park *et al.* [115] implement a 3D-CNN to extract spatio-temporal features from both saliency and optical flow to predict future viewing directions. The predicted trajectory is then exploited in a RL model that determines the downloading order and the downloading bitrate for tile-based streaming. Xu *et al.* [111] also exploit the CNN architecture but for predicting users' gaze. With this goal in mind, authors created an eye-tracking dataset captured from dynamic scenes. The computed saliency maps and the content motion maps are in two spatial scales: at the entire panorama image; and at a sub-image centred at the current gaze point. Both saliency and motion maps feed a CNN for feature extraction, and then a LSTM predicts gaze direction using the current time and gaze point. Moreover, other works use individually detected objects' movements, mostly following the success of YOLO (You Only Look Once) [116] for objection recognition Chopra *et al.* [113] propose an online regression model based on trajectories of both users and the main objects, which are extracted online from the detection model. They claim that the user's head movement highly depends on objects trajectories. Their experiments highlight 34% model weighted given the object's trajectories. Park *et al.* [114] argued that the semantic video information of motion objects (*e.g.*, people, cars) is also useful to predict the users' behaviour. Given extracting video semantic information is a high-computing task, something not common in mobile devices or HMD, they proposed a server-side analysis shared with clients in DASH MPD files. The shared information contains objects presented in tiles and the probabilities of each object being seen given other users viewing them. In other words, objects that were of interest to most of the users will have higher probability to be followed by a new user.

1.6 Conclusions and Perspectives

Omnidirectional videos (ODVs) have become widely spread since 2007 with their first commercial application (*i.e.*, Google street view) and have attracted a growing attention in the multimedia community. This novel multimedia format has revolutionised how users engage and interact with media content, going beyond the passive paradigm of traditional video technology, and offering higher degrees of presence and interaction. Thus, many new challenges have risen over the entire end-to-end communication chain due to the novel role of the user and the new geometry. For example, the spherical content need to be efficiently delivered to the viewer taking into account also the aspect of user-content interaction and bandwidth limitation. In this context, this chapter presents a summary of research advances in ODV adaptive streaming, clearly distinguishing works in terms of *system-centric* and *user-centric* streaming solutions. System-centric approaches come from a quite straightforward extension of well-established solutions for the 2D video pipeline, adding value in the streaming strategy making it user-aware (*i.e.*, tile-based viewport-dependent streaming). Given the key role of the users, behavioural investigations on how viewers navigate within ODV have attracted a lot of interest, showing the benefit of understanding users' behaviour, and enable personalised ODV streaming solutions (*i.e.*, user-centric streaming system).

Despite the intense efforts, there are still several potential research directions in this area. Looking at the streaming pipeline, advances can be done in making ODV processing steps even more geometry- and user-aware. In the first direction, for instance, on-sphere encoding methods instead of plane ones has shown promising results for omnidirectional images [52], and many future works could be done also on ODVs. Coding efficiency can be improved further also advancing tile-based design. To the best of our knowledge, user-behaviour, spherical geometry and content motion are not simultaneously taken into account when optimizing the design of the tiles. On a parallel direction, viewport prediction research has been experiencing still many open challenges. Novel approaches are now emerging to improve the prediction accuracy, considering different information, such as spatial audio and user emotion. For instance, authors in [117] propose to improve prediction incorporating spatial audio characteristics of the video content. However, there is a lack of public ODV dataset with spatial audio information to enable such direction [118]. Regarding emotion, there are efforts to enable labelling emoting during VR presentation [119] to create ground truth for user emotion during immersion in ODV. However, a trained model with such data can automatically detect emotion from recorded pupils to perform predictions [120].

As emerged in Section 1.4.2, an increasing interest in modeling and understanding users' behaviour in ODV systems has been shown recently. Even if some key-questions have been addressed, there are still open issues. For instance, a better understanding of common behavioural information among users is still missing and could be eventually exploited in a proper tool that identifies *ODV viewers' profiling*. A user profile is a collection of information that describes the behavioural features of a user and that is used to identify key behaviours. In the ODV context, users' profiles can be utilized for different purposes such as enabling new modalities for *viewport prediction* and *streaming services optimised for users' profiles*. Of strong interest for industry would be the live-streaming scenarios (*e.g.*, live sports) in which users' profile might have a high impact especially in reducing the cold-start problem in online prediction models. Even if user profiling can benefit for providing better services, it raises significant concerns toward user *privacy*. Some very recent works have been focused on this specific issue, aiming at developing user-centric solutions that are privacy-compliant. For instance, Wei *et al.* point out two directions to protect user viewport [121]. First, by using the federated learning method during training, only the model parameters are sent to the prediction server. Another way is obfuscating the real user position by performing a normalised number of camouflaged tile requests. Their experiments have shown that the scale of camouflaged requests is determinant in achieving better prediction, and users might decide such scale of data sharing (trade off between privacy and service personalisation).

Finally, it is also worth to mentioning the growing interest for *volumetric video*. This content, also known as holograms, will completely revolutionize the way we communicate. Moving from 3-DoF to 6-DoF systems, the user is not more limited to head movements (as in omnidirectional content) but can perform freely body movements among holograms, dynamically changing the both perspective and proximity from which he/she sees them. Research challenges linked to volumetric visual data are still numerous and in their infant phase, as discussed in part IV of this book [122]. Volumetric signals are typically represented as dynamic polygon meshes or point clouds. Therefore, the regular pixel-grids (frames in traditional multimedia content) or the sphere-based acquisition pattern (typical of 360° contents) are left to more abstract data structure. With volumetric videos, a collection of colour points information evolving over time is acquired, with no hard constraints on the preservation of geometry over time. Even if the signal has a completely different structure, volumetric videos share some common challenges with omnidirectional content such as a large volume of data, ultra-low delay application and the uncertainty of user requests. Thus, all the studies done for the user-centric system in ODV and presented in this work can be exploited in a new solution for a more immersive system with volumetric representation.

Bibliography

[1] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, S. Wee, The road to immersive communication, Proceedings of the IEEE 100 (4) (2012) 974–990.

[2] I. Grand View Research, Virtual Reality Market Size, Share & Trends Analysis Report By Technology (Semi & Fully Immersive, Non-immersive), By Device (HMD, GTD), By Component (Hardware, Software), By Application, And Segment Forecasts, 2021 - 2028, [Online] `https://www.grandviewresearch.com/industry-analysis/virtual-reality-vr-market` (2021).

[3] C. Flavián, S. Ibáñez-S ánchez, C. Orús, The impact of virtual, augmented and mixed reality technologies on the customer experience, Journal of business research 100 (2019) 547–560.

[4] M. Slater, M. V. Sanchez-Vives, Enhancing our lives with immersive virtual reality, Frontiers in Robotics and AI 3 (2016) 74.

[5] J. L. Rubio-Tamayo, M. Gertrudix Barrio, F. García García, Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation, Multimodal Technologies and Interaction 1 (4) (2017) 21.

[6] M. V. Sanchez-Vives, M. Slater, From presence to consciousness through virtual reality, Nature Reviews Neuroscience 6 (4) (2005) 332–339.

[7] M.-L. Ryan, Immersion vs. interactivity: Virtual reality and literary theory, SubStance (1999) 110–137.

[8] J. Mütterlein, The three pillars of virtual reality? investigating the roles of immersion, presence, and interactivity, in: Proceedings of the 51st Hawaii international conference on system sciences, 2018.

[9] A. Perkis, C. Timmerer, S. Baraković, J. B. Husić, S. Bech, S. Bosse, J. Botev, K. Brunnström, L. Cruz, K. De Moor, et al., QUALINET white paper on definitions of immersive media experience (IMEx), European Network on Quality of Experience in Multimedia Systems and Services, 14th QUALINET meeting (online) (2020).

[10] M. T. Vega, J. van der Hooft, J. Heyse, F. De Backere, T. Wauters, F. De Turck, S. Petrangeli, Exploring New York in 8K: an adaptive tile-based virtual reality video streaming experience, in: Proceedings of the 10th Multimedia Systems Conference, ACM, 2019, pp. 330–333.

[11] B. Han, Mobile immersive computing: Research challenges and the road ahead, Communications Magazine 57 (10) (2019) 112–118.

[12] P. Cremonesi, F. Garzotto, R. Turrin, User-centric vs. system-centric evaluation of recommender systems, in: IFIP Conference on Human-Computer Interaction, Springer, 2013, pp. 334–351.

[13] R. Stankiewicz, P. Cholda, A. Jajszczyk, Qox: What is it really?, Communications Magazine 49 (4) (2011) 148–158.

[14] M. Agiwal, A. Roy, N. Saxena, Next generation 5G wireless networks: A comprehensive survey, Communications Surveys & Tutorials 18 (3) (2016) 1617–1655.

[15] Z. Chen, Y. Li, Y. Zhang, Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation, Signal Processing 146 (2018) 66–78.

[16] D. He, C. Westphal, J. J. Garcia-Luna-Aceves, Network Support for AR/VR and Immersive Video Application: A Survey, in: Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, Science and Technology Publications, 2018, pp. 359–369.

[17] C.-L. Fan, W.-C. Lo, Y.-T. Pai, C.-H. Hsu, A Survey on 360° Video Streaming: Acquisition, Transmission, and Display, ACM Computing Surveys 52 (4) (2019) 1–36.

[18] M. Zink, R. Sitaraman, K. Nahrstedt, Scalable 360° Video Stream Delivery: Challenges, Solutions, and Opportunities, Proceedings of the IEEE 107 (4) (2019) 639–650.

[19] R. G. d. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, P. Frossard, Visual Distortions in 360° Videos, IEEE Transactions on Circuits and Systems for Video Technology

30 (8) (2020) 2524–2537.

[20] A. Yaqoob, T. Bi, G.-M. Muntean, A survey on adaptive 360° video streaming: Solutions, challenges and opportunities, IEEE Communications Surveys & Tutorials 22 (4) (2020) 2801–2838.

[21] R. Shafi, W. Shuai, M. U. Younus, 360° Video Streaming: A Survey of the State of the Art, Symmetry 12 (9) (2020) 1491.

[22] M. Xu, C. Li, S. Zhang, P. Le Callet, State-of-the-art in 360° video/image processing: Perception, assessment and compression, IEEE Journal of Selected Topics in Signal Processing 14 (1) (2020) 5–26.

[23] J. Ruan, D. Xie, Networked VR: State of the Art, Solutions, and Challenges, Electronics 10 (2) (2021) 166.

[24] F. Chiariotti, A survey on 360-degree video: Coding, quality of experience and streaming, Computer Communications 177 (2021) 133–155.

[25] Google, Street View's 15 favorite Street Views, [Online] https://blog.google/products/maps/street-views-15-favorite-street-views/ (2020).

[26] I. Sodagar, The MPEG-DASH Standard for Multimedia Streaming Over the Internet, IEEE MultiMedia 18 (4) (2011) 62–67.

[27] Meta, Facebook to Acquire Oculus, [Online] https://about.fb.com/news/2014/03/facebook-to-acquire-oculus/ (2014).

[28] Google, Open sourcing google cardboard, [Online] https://developers.googleblog.com/2019/11/open-sourcing-google-cardboard.html (2019).

[29] Google, A new way to see and share your world with 360-degree video, [Online] https://blog.youtube/news-and-events/a-new-way-to-see-and-share-your-world/ (2015).

[30] Meta, Insights from a year of 360 videos on facebook, [Online] https://facebook360.fb.com/2017/02/16/insights-from-a-year-of-360-videos-on-facebook/ (2017).

[31] ISO Central Secretary, Spatial relationship description, generalized URL parameters and other extensions, Standard Tech. Rep., ISO/IEC 23009-1:2014/Amd 2, International Organization for Standardization, Geneva, CH (2015).

[32] BBC, Click: How we made BBC's first fully 360-degree show, [Online] https://www.bbc.com/news/technology-35752662 (2016).

[33] DigitalTV Europe, Arte launches virtual reality TV app, [Online] https://www.digitaltveurope.com/2016/01/14/arte-launches-virtual-reality-tv-app/ (2016).

[34] E. Kuzyakov, D. Pio, Next-generation video encoding techniques for 360° video and VR, [Online] https://engineering.fb.com/2016/01/21/virtual-reality/next-generation-video-encoding-techniques-for-360-video-and-vr/ (2016).

[35] Y.-K. Wang, M. K. Hendry, Viewport dependent processing in VR: partial video decoding, in: Proceedings of the 116th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, MPEG, Vol. 116, 2016, [Online] https://mpeg.chiariglione.org/meetings/116.

[36] Vimeo, Vimeo 360: A home for immersive storytelling, [Online] https://vimeo.com/blog/post/introducing-vimeo-360/ (2017).

[37] Google, Improving VR videos, [Online] https://youtube-eng.googleblog.com/2017/03/improving-vr-videos.html (2017).

[38] M. M. Hannuksela, Y.-K. Wang, An Overview of Omnidirectional MediA Format (OMAF), Proceedings of the IEEE 109 (9) (2021) 1590–1606.

[39] Facebook, Introducing meta: A social technology company, [Online] https://about.fb.com/news/2021/10/facebook-company-is-now-meta/ (2021).

[40] J. Le Feuvre, C. Concolato, Tiled-based adaptive streaming using MPEG-DASH, in: Proceedings of the 7th International Conference on Multimedia Systems, MMSys '16, Association for Computing Machinery.

[41] ISO Central Secretary, Information technology — Coding of audio-visual objects — Part 12: ISO base media file format (2020).

[42] T. Thanh Le, J.-B. Jeong, S. Lee, J. Kim, E.-S. Ryu, An Efficient Viewport-Dependent 360 VR System Based on Adaptive Tiled Streaming, in: Computers, Materials and Continua, Vol. 66, Tech Science Press, 2021, pp. pp.2627 – 2643.

[43] M. Viitanen, A. Koivula, A. Lemmetti, A. Ylä-Outinen, J. Vanne, T. D. Hämäläinen, Kvazaar: Open-source hevc/h.265 encoder, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, [Online].
URL `http://doi.acm.org/10.1145/2964284.2973796`

[44] K. K. Sreedhar, I. D. D. Curcio, A. Hourunranta, M. Lepistö, Immersive media experience with MPEG OMAF multi-viewpoints and overlays, in: Proceedings of the 11th Multimedia Systems Conference, ACM, pp. 333–336.

[45] H. S. Kim, S. B. Nam, S. G. Choi, C. H. Kim, T. T. K. Sung, C.-B. Sohn, HLS-based 360 VR using spatial segmented adaptive streaming, in: IEEE International Conference on Consumer Electronics, 2018, pp. 1–4.

[46] T. Maugey, Acquisition, representation and rendering of omnidirectional videos, in: Immersive Video Technologies, Elsevier, 2022.

[47] D. Pio, E. Kuzyakov, Under the hood: Building 360 video, [Online] `https://engineering.fb.com/2015/10/15/video-engineering/under-the-hood-building-360-video/` (2015).

[48] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, Overview of the High Efficiency Video Coding (HEVC) Standard, IEEE Transactions on Circuits and Systems for Video Technology 22 (12) (2012) 1649–1668.

[49] S. Afzal, J. Chen, K. K. Ramakrishnan, Characterization of 360-degree videos, in: Proceedings of the Workshop on Virtual Reality and Augmented Reality Network, ACM, 2017, p. 1–6.

[50] P. R. Alface, J.-F. Macq, N. Verzijp, Interactive omnidirectional video delivery: A bandwidth-effective approach, Bell Labs Technical Journal 16 (4) (2012) 135–147.

[51] B. Vishwanath, T. Nanjundaswamy, K. Rose, Rotational motion model for temporal prediction in 360° video coding, in: 19th International Workshop on Multimedia Signal Processing, IEEE, 2017.

[52] N. M. Bidgoli, R. G. d. A. Azevedo, T. Maugey, A. Roumy, P. Frossard, OSLO: On-the-sphere learning for omnidirectional images and its application to 360-degree image compression, arXiv preprint arXiv:2107.09179 (2021).

[53] K. K. Sreedhar, A. Aminlou, M. M. Hannuksela, M. Gabbouj, Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications, in: International Symposium on Multimedia, IEEE, 2016, pp. 583–586.

[54] X. Corbillon, A. Devlic, G. Simon, J. Chakareski, Optimal set of 360-degree videos for viewport-adaptive streaming, in: Proceedings of the 25th International conference on Multimedia, ACM, 2017, pp. 943–951.

[55] M. Graf, C. Timmerer, C. Mueller, Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP: Design, implementation, and evaluation, in: Proceedings of the 8th ACM on Multimedia Systems Conference, ACM.

[56] C. Ozcinar, J. Cabrera, A. Smolic, Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9 (1) (2019) 217–230.

[57] S. Rossi, C. Ozcinar, A. Smolic, L. Toni, Do users behave similarly in VR? Investigation of the user influence on the system design, ACM Transactions on Multimedia Computing, Communications, and Applications 16 (2) (2020) 1–26, [Online] `https://github.com/V-Sense/VR_user_behaviour`.

[58] S. Ahsan, A. Hourunranta, I. D. D. Curcio, E. Aksu, FriSBE: adaptive bit rate streaming of immersive

tiled video, in: Proceedings of the 25th Workshop on Packet Video, ACM, 2020, pp. 28–34.

[59] J. Fu, X. Chen, Z. Zhang, S. Wu, Z. Chen, 360SRL: A Sequential Reinforcement Learning Approach for ABR Tile-Based 360 Video Streaming, in: International Conference on Multimedia and Expo, IEEE, 2019, pp. 290–295.

[60] S. Rossi, L. Toni, Navigation-aware adaptive streaming strategies for omnidirectional video, in: 19th International Workshop on Multimedia Signal Processing, IEEE, 2017, pp. 1–6.

[61] F. Duanmu, E. Kurdoglu, Y. Liu, Y. Wang, View direction and bandwidth adaptive 360 degree video streaming using a two-tier system, in: International Symposium on Circuits and Systems, IEEE, 2017, pp. 1–4.

[62] M. Nguyen, D. H. Nguyen, C. T. Pham, N. P. Ngoc, D. V. Nguyen, T. C. Thang, An adaptive streaming method of 360 videos over HTTP/2 protocol, in: 4th NAFOSTED Conference on Information and Computer Science, IEEE, 2017, pp. 302–307.

[63] D. V. Nguyen, T. T. Le, S. Lee, E.-S. Ryu, SHVC tile-based 360-degree video streaming for mobile VR: PC offloading over mmWave, Sensors 18 (11) (2018) 3728.

[64] P. Maniotis, N. Thomos, Viewport-Aware Deep Reinforcement Learning Approach for 360° Video Caching, IEEE Transactions on Multimedia (2021).

[65] J. Son, D. Jang, E.-S. Ryu, Implementing Motion-Constrained Tile and Viewport Extraction for VR Streaming, in: Proceedings of the 28th SIGMM Workshop on Network and Operating Systems Support for Digital Audio and Video, ACM, 2018, pp. 61–66.

[66] N. M. Bidgoli, T. Maugey, A. Roumy, Fine granularity access in interactive compression of 360-degree images based on rate-adaptive channel codes, IEEE Transactions on Multimedia 23 (2020) 2868–2882.

[67] X. Corbillon, F. De Simone, G. Simon, 360-degree video head movement dataset, in: Proceedings of the 8th on Multimedia Systems Conference, ACM, 2017, pp. 199–204, [Online] https://doi.org/10.1145/3193701.

[68] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, C.-H. Hsu, 360° video viewing dataset in head-mounted virtual reality, in: Proceedings of the 8th on Multimedia Systems Conference, ACM, 2017, pp. 211–216, [Online] https://doi.org/10.1145/3192927.

[69] C. Wu, Z. Tan, Z. Wang, S. Yang, A dataset for exploring user behaviors in VR spherical video streaming, in: Proceedings of the 8th on Multimedia Systems Conference, ACM, 2017, pp. 193–198, [Online] https://doi.org/10.1145/3192423.

[70] M. Xu, C. Li, Y. Liu, X. Deng, J. Lu, A subjective visual quality assessment method of panoramic videos, in: International Conference on Multimedia and Expo, IEEE, 2017, pp. 517–522, [Online] https://github.com/Archer-Tatsu/head-tracking.

[71] C. Ozcinar, A. Smolic, Visual attention in omnidirectional video for virtual reality applications, in: 2018 Tenth international conference on quality of multimedia experience, IEEE, 2018, pp. 1–6, [Online] https://github.com/cozcinar/omniAttention.

[72] S. Fremerey, A. Singla, K. Meseberg, A. Raake, AVtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD, in: Proceedings of the 9th Multimedia Systems Conference, ACM, 2018, pp. 403–408, [Online] https://github.com/acmmmsys/2018-AVTrack360.

[73] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, Z. Wang, Predicting head movement in panoramic video: A deep reinforcement learning approach, IEEE Transactions on pattern analysis and machine intelligence 41 (11) (2018) 2693–2708, [Online] https://github.com/YuhangSong/DHP.

[74] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, P. L. Callet, A dataset of head and eye movements for 360° videos, in: Proceedings of the 9th Multimedia Systems Conference, ACM, 2018, pp. 432–437, [Online] https://salient360.ls2n.fr/datasets/.

[75] Z. Zhang, Y. Xu, J. Yu, S. Gao, Saliency detection in 360° videos, in: Proceedings of the European

conference on computer vision, 2018, pp. 488–503, [Online] https://github.com/xuyanyu-shh/Saliency-detection-in-360-video.

[76] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, S. Gao, Gaze prediction in dynamic 360 immersive videos, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 5333–5342, [Online] https://github.com/xuyanyu-shh/VR-EyeTracking.

[77] A. T. Nasrabadi, A. Samiei, A. Mahzari, R. P. McMahan, R. Prakash, M. C. Farias, M. M. Carvalho, A taxonomy and dataset for 360° videos, in: Proceedings of the 10th Multimedia Systems Conference, ACM, 2019, pp. 273–278, [Online] https://github.com/acmmmsys/2019-360dataset.

[78] M. F. R. Rondón, L. Sassatelli, R. Aparicio-Pardo, F. Precioso, A unified evaluation framework for head motion prediction methods in 360° videos, in: Proceedings of the 11th Multimedia Systems Conference, ACM, 2020, pp. 279–284, [Online] https://gitlab.com/miguelfromeror/head-motion-prediction/tree/master.

[79] A. Dharmasiri, C. Kattadige, V. Zhang, K. Thilakarathna, Viewport-aware dynamic 360° video segment categorization, in: Proceedings of the 31st Workshop on Network and Operating Systems Support for Digital Audio and Video, ACM, 2021, pp. 114–121, [Online] https://github.com/theamaya/Viewport-Aware-Dynamic-360-Video-Segment-Categorization.

[80] Y. Bao, H. Wu, T. Zhang, A. A. Ramli, X. Liu, Shooting a moving target: Motion-prediction-based transmission for 360-degree videos, in: International Conference on Big Data, IEEE, 2016.

[81] Y. Guan, C. Zheng, X. Zhang, Z. Guo, J. Jiang, Pano: Optimizing 360° video streaming with a better understanding of quality perception, in: Proceedings of the ACM Special Interest Group on Data Communication, 2019, pp. 394–407.

[82] J. Chakareski, R. Aksu, V. Swaminathan, M. Zink, Full UHD 360-Degree Video Dataset and Modeling of Rate-Distortion Characteristics and Head Movement Navigation, in: Proceedings of the 12th Multimedia Systems Conference, ACM, 2021, pp. 267–273.
URL https://alabama.app.box.com/v/8k-360-dataset

[83] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, Y. Wang, A subjective study of viewer navigation behaviors when watching 360-degree videos on computers, in: International Conference on Multimedia and Expo, IEEE, 2018, pp. 1–6.

[84] M. V. d. Broeck, F. Kawsar, J. Schöning, It's all around you: Exploring 360° video viewing experiences on mobile devices, in: Proceedings of the 25th International conference on Multimedia, ACM, 2017, pp. 762–768.

[85] M. Almquist, V. Almquist, V. Krishnamoorthi, N. Carlsson, D. Eager, The prefetch aggressiveness tradeoff in 360° video streaming, in: Proceedings of the 9th Multimedia Systems Conference, ACM, 2018, pp. 258–269.

[86] K. Srivastava, R. C. Das, S. Chaudhury, Virtual reality applications in mental health: Challenges and perspectives, Vol. 23, Wolters Kluwer–Medknow Publications, 2014.

[87] R. F. Martin, P. Leppink-Shands, M. Tlachac, M. DuBois, C. Conelea, S. Jacob, V. Morellas, T. Morris, N. Papanikolopoulos, The Use of Immersive Environments for the Early Detection and Treatment of Neuropsychiatric Disorders, Frontiers in Digital Health 2 (2021) 40.

[88] S. Petrangeli, G. Simon, V. Swaminathan, Trajectory-based viewport prediction for 360-degree virtual reality videos, in: International Conference on Artificial Intelligence and Virtual Reality, IEEE, 2018, pp. 157–160.

[89] S. Atev, G. Miller, N. P. Papanikolopoulos, Clustering of vehicle trajectories, IEEE transactions on intelligent transportation systems 11 (3) (2010) 647–657.

[90] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in neural information processing systems, 2002, pp. 849–856.

[91] L. Xie, X. Zhang, Z. Guo, CLS: A cross-user learning based system for improving QoE in 360-degree video adaptive streaming, in: Proceedings of the 26th international conference on Multimedia, ACM,

2018, pp. 564–572.

[92] S. Rossi, F. De Simone, P. Frossard, L. Toni, Spherical clustering of users navigating 360° content, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 4020–4024.

[93] T. Xu, B. Han, F. Qian, Analyzing viewport prediction under different VR interactions, in: Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies, ACM, 2019, pp. 165–171.

[94] S. Rossi, L. Toni, Understanding user navigation in immersive experience: an information-theoretic analysis, in: Proceedings of the 12th International Workshop on Immersive Mixed and Virtual Environment Systems, ACM, 2020, pp. 19–24.

[95] F. Qian, L. Ji, B. Han, V. Gopalakrishnan, Optimizing 360° video delivery over cellular networks, in: Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges, ACM, 2016, pp. 1–6.

[96] S. Petrangeli, V. Swaminathan, M. Hosseini, F. De Turck, An HTTP/2-Based Adaptive Streaming Framework for 360° Virtual Reality Videos, in: Proceedings of the 25th international conference on Multimedia, ACM, 2017, pp. 306–314.

[97] D. V. Nguyen, H. T. Tran, A. T. Pham, T. C. Thang, An optimal tile-based approach for viewport-adaptive 360-degree video streaming, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9 (1) (2019) 29–42.

[98] A. T. Nasrabadi, A. Mahzari, J. D. Beshay, R. Prakash, Adaptive 360-degree video streaming using scalable video coding, in: Proceedings of the 25th international conference on Multimedia, ACM, 2017, pp. 1689–1697.

[99] Y. Ban, L. Xie, Z. Xu, X. Zhang, Z. Guo, Y. Wang, Cub360: Exploiting cross-users behaviors for viewport prediction in 360 video adaptive streaming, in: International Conference on Multimedia and Expo, IEEE, 2018, pp. 1–6.

[100] A. T. Nasrabadi, A. Samiei, R. Prakash, Viewport prediction for 360° videos: a clustering approach, in: Proceedings of the 30th Workshop on Network and Operating Systems Support for Digital Audio and Video, ACM, 2020, pp. 34–39.

[101] M. Hu, J. Chen, D. Wu, Y. Zhou, Y. Wang, H.-N. Dai, TVG-Streaming: Learning User Behaviors for QoE-Optimized 360-Degree Video Streaming, IEEE Transactions on Circuits and Systems for Video Technology (2020).

[102] L. Xie, Z. Xu, Y. Ban, X. Zhang, Z. Guo, 360ProbDASH: Improving QoE of 360 Video Streaming Using Tile-based HTTP Adaptive Streaming, in: Proceedings of the 25th International conference on Multimedia, ACM, 2017, pp. 315–323.

[103] Z. Xu, X. Zhang, K. Zhang, Z. Guo, Probabilistic viewport adaptive streaming for 360-degree videos, in: International Symposium on Circuits and Systems, IEEE, 2018, pp. 1–5.

[104] F.-Y. Chao, C. Ozcinar, A. Smolic, Transformer-based Long-Term Viewport Prediction in 360° Video: Scanpath is All You Need, in: 23th International Workshop on Multimedia Signal Processing, IEEE, 2021.

[105] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ACL, 2020, pp. 38–45.

[106] A. D. Aladagli, E. Ekmekcioglu, D. Jarnikov, A. Kondoz, Predicting head trajectories in 360° virtual reality videos, in: International Conference on 3D Immersion, IEEE, 2017, pp. 1–6.

[107] A. Nguyen, Z. Yan, K. Nahrstedt, Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction, in: Proceedings of the 26th international conference on Multimedia, ACM, 2018, pp. 1190–1198.

[108] M. F. R. Rondon, L. Sassatelli, R. Aparicio-Pardo, F. Precioso, Track: A new method from

a re-examination of deep architectures for head motion prediction in 360-degree videos, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[109] X. Zhang, G. Cheung, Y. Zhao, P. Le Callet, C. Lin, J. Z. G. Tan, Graph learning based head movement prediction for interactive 360 video streaming, IEEE Transactions on Image Processing 30 (2021) 4622–4636.

[110] C.-L. Fan, J. Lee, W.-C. Lo, C.-Y. Huang, K.-T. Chen, C.-H. Hsu, 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, Association for Computing Machinery, New York, NY, USA, 2017, p. 67–72.

[111] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, S. Gao, Gaze prediction in dynamic 360° immersive videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5333–5342.

[112] X. Feng, V. Swaminathan, S. Wei, Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking, Vol. 3, ACM, 2019, pp. 1–22.

[113] L. Chopra, S. Chakraborty, A. Mondal, S. Chakraborty, PARIMA: Viewport Adaptive 360-Degree Video Streaming, in: Proceedings of the Web Conference, ACM, 2021, pp. 2379–2391.

[114] J. Park, M. Wu, K.-Y. Lee, B. Chen, K. Nahrstedt, M. Zink, R. Sitaraman, SEAWARE: Semantic Aware View Prediction System for 360-degree Video Streaming, in: International Symposium on Multimedia, IEEE, 2020, pp. 57–64.

[115] S. Park, M. Hoai, A. Bhattacharya, S. R. Das, Adaptive streaming of 360-degree videos with reinforcement learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1839–1848.

[116] A. Farhadi, J. Redmon, Yolov3: An incremental improvement, in: Computer Vision and Pattern Recognition, Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–2767.

[117] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, A. Smolic, Towards Audio-Visual Saliency Prediction for Omnidirectional Video with Spatial Audio, in: International Conference on Visual Communications and Image Processing, IEEE, 2020, pp. 355–358.

[118] T. L. Pedro Morgado, Nuno Vasconcelos, O. Wang, Self-supervised generation of spatial audio for 360° video, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, ACM, 2018, p. 360–370.

[119] T. Xue, A. E. Ali, T. Zhang, G. Ding, P. Cesar, RCEA-360VR: Real-time, Continuous Emotion Annotation in 360° VR Videos for Collecting Precise Viewport-dependent Ground Truth Labels, in: Proceedings of the International Conference on Human Factors in Computing Systems 2021, ACM, 2021, pp. 1–15.

[120] L. J. Zheng, J. Mountstephens, J. Teo, Four-class emotion classification in virtual reality using pupillometry, Journal of Big Data 7 (1) (2020) 1–9.

[121] X. Wei, C. Yang, FoV Privacy-aware VR Streaming (2021). arXiv:2110.10417.

[122] M. Quach, G. Valenzise, D. Tian, F. Dufaux, Geometry-based pcc + video-based pcc, in: Immersive Video Technologies, Elsevier, 2022.