Do Users Behave Similarly in VR? Investigation of the User Influence on the System Design

SILVIA ROSSI, University College London (UCL), UK
CAGRI OZCINAR, Trinity College Dublin (TCD), Ireland
ALJOSA SMOLIC, Trinity College Dublin (TCD), Ireland
LAURA TONI, University College London (UCL), UK

With the overarching goal of developing user-centric Virtual Reality (VR) systems, a new wave of studies focused on understanding how users interact in VR environments has recently emerged. Despite the intense efforts, however, current literature still does not provide the right framework to fully interpret and predict users trajectories while navigating in VR scenes. This work advances the state-of-the-art on both the study of users' behaviour in VR and the user-centric system design. In more details, we complement current datasets by presenting a public available dataset that provides navigation trajectories acquired for heterogeneous omnidirectional videos *and* different viewing platforms, namely, head-mounted display, tablet and laptop. We then present an exhaustive analysis on the collected data, to better understand navigation in VR across users, content, and for the first time across viewing platforms. The novelty lies in the user-affinity metric, proposed in this work to investigate users' similarities when navigating within the content. The analysis reveals useful insights on the effect of device and content on the navigation, which could be precious considerations from the system design perspective. As a case study of the importance of studying users' behaviour when designing VR systems, we finally propose an user-centric server optimisation. We formulate an integer linear program that seeks the best stored set of omnidirectional content that minimises encoding and storage cost while maximises the user's experience. This is posed while taking into account network dynamics, type of video content, but also user population interactivity. Experimental results prove that our solution outperforms commonly company recommendations in terms of experienced quality but also in terms of encoding and storage, achieving a saving up to 70%. More importantly, we highlight a strong correlation between the storage cost and the user-affinity metric, showing the impact of the latter in the system architecture design.

CCS Concepts: • **Information systems** → **Multimedia databases**; *Multimedia streaming*; Storage management; • **Human-centered computing** → **User studies**; *Virtual reality*; • **Mathematics of computing** → *Mathematical optimization.*

Additional Key Words and Phrases: Omnidirectional video dataset, user behaviour analysis, integer linear program, viewport-based adaptive streaming

ACM Reference Format:
Silvia Rossi, Cagri Ozcinar, Aljosa Smolic, and Laura Toni. 2020. Do Users Behave Similarly in VR? Investigation of the User Influence on the System Design. *ACM Trans. Multimedia Comput. Commun. Appl.* , (2020), 26 pages. https://doi.org/10.1145/3381846

Authors' addresses: Silvia Rossi, s.rossi@ucl.ac.uk, University College London (UCL), London, UK; Cagri Ozcinar, ozcinarc@scss.tcd.ie, Trinity College Dublin (TCD), Dublin, Ireland; Aljosa Smolic, smolica@scss.tcd.ie, Trinity College Dublin (TCD), Dublin, Ireland; Laura Toni, l.toni@ucl.ac.uk, University College London (UCL), London, UK.

1551-6857/2020/-ART $15.00
https://doi.org/10.1145/3381846

1 INTRODUCTION

Since its conception 50 years ago, Virtual Reality (VR) technology has been increasingly developed, with a disruptive impact envisioned across many sectors, such as gaming and entertainment, but also healthcare, education, sport, journalism and automotive [48]. The revolutionary novelty introduced by VR is the possibility to interact with the content provided to users, empowering viewers with a feeling of engagement and presence in the virtual space, even if they are not physically there [45, 48]. This immersion sensation is provided to users by a new multimedia format, namely the omnidirectional video (ODV) or spherical video, defined as a visual signal depicting the 360° surrounding scene on a virtual sphere. The viewer, virtually positioned at the centre of the virtual space, dynamically interacts with the content and changes the rendered portion of the spherical content (*i.e.*, *viewport*) based on his/her viewing direction. Typically, users enjoy this new multimedia format by head-mounted display (HMD) or tablet/smartphone. The dynamic level of interactivity and the ability for viewers to display only a desired portion of the content has pushed toward a technological paradigm shift, in which the user is at the center of the content consumption (as opposed to more classical fully passive content consumption). This ensures immersion, presence and interactivity, which are the three crucial factors to guarantee high Quality of Experience (QoE) in a VR system [19]. However, this paradigm shift has introduced few main challenges: VR systems are highly data intensive and require ultra-low latency. Both requirements imply a very high amount of data to be transmitted in real time for the millions of VR users envisioned in the near future, also pushing connectivity boundaries. To overcome these limitations and challenges, VR systems need to evolve in a personalised manner, implying a fundamental revolution of the media delivery chain, from coding to rendering. The interactive user has to be put at the heart of the next generation of VR system rather than at the end of the chain.

It is therefore clear the urgent need for understanding and anticipating user's movements to develop *user-centric* solutions. This has been proven by an increasing interest on users' behaviour analysis and classification in VR [39, 43, 47]. Many public datasets have appeared presenting data of head and/or eye movements collected while viewers were displaying VR images/videos by HMD [7, 10, 17, 25, 26, 53], and computer [13]. Their focus has been mainly on the analysis and prediction of the most salient areas within the content. At the same time, research interests have expanded toward psychological and emotional aspects related with VR applications. Since ODVs can be experienced by heterogeneous apparatus, such as smartphones, tablet and HMD, recent psychological investigations on users' experiences suggest that viewers prefer different devices based on content category and his/her current location (*i.e.*, travelling or at home) [52]. Moreover, human perception is strongly dependent on the selected viewing platform [5]. From a technical perspective, the investigation of users' behaviour in relation with selected content and device could be the key to optimise the system design of VR applications. However, currently, this is not possible since the behaviour of interactive users across devices is highly overlooked in the literature.

To overcome this issue, this paper introduces a dataset of navigation trajectories of users watching 15 ODVs on different devices (HMD, tablet and laptop) and analyses the users' behaviour across content *and* across viewing device. As main novelty, we investigate different conditions of ODVs exploration based on the viewing device: traditional VR-based navigation enabled with HMD, touch-based navigation with tablet and mouse-based navigation with laptop. Based on this collected dataset, we then present an extensive user data analysis. A first analysis is carried out with conventional metrics such as angular velocity and viewport center distribution, and it highlights the dependency of the users navigation from the displaying device. However, this first part of the analysis fails in detecting how much users interact in harmony among themselves; key information to understand users predictability. Therefore, we expand the dataset analysis including a novel

metric aimed at evaluating the affinity among users – i.e., the similarity among them in terms of viewport displayed overtime. Namely, we introduce the *User Affinity Index* metric, which is based on a recent clique-based clustering tool [43]. This allows us to move a step forward in the direction of better understanding how users interact with the VR technology, with a substantial impact on the efficiency of VR systems. Our study is therefore vital for the community to design reliable user-centric systems, as the recently proposed coding and streaming strategies for ODV [55].

To emphasise the importance of the proposed dataset and associated analysis, we propose a case study on user-centric optimisation for coding and storing ODVs at the server. Due to the increasing cost of storage and coding, optimising the storage space at the main server has become fundamental, especially for VR content – highly data intensive. Works focused on server optimisation for classical adaptive streaming platforms have been already proposed in literature [51], tuning the coding rate and resolution depending on both the population features and the type of content. In the context of ODV, only [37] introduces a content-aware encoding ladder estimation that achieves cost-optimal and higher objective quality compared to recommended encoding ladders. However, information about users' navigation within the content is not considered. Hence, to carry out our case study we also bring-in the novelty of formulating a *user-centric server optimisation* for ODV adaptive streaming systems. In particular, we evaluate the optimal set of coding parameters to store ODVs at the main server minimising the total cost and maximising user's experience, taking into account the users' behaviour and network characteristics. Results show that our solution performs well in terms of total cost (i.e., encoding and storage cost) and quality experienced by users. Most importantly, results reveal also a correlation between the optimal set and the user affinity index. This insight suggests that user affinity index could be a key metric in the design of the next generation systems.

In conclusion, our work contributes to the overall open problem of optimally designing a VR system, with the following main contributions:

(i) A new public dataset of 15 ODVs with associate navigation trajectories collected in task-free experiments using 3 different devices such as HMD, tablet and laptop.
(ii) An exhaustive analysis of the aforementioned collected data, showing that users navigate differently based on the device, and introducing a novel affinity metric able to quantify user navigation similarities.
(iii) A case study of VR systems optimised from the server perspective, with a two-folds novelty: *i)* the proposed problem formulation; *ii)* the translation of the users' behaviour analysis into gain for a system provider.

The remainder of this paper is organised as follows. Related works on users' behaviour analysis and streaming strategies in VR system are reported in Section 2. The data collection campaign is described in Section 3 while the associated analysis is depicted in Section 4. The case study is formulated in Section 5. Section 6 and Section 7 describe metrics and simulation settings, respectively. In Section 8, the performance of the proposed optimisation algorithm is first compared with the set of recommended representations and then, the results are further analysed to reveal the effect of the user behaviour. Finally, conclusions are summarized in Section 9.

2 RELATED WORKS

Although streaming strategies have been widely investigated in recent decades, many open challenges are still unsolved in the context of user-centric immersive communications. We now describe the latest contributions mostly related to our work, which are focused on: *i*) analysis of users' behaviour within ODV; *ii*) user-depended streaming strategies for ODV. For a comprehensive literature review on ODV analysis and communication, we refer the reader to [55] and [14], respectively.

2.1 Studies related to ODV dataset and user behaviour analysis

The main contributions in the area of users' behaviour can be categorised in *i*) dataset collection; *ii*) analysis of the acquired dataset. Ideally, the data collection should be as much exhaustive as possible, while the data analysis should identify the hidden patterns of users while navigating. The collected users' head and eye movement data show the most salient regions of ODVs. In particular, head movement determines field of view (FoV) as the pixel region of ODV to be seen by the HMD over time. Eye movement datasets contain the regions of a given ODV that are salient within the FoV. To understand how people observe and explore ODVs, David *et al.* [10] established a dataset of head and eye movements using an HMD across various content types of ODVs. In their study, statistics related to ODV exploration behaviours are presented using the distribution of eye fixations. Salient360! grand challenges at the ICME 2017 and 2018 fostered further research works in this domain by providing benchmark platform for visual attention models [46]. For instance, Chao *et al.* [6] demonstrated state-of-the-art saliency prediction accuracy using generative adversarial networks. Furthermore, Zhang *et al.* [57] presented a large-scale HMD eye-tracking dataset using only sport-related ODVs. In their work, the performance of spherical convolutional neural network architecture is analysed with state-of-the-art image saliency detection methods. Similarly, the work in Ozcinar *et al.* [39] analysed the performance of standard video saliency detection methods using 6 ODVs rendered by an HMD. Results reveal that the quantity of fixations depends on motion complexity of ODV. A HMD-based Director's Cut dataset has been proposed in Knorr *et al.* [24] to evaluate the users' attention in storyteller ODV. An interactive storytelling perspective was then presented in [15, 16]. More recently, Nasrabadi *et al.* [31] investigated the impact of camera motion on HMD navigation trajectories using the clique-based clustering presented in [44]. Another dataset was established by Corbillon *et al.* [7] using 7 different ODVs viewed by an HMD. In their work, statistical analysis was performed by using maximum and average angular speeds of navigation trajectories under various video segment lengths. Similarly, Lo *et al.* [26] published a dataset intending to optimise ODV streaming for an HMD. This work, however, excluded the analysis of users' behaviour for ODV viewing. Furthermore, Wu *et al.* [53] investigated what content users remember after each viewing session on their proposed head orientation dataset. According to their analysis, users share certain common patterns in ODV streaming, which are different from those in conventional video streaming.

The above works have set a solid background for understanding users' behaviour in VR systems. However, they do not completely provide the right framework to develop fully personalised VR streaming solutions, which take into account both content and user device. Specifically, none of the above works provide an open-data set where users' trajectories are acquired with different displaying devices. Moreover, in most of the data analysis studies the popular metrics show an average behaviour of the users (for example, the mean angular velocity) but do not necessarily reveal quantitative information of users' similarity. Hence, the novelty of this paper is in tracking users navigation across three different VR devices and in analysing the acquired dataset with new users similarity metrics.

2.2 Studies related to viewport-oriented ODV streaming strategies

In recent years, user-centric systems have been developed, optimising every step of the ODV video delivery chain: coding [30], streaming [36, 44], caching [28, 29], and rendering [40]. In particular, tile-based coding systems [30, 34] were utilised using viewport adaptive streaming algorithms [8, 33, 38] to provide smooth VR video experience [36, 44]. For instance, a navigation-aware adaptation logic was developed in [44] to optimise the downloading rate for each tile of ODV streaming. The results reveal that the final quality is strongly affected by the video content and

users' navigation trajectories. Furthermore, Nguyen *et al.* [33] presented an adaptation logic for ODV streaming to decide an optimal version of each tile according to users' head movements and network bandwidth. Their analysis emphasised the need of accurately predicting future viewport position for ODV streaming. In the aspect of the prediction of future viewport, Petrangeli *et al.* [41] proposed a prediction algorithm for long-term prediction of user viewport. In their work, the navigation trajectories of a given user are modeled over time such that future viewports can be predicted based on the navigation patterns of users stored in the system. According to their results, their proposed algorithm can increase prediction accuracy of the expected viewport area by 13% on average compared to previous algorithms. By looking more at the server-side (i.e., coding optimisation), Ozcinar *et al.* [36] proposed a visual attention-based ODV streaming system optimising the tile-based design taking into account users saliency maps. The work showed the importance of being user-centric also at encoding side without focusing a design of cost-aware VR system. In contrast, Xiao *et al.* [54] optimized the tile-based encoding design of ODVs seeking the best trade-off between storage costs and overall quality of the panorama. However, the storage cost was not formally optimized and the users trajectories were neglected in the problem formulation. There are also some activities in the sense of standardization bodies, such as MPEG-I [50]. For instance, a practical study by Graf *et al.* [18] examined several adaptive streaming strategies and evaluated bitrate overhead with quality requirements in VR. To find the optimal set of quality-variable video versions for ODV streaming, Corbillon *et al.* [8] presented an optimisation model for the concept of quality regions of ODVs. Their main contribution is to consider the surface bitrate and users' head movement data within the proposed optimisation framework. However, their study was restricted to using the concept of quality-emphasised regions, with the employed constraints being the number of quality-variable video versions and the bandwidth. Also, Zue *et al.* [58] proposed a server-side rate adaptation problem for the tile-based adaptive ODV streaming. They aimed to maximise the QoE of multiple users who are competing for transmission resources at the network bottleneck. Furthermore, Chakareski *et al.* [23] maximised the QoE for given network resources at the server side. Their work consider user navigation trajectories and spatio-temporal rate-distortion characteristics of a given video. However, the proposed formulation is based on the traditional Mean Square Error (MSE), which does not take the spherical distortion of ODV representation into account. In summary, from the literature it is clear the importance and the gain in being user-, cost-, and geometry-aware when designing VR systems. However, such a complete design at the server side is missing.

Our work goes beyond the state-of-the-art as we take into account our users' behaviour analysis, formulating a novel user-centric server optimisation system, which minimize the user-centric spherical quality and the coding and storage costs. In particular, we developed an optimisation algorithm to determine the optimal set of coding parameters to store ODVs at the server minimising the total cost and maximising users experience. Differently from the aforementioned works, the main novelty of our algorithm is to take into consideration users' behaviour beyond the spherical geometry and content information, minimising the total cost and yet maximising the final quality for ODV adaptive streaming systems. A further novelty is to link the optimal design with the affinity of users navigation patterns.

3 COLLECTION OF USERS' NAVIGATION TRAJECTORIES

In this work, we are primarily interested in understanding users' navigation across space and time when interacting with different ODVs and the impact that different devices might have on the actual interaction. With this aim in mind, we collected a dataset with head-trajectories across different viewing platforms. In particular, we conducted subjective experiments across two universities, namely, Trinity College Dublin (TCD) and University College London (UCL). In this section, we

Table 1. Description of the ODVs used for the subjective experiment. The dataset contains three content categories (documentary, action, and movie). Each content category has a training ODV and five test ODVs.

	Dataset ID	Name	Fps	YouTube Id	Selected Segment
Documentary	**Test**	*WildDolphins*	25	BbT_e8lWWdo	00:44 – 01:04
	01	*BabyPandas*	24	0XrH2WO1Mzs	02:05 – 02:25
	02	*Symphony*	30	LZlNCAGWtwE	01:10 – 01:30
	03	*Ocean Shark*	24	aQd41nbQM-U	00:40 – 01:00
	04	*Dancing*	24	raCda6VRrE8	00:00 – 00:20
	05	*Survivorman*	30	OLQzLOd7Xpk	00:30 – 00-50
Action	**Test**	*LaRonde*	25	r-qmDDi8S5I	00:10 – 00:30
	06	*FighterJet*	25	NdZ02-Qenso	00:00 – 00:20
	07	*HollywoodRockit*	25	Js_Jv5EzOv0	00:10 – 00:30
	08	*GetBarreled*	30	7gjR60TSn8Q	01:22 – 01:42
	09	*KITZ*	30	KS9S1Hgx2co	00:00 – 00:20
	10	*Knockout*	30	0x16ngo8xfY	01:22 – 01:42
Movie	**Test**	*Starwars*	25	SeDOoLwQQGo	02:23 – 02:43
	11	*Back2theMoon*	30	BEePFpC9qG8	00:11 – 00:31
	12	*Help*	30	G-XZhKqQAHU	01:20 – 01:40
	13	*Nick*	24	Au5ro1NOnh	03:25 – 03:45
	14	*Invasion*	25	QolJrTXr7PA	00:44 – 01:04
	15	*InvisibleMan*	25	I_FUpUi2LBk	01:55 – 02:15

describe the technical details of the experiments. The collected navigation trajectories and the used tools are shared in a public repository[1] under the MIT open source license.

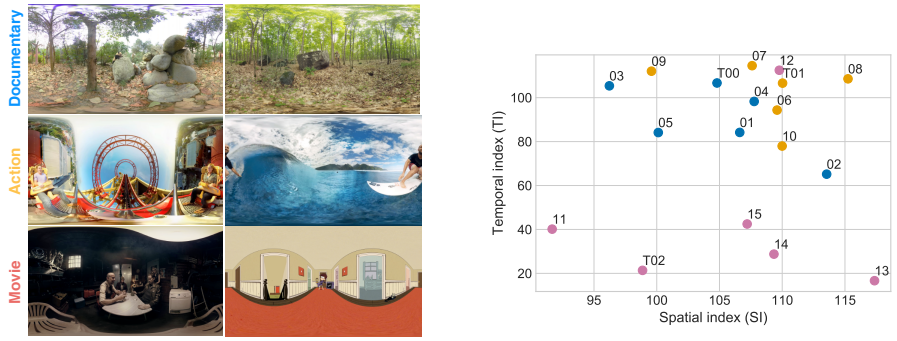
3.1 Material

To ensure diversity in terms of content, we selected 18 ODVs with diverse content characteristics and representative of three video categories: *Documentary*, *Action*, and *Movie*. These categories are diverse enough to maximise the number of subjective experiments to carry out, and yet they span various content characteristics. Moreover, these categories are widely used in the classification of ODV content types. Fig. 1 (a) depicts a snapshot of two randomly sampled ODV for each category. Specifically, we selected videos to span a wide range of content characteristics, such as spatial and temporal complexities. Fig 1 (b) visually reflects the diversity that each video exhibits in terms of spatial and temporal information measures [22], SI and TI, respectively.

Each ODV was downloaded from YouTube in the Equirectangular Projection (ERP) format at the maximum available bitrate and resolution, which is 2560×1440. These ODVs were selected by a consideration of downloading ODVs with high quality. Then, a visual segment of 20 *sec.* duration was extracted from each video, and the audio signal was discarded from each ODV. Our work focuses only visual (texture) part of ODV by ignoring audio in every step of the delivery pipeline. In particular, we are interested at studying the effect of visual content on the trajectories, which has been the case in many other related works (*e.g.*, [7, 10, 39]).

Each 20 *sec.* segment was selected in a pilot test with two experts. The experts selected the 20 *sec.*, making sure that the selection exhibits its content category and contains at least one salient object. This duration was chosen as it is the most commonly used in visual attention studies [10]; specifically, it is a meaningful duration for the visual attention experiments as it is long enough for users to engage with the content, and yet short enough to maximise the number of experiments to carry out. Finally, an ODV from each category (out of the 18 ODVs) was used as training content for participants to familiarise with the setup of each device. Table 1 summarises characteristics of ODVs used in this work, where *Test* denotes the training content, one for each category.

[1]https://v-sense.scss.tcd.ie/research/3dof/vr_user_behaviour_system_design/



(a) ODV Dataset: Sample thumbnail frames of the ODVs (two per category) used in the experiments.

(b) Spatial information (SI) and temporal information (TI) [22] of ODVs used in subjective experiments. Each category is visualised using different color: Blue: documentary; yellow: action, pink: movie.

Fig. 1. Sample frames and statistics for the used ODVs in this work.

3.2 Apparatus

We modified the JavaScript-based test-bed developed in [39] allowing users to display ODVs on three different devices, namely, HMD, laptop, and tablet, while recording their navigation (viewport) trajectories for the whole duration of the experiment. The developed test-bed records participants' viewport positions with the current time-stamp and ODV name. Here, a given set of ODVs is first loaded using the playlist file, and a given video is played while the recorded data is transmitted to the server with the refresh rate of the device's graphics card. At the server side, the HTTP server was implemented using the Apache web server with the MySQL database, where the device-related (*e.g.*, HMD, laptop, and tablet), sensor-related (*e.g.*, viewing direction), and user-related (*e.g.*, user ID, age, and gender) data are stored on the database.

We conducted ODV subjective experiments with VR-based navigation enabled with HMD, touch-based navigation with tablet and mouse-based navigation with laptop. As HMD, we used the Oculus Rift consumer version that allows rendering of scene with a nearly 110 FoV at 90 Hz refresh rate. Each ODV is displayed in the HMD using the Firefox Nightly (*ver.* 67.0a1) Web browser. Finally, Alienware 15 Gaming Laptop and Apple iPad Pro 10.5 tablet were used. In both devices, we utilised Google Chrome (*ver.* 71.0.3578.98) as a web browser to play ODVs. We considered two different web browsers due to hardware and video codec compatibility issues at the time of subjective experiments.

3.3 Participants

In all, 94 participants (65 males and 29 females - about 30% women) took part in our subjective experiments. Participants were aged between 21 to 54, with an average of 31 years. Nine of the participants (about 10%) were familiar with ODV, and the others were naive viewers. Furthermore, 43 participants wore glasses during the experiment, and all of the viewers were screened and reported normal or corrected-to-normal visual acuity. Each participant watched a total of 18 ODVs (5 test ODVs plus 1 training ODV per device).

3.4 Viewing procedure

To ensure diversity in participants (*e.g.*, ODV familiarity) and maximise the number of navigation trajectories, we performed the data collection campaign using the same apparatus at TCD and UCL. Each subjective test was performed as *task-free* viewing sessions in laboratory condition,

where each participant was asked to naturally look at each ODV. The task-free viewing is the most common procedure for analysing visual attention [11, 39]. Participants were seated in a swivel chair and allowed to turn freely. During experiments, participants were alone in the environment to avoid any influences given by the presence of instructor.

The subjective test was divided into three phases, where in each phase a viewing session with a different device was conducted. The order of the devices was in a random order. To get familiar with the new device, a training video was displayed at the beginning of each phase (one for each device). Then, after the test session, 5 test ODVs were played in a random order while the individual navigation trajectories were recorded using the implemented test-bed. To avoid motion sickness and eye fatigue, we inserted a 5 *sec.* rest period with a grey screen between two successive ODVs. Also, before playing each video, we reset the sensor to return to the centre of the ERP. In total, each viewing session lasted 2 *min.* and 25 *sec.*. We also set a 3 min break between each viewing session.

To ensure both the balance among the collected dataset (*i.e.*, balanced amount of viewport trajectories per device per video) and that each user watches each video only once, a set of playlists was prepared. Each playlist included a training and 5 test ODVs per device, and in total there were three different playlists for the three different phases of the test (*e.g.*, three different devices). The total number of 15 test videos were thus divided into three playlists, and each user was shown three playlists for three different devices. These playlists were randomly selected for each user at the beginning of the subjective experiment. It is worth noting that the avoidance of repetition of the same video within the same playlist avoids the memory bias effect, that could affect the navigation trajectories [42]. Therefore, during one experiment, a user switches devices every 5 ODVs. In total, she/he watches 15 different ODVs.

3.5 Post-processing

In order to analyse the collected navigation trajectories, all recorded data was re-sampled based on the frame rate of the corresponding video. In this way, a fair comparison is allowed having a single value per user in each frame. Since roll movements are permitted only with HMD, our following investigations are based only on viewport's movements in longitude and latitude coordinates. Previous works [1] showed that most of the users' movements happen mainly along with these directions and the roll movements are at minimum. Therefore, this choice will not compromise the validity of the analysis presented in the following.

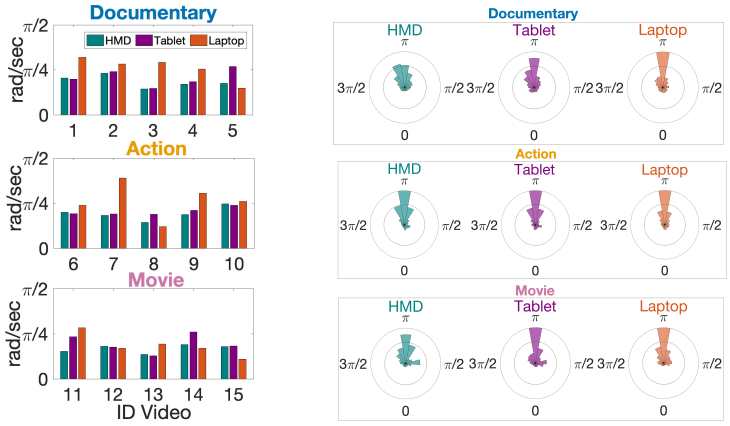
4 USER BEHAVIOUR ANALYSIS

We now present an analysis of the collected navigation trajectories across video content and devices. Specifically, we propose two lines of analysis: one more traditional aimed to show similar features of navigation among users, and the second focused on quantifying similarity among users' behaviour. Over the entire section, we will also underline the key insights that we observe when users navigate in different video categories and with different devices.

4.1 A Conventional Data Analysis

We take the liberty to denote this first analysis of the collected trajectories as "conventional" data analysis since we adopt well know metrics such as angular velocity and spatial distribution of viewport center. Here, the key novelty is to investigate the users' behaviour across categories *and* devices, leading to the following observations (supported in the remaining of the subsection)

- **Observation 1**: Users tend to be more dynamic with laptop compared to other devices.



(a) Angular velocity (b) Viewport center distribution

Fig. 2. Traditional analysis of users' behaviour across devices and video categories. (a) Angular velocity per video and device - Video ID refers to Table 1. (b) Viewport center distribution on the longitude direction per video category and device.

- **Observation 2**: In contents characterised by a dominant focus of attention, the level of interactivity is negligibly affected by the displaying device (highlighting the dominance of the focus of attention).
- **Observation 3**: On the contrary, in contents with no focus of attention, users have highly exploratory trajectories and, a strong dynamic with HMD.

In Fig. 2(a), we analyse the users' behaviour via the mean angular velocity per user for different devices and video categories. This analysis reveals the dynamicity of users navigation, measuring how fast each participant moves his/her head inside a given ODV. It is worth noting that the angular velocity is typically lower using HMD rather than other devices; on the contrary, users experience the highest mean angular speed when displaying ODV on laptop. This can be motivated by the physical constraints imposed by HMDs (*i.e.*, limited head movements), but also by a deeper feeling of immersion experienced with HMD compared to the laptop. Implicitly, a drop of attention or immersion sensation leads to a more scattered navigation paths. Authors in [27] show how film editing and style influence user gaze movements during the vision of standard 2D movies. Therefore comparing different video categories in our analysis, we can observe a slower angular velocity for users displaying *Movie* videos. This confirms that film maker manages to drive users' visual attention toward the main subject of interest also in ODVs. On the contrary, *Documentary* videos usually lack of a main focus of attention; hence, viewers tend to explore more the content.

Beyond the velocity of participants' movements, we are also interested in detecting the areas of saliency, i.e., the most interesting areas in which users look at. Fig. 2(b) shows the distribution of viewport centers in the longitudinal direction for all video categories across devices. Each slice, spanning a $\pi/10$ angle, represents the popularity of each direction across the entire video. The extension of each slice is proportional to the times in which - on average - users centred their viewports in the longitude direction identified by the slice. In particular, a single triangle predominant over others reflects that most of the users tend to center their displayed viewport in the same region of the ODV, identifying a clear focus of attention. From the figure, it is evident that the privileged area in terms of longitude is not really affected by video category and device. Viewers indeed tend to spend most of the time in a restricted portion of the central area around

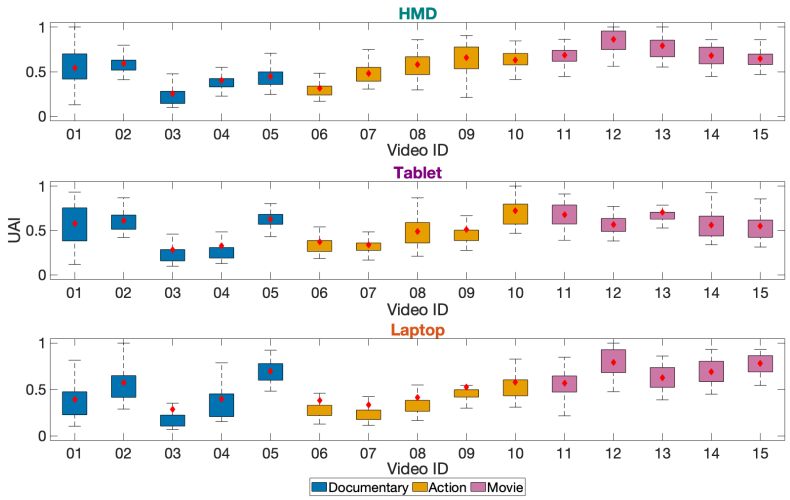


Fig. 3. Boxplots per viewing device of Users' Affinity Index (UAI) for each video in the dataset. The lower and upper side of the rectangular represents 25% and 75% percentile, respectively. While diamond is the mean value of User Affinity Index (UAI) per the entire video.

π in all different settings and video of the test. As expected, this is evident in *Action* and *Movie* categories, while less present in the *Documentary* contents, that usually have a less dominant focus of attention. In this latter case, the interaction is device-dependant, with a more spread distribution of viewport's centers with HMD when compared to laptop and devices.

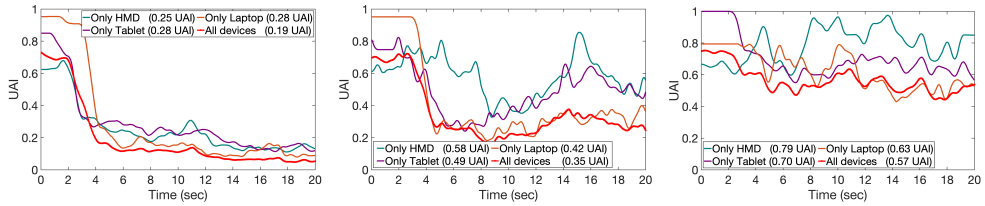
4.2 Looking for Users' Similarities

The metrics studied in Section 4.1 reveal general and useful features of users' behaviour, however they do not necessary provide an answer to one simple and yet crucial question: *"Can we predict users' behaviour?"*. Without pretending to fully answer to this question with the following data analysis, we truly believe that a key information to grasp is *"Do users behave similarly?"*. This is the key as users with poor similarity in the navigation are highly challenging to predict. This motivates the following analysis, aimed at identifying behaviour similarities among users, across video content and/or devices; hence, the importance of developing metrics able to capture this information. Specifically, we analyse our dataset with the clique-based clustering algorithm presented in [44], which is able to identify users clusters based on their consistency in the navigation. In practice, the algorithm detects and puts together users that consistently display similar viewports over time while consuming the ODV content. Also, this is done by taking into account the spherical geometry of the ODVs. We therefore introduce a novel metric (based on the clique-based clustering algorithm) to better reflect similarity among users' navigation trajectories within the same given ODV. We define this metric as the *User Affinity Index* (UAI), given as follows:

$$UAI = \frac{\sum_{i=1}^{C} x_i \cdot w_i}{\sum_{i=1}^{C} w_i} \tag{1}$$

where C is the number of clusters detected in a frame by the clique-clustering[2], x_i is the % of users (*i.e.*, out of the whole population/users sampled) in cluster i and w_i is the number of users in cluster i. In other words, the UAI represents the weighted average of cluster popularity (*i.e.*, how many users

[2]The clique-based clustering is applied with a geodesic distance threshold equal to $\pi/8$.



(a) Documentary (Video ID 03) (b) Action (Video ID 08) (c) Movie (Video ID 13)

Fig. 4. User Affinity Index (UAI) over time for three different videos (one per category) and for all devices. The mean value over time is reported on bracket in the legend for each analysed clustering condition.

per cluster). The UAI approaches 1 when a small number of clusters with a large number of users per cluster are detected. This shows high affinity among users (*i.e.*, users share strong similarity in how they navigate the content). On the contrary, UAI tends towards 0 when participants experience highly scattered navigation patterns, and they cannot be clustered together. In the Appendix A, we compare our proposed metric for detecting similarities in users' behaviour over time with a well-known metric (*i.e.*, entropy of saliency map [12]).

Fig. 3 shows the range and mean value (*i.e.*, box and red diamond, respectively) of UAI distributions obtained for each ODV of the entire database. Different behaviour can be identified based on the device and the category of video. For instance, the affinity for *Documentary* videos is lower than the one experienced with ODVs from the *Movie* category. We can also generalise that the navigation affinity within *Documentary* videos is not really influenced by the viewing device. On the contrary, HMD enable users to enjoy very similar experiences within ODVs, mainly for *Movie* and *Action* sequences. For example, users that display *Action* video with HMD have an UAI higher than 0.5 (except for Video ID 06). These findings are strongly evident in Fig. 4 that shows the UAI over time for three selected videos, one per category (*i.e.*, ID 03, 08 and 13). After an initial phase where most of the users are focused on the same area, people start exploring the scene and behave differently based on the content (or video category). Specifically, in *Documentary* sequence (ID 03) users have a very low affinity, while they navigate in a much more compact way in *Movie* video (ID 13) leading to higher UAI for all devices. Moreover, HMD leads to more similar navigation paths compared to the laptop, see Fig. 4(b) and Fig. 4(c). Finally as a further comparison, we also apply the clique-clustering to all the recorded data without distinguishing them based on viewing device. We then evaluate the corresponding UAI (labelled as "All devices" in Fig. 4) and notice that the affinity drops drastically, with respect to the case in which the clusters were formed per device. The "All devices" curve seems to be a worst-case scenario, showing that the users navigation has a strong affinity when looking at data from the same viewing device but this affinity drops when analysing data for the same content but across devices.

In summary, from this second analysis we can conclude the following:

- **Observation 4**: In content with no main focus of attention, users experience a low affinity, which is interestingly not perturbed by the viewing device.
- **Observation 5**: Users tend to explore content characterised by a dominant focus of attention in a very similar way.
- **Observation 6**: In content with a main focus of attention, the user affinity is strongly related to the selected viewing device. In particular, the HMD leads to quite similar navigation among users.

These outcomes highlight the importance of studying navigation trajectories in VR systems per viewing platform. Specifically, we argue that similar users' behaviours (*i.e.*, high value of affinity) identify predictable patterns that can be used to properly optimise user-centric streaming systems.

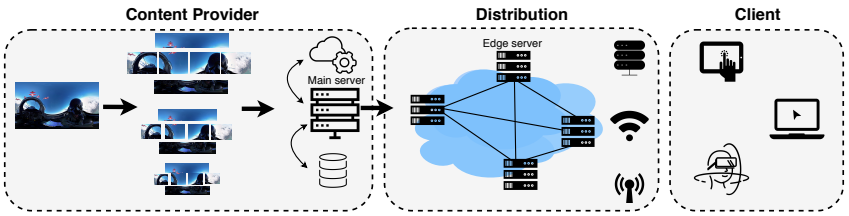


Fig. 5. Schematic of the adopted tile-based adaptive ODV streaming system.

In the following, we show the impact of this study when applied to adaptive streaming strategies for VR, with a focus on a user-centric server optimisation design.

5 CASE STUDY: USER-CENTRIC SERVER OPTIMISATION

We now show the importance of considering users' behaviour when designing an ODV streaming system defining a *user-centric server optimisation* that considers multiple VR devices. In particular, we focus on optimising the set of tile-representations to store at the server, considering spherical geometry, content complexity of ODVs and network capacity beyond users' navigation features. First, we introduce the system model for the tile-based adaptive ODV streaming scenario adopted in this work. Then, we formulate an Integer Linear Programming (ILP) used to evaluate the optimal set of tile-representations that maximises the quality perceived by users while minimises the total cost of encoding and storage. In Table 2, we summarise the main notations adopted in the following problem formulation.

5.1 System model

Fig. 5 illustrates the adopted tile-based adaptive ODV streaming system. Namely, each video sequence is spatially decomposed into tiles, which are encoded at different coding rates and resolutions. The generated representations of each tile are then temporally segmented into chunks of fixed duration (*i.e.*, typically 2 *sec.*) and stored at the main server. Out of the many representations stored at the server, only one per tile is actually distributed through edge servers to the final user. The selection of the representation is usually performed at the client side. Specifically, any final users, while navigating inside an ODV, will periodically requests to download the most suitable set of tile-representations (*i.e.*, such that to cover the entire panorama), based on the available bandwidth and his/her current position inside the ODV – usually the best quality that meets bandwidth constraints. In particular, we consider users downloading the entire panorama at each downloading opportunity but at heterogeneity quality levels. Specifically, the more probable a tile is the higher quality at which it is downloaded. In this contest, we are interested in investigating how to design an optimal representations set at the server side able to satisfy the requests from a potential VR population.

More formally, let $\mathcal{V}$ be the set of ODVs available at the *main server*. Each video $v \in \mathcal{V}$ is decomposed into N tiles. We denote by $j \in \mathcal{J}_v = 1, 2, .., N$ the set of tiles belonging to v. Then, each tile is encoded independently into different representations characterised by bitrate levels, $r \in \mathcal{R}$ and, spatial resolutions $s \in \mathcal{S}$. Note that $\mathcal{R}$ and $\mathcal{S}$ are sets of admissible bitrates and spatial resolution values. All variables v, j, r and s are integer values that represent the index in their corresponding set. In particular, the nominal value (in *kbps*) of the encoding rate r is denote by b_r and $\mathcal{B}$ is the set of available bitrates. Each representation is temporally divided into chunks of a fixed duration. Let $\mathcal{L}^v = \{(j, r, s) | j \in \mathcal{J}_v, r \in \mathcal{R}, s \in \mathcal{S}\}$ be the set of *representations* per chunk of a video $v \in \mathcal{V}$; the triple (j, r, s) indicates the representation of tile j encoded at bitrate r and resolution s.

Table 2. Notation adopted in the problem formulation.

Name	Description
$\mathcal{U}, u \in \mathcal{U}$	set of all users' type and the actual user served in the system, respectively
$\mathcal{V}, v^u \in \mathcal{V}$	set of ODV and video content requested by user u, respectively
$\mathcal{J}_v, j \in \mathcal{J}_v$	set of all tiles of the video v and the selected j tile, respectively
$\mathcal{R}, r \in \mathcal{R}$	set of all possible coding rate and the actual coding rate at which a tile can be encoded, respectively
$\mathcal{B}, b_r \in \mathcal{B}$	set of all available values of encoding rate and the nominal value of r (kbps),
$\mathcal{S}, s \in \mathcal{S}$	set of possible spatial resolution and actual spatial resolution s at which a tile j can be encoded,
(j, r, s)	representation of a tile j encoded at rate r and spatial resolution s,
$\mathcal{L}^v$	set of all possible tile-representations for a video v,
$\mathcal{T}^* \subseteq \mathcal{L}$	optimal set of representations stored at the main server,
$\mathcal{D}, d^u \in \mathcal{D}$	set of available device and actual device selected by user u
$\mathcal{S}, s^u \in \mathcal{S}$	set of available spatial resolution (i.e., screen size) and actual resolution of device selected by user u
$\mathcal{N}, n^u \in \mathcal{N}$	set of available networks and actual network selected by user u, respectively
BW^u	available bandwidth throughput for user u,
p_j^u	probability of tile j to be displayed by users of type-u,
δ^u	portion of users of type-u,
$D_j^u(r, s)$	distortion value of tile j encoded at rate r and resolution s requested by user of type-u
$C^{TOT}(r, s)$	total costs (encoding and storage costs) for a tile-representation encoded at rate r and resolution s

Given the heterogeneity of users downloading ODV (*i.e.*, different type of network and devices), all the possible representations $\bigcup_v \mathcal{L}_v$ should be stored at the main server. This would ensure to serve each users' request at the best. In practice, coding and storage costs can be unbearable when all representations are stored. Hence, the need to select a subset of representations $\mathcal{L} \subseteq \bigcup_v \mathcal{L}_v$ to store at the main server. Our goal is then to seek the optimal subset $\mathcal{T}^*$ to be available at the server in order to maximise the QoE given constraints from both the server and client perspectives. We argue that in this system design optimisation, the knowledge of displaying device and video category as well as the user navigation trajectories is the key for any efficient optimal set.

Let $\mathcal{U}$ be the set of all clients served in our ODV streaming system. We assume that all final users can be categorised based on the selected video content, viewing device and the kind of network connection (*i.e.*, capacity of each user connection). Namely, a user of type $u \in \mathcal{U}$ is defined by the desired video $v^u \in \mathcal{V}$ displayed at the resolution of the selected device $m^u \in \mathcal{M}$, downloaded based on the kind of network $n^u \in \mathcal{N}$. Without loss of generality, we make the assumption that each device is associated with a single display resolution. The type of network n selected by user u defines the range of available throughput value BW^u. Finally, each type of users has an own navigation path inside the ODV, that depends on the selected device d^u as well as the content and the user itself. Therefore, we define p_j^u as the probability of tile j to be displayed by user's category u. Finally, we denote by δ^u the portion of users of type-u, with $\sum_{u \in \mathcal{U}} \delta^u = 1$.

5.2 Problem Formulation

Given the set $\mathcal{L}$ of all possible representations for all videos $v \in \mathcal{V}$, we seek the optimal subset of representations $\mathcal{T}^* \subseteq \mathcal{L}$, which maximises the perceived quality during the navigation and yet it minimises the total price of storage and encoding for the selected tile-set. Our *user-centric server*

optimisation problem can be defined as follows:

$$\begin{aligned} \mathcal{T}^* : \arg\min_{\mathcal{T}} & \sum_{u\in\mathcal{U}} D^u(\mathcal{T})p^u + \lambda C^{TOT}(\mathcal{T}) \\ \text{s.t.} & \sum_{(j,r,s)\in\mathcal{T}} b_r \le BW^u \qquad\qquad \forall u \end{aligned} \tag{2}$$

where $D^u(\mathcal{T})$ is the spherical distortion experienced by type-u user achieved when the $\mathcal{T}$ representation set is available at the main server and p^u is the probability to experience a type-u user. Finally, λ is the regularisation term and $C^{TOT}(\mathcal{T})$ is the total cost to store and code $\mathcal{T}$. In particular, the distortion $D^u(\mathcal{T})$ is defined as follows:

$$D^u(\mathcal{T}) = \sum_{(r,s)\in\mathcal{T}} D^u(r,s) = \sum_{(j,r,s)\in\mathcal{T}} D^u_j(r,s)\hat{S}_j p^u_j \tag{3}$$

where j is a generic tile on the planar encoded at type-u-th rate and s-th resolution. To take into account the spherical geometry, the spherical distortion is weighted by $\hat{S}_j$ that is the normalised portion of the sphere covered by tile j. Finally, p^u_j is the probability for the tile j to be displayed by a type-u user. Storing the video on the main server provider has a cost (\$), which depends on both the content complexity (affecting the total file size) and the resolution of representations. We estimate this total cost $C^{TOT}(\mathcal{T})$ in Eq. (2) as sum of the cost per each encoded ($\sum C_j(r,s)$). Since no prior assumption on distortion function (such as linear, quadratic, or convex function) is imposed, we preserve a general solving method and we cast the optimisation problem presented in Eq. (2) as ILP problem introducing the following binary decision variables:

$$\begin{aligned} \alpha^u_{j,r,s} &= \begin{cases} 1, \text{ if user } u \text{ requests the representation } (j,r,s) \\ 0, \text{ otherwise} \end{cases} \\ \beta_{j,r,s} &= \begin{cases} 1, \text{ if any user request a representation } (j,r,s) \\ 0, \text{ otherwise.} \end{cases} \end{aligned} \tag{4}$$

Without loss of generality, we suppose that each user can request only tile-representation encoded at resolution s corresponding to the display resolution (*i.e.*, spatial resolution at which the content will be displayed) of the selected device m^u. Therefore, we also define the following auxiliary variable:

$$\gamma^u_s = \begin{cases} 1, \text{ if user } u \text{ requests representations at resolution } s \\ 0, \text{ otherwise.} \end{cases} \tag{5}$$

This leads to the problem formulation shown in problem (6) equivalent to the problem showed in Eq. (2). The constraints (6a)-(6c) set up a consistent relation between the two decision variables. The constraints (6d)-(6f) makes homogeneous the resolution constraint by auxiliary variable γ. The constraint (6g) imposes bandwidth constraints. Finally, constraints (6h)-(6j) limit the decision variables to binary values.

The optimal solution of the ILP problem proposed in Eq. (6) is NP-hard and it can be evaluated by a generic solver IBM ILOG CPLEX [20] using a branch-and-cut algorithm. The method of branch-and-cut consists of a search tree technique, and the application of cuts at the nodes of the tree. In particular, each node represents a LP subproblem to be solved, and the creation of two new nodes from a parent node is a branch. It is worth mentioning that the branch-and-cut algorithm generally requires exponential computational complexity $O(2^E)$ to achieve the optimal solution, with E being the cardinality of decision variables. In our case with the binary decision variables α,β and γ, we obtain E$\sim |\mathcal{U}|^2|\mathcal{J}_v|^2|\mathcal{R}|^2|\mathcal{S}|^3$.

Integer Linear Programming 1

$$\min_{\alpha,\beta,\gamma} \sum_{u\in\mathcal{U}} \sum_{j\in\mathcal{J}_v} \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} D_j^u(r,s) p_j^u \alpha_{jrs}^u + \lambda \sum_{j\in\mathcal{J}_v} \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \beta_{jrs} D_j(r,s)$$

$$\text{s.t.} \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \alpha_{jrs}^u \leq 1 \qquad \forall u,j \qquad (6a)$$

$$\alpha_{jrs}^u \leq \beta_{jrs} \qquad \forall u,j,r,s \qquad (6b)$$

$$\beta_{jrs} \leq \sum_{u\in\mathcal{U}} \alpha_{jrs}^u \qquad \forall j,r,s \qquad (6c)$$

$$\sum_{s\in\mathcal{S}} \gamma_s^u \leq 1 \qquad \forall u \qquad (6d)$$

$$\alpha_{jrs}^u \leq \gamma_s^u \qquad \forall u,j,r,s \qquad (6e)$$

$$\gamma_s^u \leq \sum_{j\in\mathcal{J}_v} \sum_{r\in\mathcal{R}} \alpha_{jrs}^{dn} \qquad \forall u,s \qquad (6f)$$

$$\sum_{j\in\mathcal{J}_v} \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \alpha_{jrs}^u b_r \leq BW^u \qquad \forall u \qquad (6g)$$

$$\alpha_{jrs}^u \in \{0,1\} \qquad \forall u,j,r,s \qquad (6h)$$

$$\beta_{jrs} \in \{0,1\} \qquad \forall j,r,s \qquad (6i)$$

$$\gamma_s^u \in \{0,1\} \qquad \forall u,s \qquad (6j)$$

6 METRICS AND USER POPULATION

In the following, we describe the objective functions used in this work to validate the optimisation problem proposed in Section 5.2. First, we present the distortion function and cost model that we consider to minimise storage capacity utilisation, ensuring a high quality of experience. Then, we define the different types of user population that reflect a wide set of clients in our simulated ODV adaptive streaming scenario.

6.1 Distortion and cost models

We now evaluate the distortion $D^u(\mathcal{T})$ as the weighted MSE (WMSE) [49] that include coding and spherical geometry (*i.e.*, spherical shape) distortions in the traditional MSE metric. Because of its pixel-based distortion estimation and low computational complexity, we adopt the WMSE metric [49] as a distortion measure. A recent subjective study for ODV has shown a good correlation between the WMSE metric values and subjective scores [9]. Given a frame with resolution $W \times H$, the WMSE is defined as following:

$$WMSE(k,l) = \sum_{k=0}^{W-1} \sum_{l=0}^{H-1} (x(k,l) - y(k,l))^2 w(k,l) \qquad (7)$$

where $x(k,l)$ and $y(k,l)$ are intensity values at the pixel position (k,l) for the reference and projected image, respectively. Instead, $w(k,l)$ represents the non-linear weights that takes into account the spherical geometry to MSE. Namely, this constant reflects the stretching ratio for pixel in position (k,l) and depends on the planar-to-spherical projection. In this framework we consider ERP, hence each pixel weight is defined as follows:

$$w(k,l) = \frac{W(k,l)}{\sum_{k=0}^{W-1} \sum_{l=0}^{H-1} W(k,l)} \qquad (8)$$

where $W(k,l)$ is the area scaling factor from equirectangular to unit spherical surface and is given by $W(k,l) = cos\left[\left(l - \frac{H}{2} + \frac{1}{2}\right)\frac{\pi}{H}\right]$.

Beyond the distortion, another important aspect that the system designer should aim to minimise is the storage and encoding costs. Storing video representations at server providers, such as Amazon, Microsoft, etc., has a price that depends on the total size of the representations (in terms of *kbps*), and while storage cost might seems negligible, it is not when scaled for the number of video contents and representations that a content provider should have. Hence there is a need for the proposed optimisation algorithm. Formally, the storage and coding costs is a function of the video complexity, resolution and encoding rate and it is defined as (cost per tile-representation) [37] $C^{TOT}(\mathcal{T}) = C^e(\mathcal{T}) + C^s(\mathcal{T})$ where C^e and C^s are the encoding and storage costs, respectively. In particular, C^e is defined per each representation set ($\mathcal{T}$) as:

$$C^e(\mathcal{T}) = \begin{cases} \mu_e, & \text{if } s \leq 720p \\ 2\mu_e, & \text{if } 720p < s \leq 1080p \\ 4\mu_e, & \text{if } 1080p < s \leq 4\text{K} \end{cases} \tag{9}$$

where μ_e (\$) is a constant defined by the service provider and s is the resolution of each representation in $\mathcal{T}$. Instead, C^s is modelled as a linear function of the representation bitrate:

$$C^s(\mathcal{T}) = \mu_s \sum_{(j,r,s)\in\mathcal{T}} b_r \tag{10}$$

where μ_s (\$/GB) is a constant defined by service provider and b_r is the bitrate of the selected representation set ($\mathcal{T}$).
In our simulation settings, we follow the price-table of a real service provider [2, 3]. Therefore, we set $\mu_e = 0.1904$ \$/minute as the price to convert a video with an optimised quality in High Efficient Video Coding (HEVC) with frame rate ≤ 30 *fps* and $\mu_s = 0.024$ \$/GB. Both costs refer to the area of Europe (London) in [2, 3].

6.2 Users population features

In a practical adaptive ODV systems, content providers serve a vast number of highly heterogeneous users. For the optimisation purposes, we categorise them based on key features. As defined in Section 5.2, a user $u \in \mathcal{U}$ is characterised by three parameters: requested video, viewing device and network type. Each of these parameters is modelled as follows.

- *Requested video content, v^u.* We consider the dataset of 15 ODVs presented in Section 3.1 composed by 3 different categories (*Documentary*, *Action* and *Movie*) with 5 video per category. We suppose that users can select each available video with the same probability (1 out of 15).
- *Selected rendering device, m^u.* Each user can display the video content on 3 viewing platforms (*i.e.*, HMD, tablet and laptop). Without loss of generality, we assume that users select device with equal probability (1 out of 3).
- *Type of network and related available bandwidth, n^u and BW^u.* We consider 3 types of networks (*i.e.*, , 4G, WiFi, and ADSL) with their specific range of throughput and probability of experiencing that connection. For each type of connection, 3 different kinds of users have been considered, which means 3 values of bandwidth BW^u is possible per connection. Further details have been provided in Appendix B

In summary, we consider 27 types of users per video (3 types of devices $\times$ 3 types of possible networks $\times$ 3 possible bandwidth values). This ensures that our proposed *user-centric server optimisation*

algorithm is tested under realistic settings with a complete and exhaustive set of clients, while preserving a limited complexity of the ILP problem.

7 SIMULATION SETTINGS

In this section, we provide the remaining details of the framework that we used to validate the proposed *user-centric server optimisation*.

7.1 Tiling and encoding

Each ODV was partitioned into six self-decodable tiles to deliver and render ODVs efficiently. Following the usual assumption of lower importance and low-motion characteristics of the poles and the dominant viewing adjacency of the equator [11, 56]. Further details are provided in Appendix B.

We used the HEVC standard [35] to encode each tile of a given omnidirectional video. For this purpose, the *libx265* codec in the FFmpeg software (*ver.* N-85291) [21] was used. As recommended in [4], each tile was encoded using two-pass with 150 percent constrained variable bitrate configurations to ensure smooth video quality frame by frame for a wide range of devices. Before encoding, we scaled each video at different resolutions, $\mathcal{S} = \{1280 \times 720, 1920 \times 1080, 2560 \times 1440\}$. For the former one, as the content is already in the 2560×1440 resolution, no scaling was applied, and the two other resolutions were obtained by down sampling using the bi-cubic scaling technique. Here, we ensured that there is a noticeable objective quality difference between each selection per ODV. Each scaled version of ODV was tiled and encoded using a set of target bitrate parameters $\mathcal{B} = \{500, 760, 1005, 1529, 2326, 3537\}$ (in terms of Kbps). Each bit-stream was then divided into 2 *sec.* streaming chunks to perform adaptive streaming.

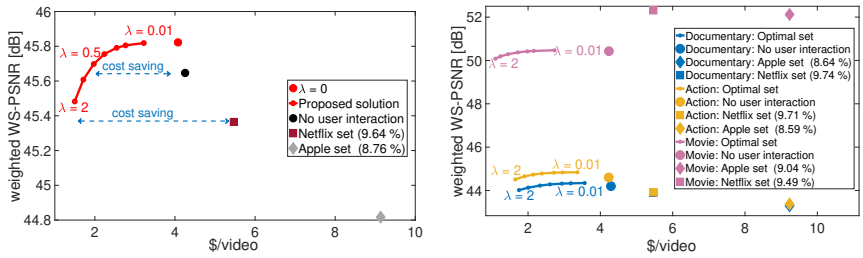
7.2 Comparative Methods

As last step of the simulation settings, we describe the benchmarking solutions for the optimisation server design. In particular, we evaluate the optimal sets of tile-representation with our user-centric algorithm (named "Optimal set" in the following plots) imposing different values of the regularisation parameter λ. In particular, we set $\lambda = [0.01, 0.05, 0.1, 0.25, 0.5, 1, 2]$. Then, we compare the performance of our optimisation with two sub-optimal solutions (*i.e.*, "$\lambda = 0$" and "optimal set - no interactivity") and two traditional recommendations sets (*i.e.*, "Netflix set" and "Apple set") [4, 32], which were originally developed for traditional 2D videos. "$\lambda = 0$" indicates the solution of our problem but neglecting the optimisation of costs, while "optimal set - no interactivity" omits also the probability p_j^u that defines where users most likely will focus their attention. The recommended bitrate sets of Apple and Netflix are defined as following: *i)* $\mathcal{B} = \{400, 480, 560, 640, 750, 900, 970, 1170, 1350, 1670\}$ Kbps for the Apple set with corresponding encoded resolutions $\mathcal{S} = \{720p, 720p, 720p, 720p, 1440p, 1440p, 1440p, 1440p, 1080p, 1080p\}$ and *ii)* for the Netflix set $\mathcal{B} = \{390, 500, 720, 980, 1300, 1920\}$ Kbps and encoding resolution $\mathcal{S} = \{720p, 720p, 1440p, 1440p, 1080p, 1080p\}$.

8 SIMULATION RESULTS

The key goals of the proposed optimisation problem are *i*) to ensure a good navigation experience within an ODV, reducing the total cost of encoding and storage; *ii*) to show the advantage of taking into account users' behaviour in this optimisation.

Fig. 6 depict the averaged quality experienced by users (in terms of Weighted Spherical PSNR (WS-PSNR) [49]) as a function of the total cost, for the proposed optimal set representations as well as the benchmark ones introduced in Section 7.2. The experienced quality has been evaluated as the average quality of each tile weighted by its probability of being displayed in a specific scenario



(a) Performance averaged per entire database (b) Performance averaged per video category

Fig. 6. Average experienced quality versus total cost of storage. In the legend on bracket, utilisation rate for non-optimal solutions.

(*i.e.*, selected video and viewing device). We consider the performance averaged across all videos of the database in Fig. 6 (a) and across content category in Fig. 6(b). As a result, the optimal set evaluated by the proposed optimisation achieves a lower distortion with respect to benchmark solutions (especially compared to the Apple set). Most importantly, the optimal set achieves a substantial saving in terms of cost. While Netflix and Apple sets spend respectively around 5.4$ and 9$ to store a short ODV of 20 seconds in length, we ensure the same performance in terms of WS-PSNR while saving 50%-70% of their cost per sequence. This translates to a gain of 50-100 to store the entire dataset of 15 videos (*i.e.*, 300 *sec.* of video content), which represents a significant saving in terms of cost even for the relatively small database presented in this work. If we imagine applying this optimisation to a bigger dataset and/or longer sequences, the financial saving could be very significant. The experienced quality is also strongly related to the video content as evident in Fig. 6(b). For instance, the *Movie* category is characterised by a reduced video complexity (see Fig. 1) and achieves higher performance with respect to the other video categories. More in general, for all video categories, the optimal set and the vendor recommendations achieve a comparable quality of experience, but with a much higher cost for the vendor ones.

Finally, it is worth noting that when the representation set is optimised without taking into account user navigation, see black dot in Fig. 6 (a), it performs almost as well as the optimal set with $\lambda = 0.5$ in terms of quality but it costs more than the double ($4.2 and $2, respectively). Overall, the optimised set of representations to store at the main server outperforms the recommended sets in terms of quality and, especially, total costs.

We are now interested in formalising the link between the data analysis provided in Section 4 and the user-centric server optimisation. For more in-depth study of the relationship between users' behaviour and the final quality, Fig. 7 depicts the total cost (per video) of the optimal tile-representation set optimised with $\lambda = 0.5$ as a function of the mean value UAI previously defined in Eq. (1). With the exception of IDs 09 and 10, the total cost increases accordingly with the value of UAI, especially when observing at the cost increase per video category. This shows that the way in which users interact with the content influences the performance of the optimal set of tile representations stored in an adaptive ODV streaming system. We investigate this intuition in greater depth, providing an exhaustive analysis of the effect of users' behaviour on the optimal set. In particular, we select three ODVs (namely, IDs 03, 08 and 13), each one coming from one category. These videos are selected as heterogeneous samples −in terms of UAI and cost value ($) − in Fig. 7. In particular, ID 03 has the lowest UAI value (0.25, mean value from Fig. 3), ID 08 has a medium value (0.58 from Fig. 3) and ID 13 has the highest value among the three (0.80 from Fig. 3). The quality distribution over time and space of the optimal set evaluated with $\lambda = 0.5$ is now further analysed. As previously highlighted in Fig. 2 (b), users tend to display the central area (*i.e.*, around

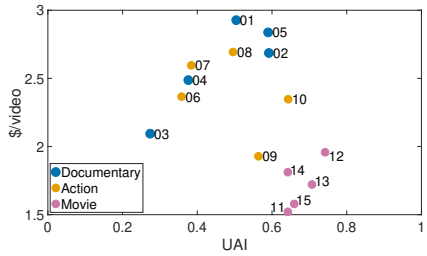
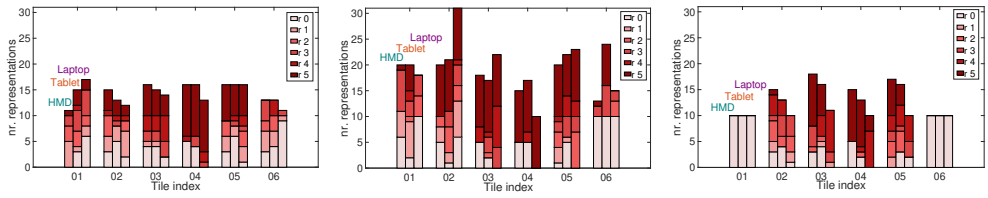


Fig. 7. Total cost of storage and coding of optimal tile-representation set ($\lambda = 0.5$) per each video and User Affinity Index (UAI) averaged per devices.

π value of latitude) of the equatorial zone in all ODVs of our database. This preference is reflected in the optimal tile set. Specifically, Fig. 8 provides the stored coding rate level averaged over time per each tile and viewing device (*i.e.*, variable r in Table 2) computed by the proposed user-centric optimisation algorithm. At first look, it can be noticed that tiles corresponding to the two poles (*i.e.*, tile indexes 1 and 6) are mainly stored with the lowest value of quality. This is extremely evident for *Movie* sequences in Fig. 8 (c). In contrast, the central area, such as tiles of index 3 and 4, have the majority of stored representations at the highest quality (*i.e.*, $r = 06$, where r is the selected coding rate as defined in Table 2). It is also worth mentioning that tile 4 (*i.e.*, one of the two frontal tiles as showed in Fig. 11) is mainly selected either with the highest or lowest quality in all three examples. This could be related with the user probability of displaying that area. The algorithm allocates the highest quality to this tile since it is the most commonly selected one during the navigation, but to ensure the streaming service in all conditions, it also picks the lowest quality, which has the lowest cost. In the following, we further investigate this behaviour by observing the quality levels stored over time. For the sake of brevity and motivated by the previous observations in Fig. 8, we only consider an HMD, and we restrict the analysed area to the equatorial zone (*i.e.*, tile index 2, 3, 4 and 5). In Fig. 9 (a,b,c), the UAI over time is compared with the total stored bitrate of the optimal tile set for Video IDs 03, 08 and 13, respectively. Interestingly, a strong correlation between these two metrics can be observed. For example, Video ID 13 of the *Movie* category has a high UAI, and the total stored bitrate is almost constant over time. In the other two examples, the amount of stored data is more sensitive to users' behaviour. A similar correlation can be noticed when comparing the UAI over time with the stored quality distribution in Fig. 9 (d,e,f). In Fig. 9 (d), we can note that diversity in terms of quality for the tile-representations is high when the affinity among users is low overall. In contrast, the Video ID 08 in Fig. 9 (e) has a medium level of affinity but the variance of the stored quality levels is lower. Interestingly comparing Fig. 9 (b) and (e), we can note that the UAI has a peak around 12–14 *sec.* leading to a drop in the stored bitrate (Fig. 9 (b)). The behaviour may seem contradictory, but it is worth mentioning that a high affinity value means a reduced uncertainty in the system. Therefore, the resources can be better allocated based on users' preferences. Indeed, observing Fig. 9 (e), the quality distributions of tile-representations is almost constant. Therefore, the lower value of stored bitrate around second 14 is due to a further reduction of stored representations in the polar area. As it is unlikely they will be selected by users, their quality level drops. The corresponding plots of Fig. 9 evaluated for the other devices (*i.e.*, tablet and laptop) are provided in the Appendix C and similar conclusions can be extracted from them.

In summary, from this user-centric server optimisation, we can deduce the following:

- **Observation 7**: A significant saving in terms of bitrate and encoding/storage cost is achieved when the stored representations are optimised based on both content and users' profiles.



(a) Documentary (ID 3) (b) Action (ID 8) (c) Movie (ID 13)

Fig. 8. Total number of stored tile-representations for all rendering devices per a single video of each category. In particular, the three column represents the optimal set for each tile for all device corresponding to HMD, tablet and laptop in order from left to right.

- **Observation 8**: The users' behaviour during the navigation generally affects the resource allocation of the optimal set (e.g., number of representations and quality levels).
- **Observation 9**: UAI provides a good representation of the existing correlation between users' behaviour and optimal set, floating the idea that UAI could be a key metric in the design of the next generation systems.

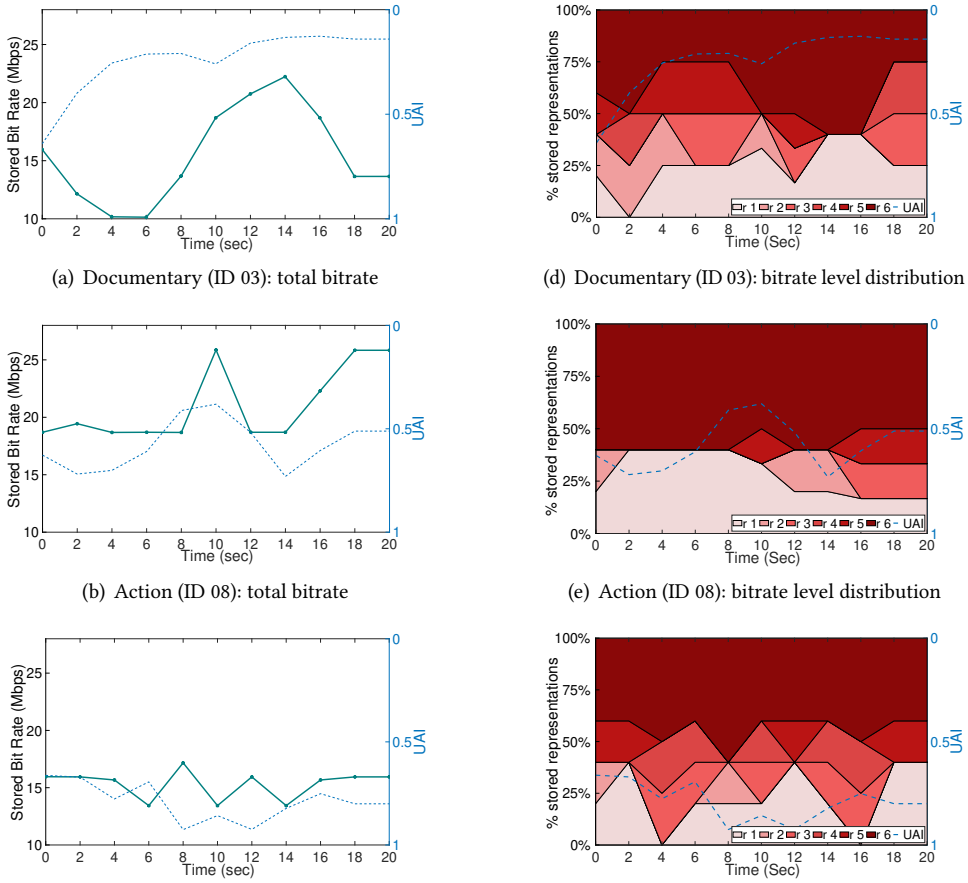
While Observation 7 has been already demonstrated in previous works examining conventional video [51] and ODVs [37], the other outcomes are novel insights that prove the importance of considering users' behaviour in the design of a VR streaming system.

9 CONCLUSION

The overall goal of this work is to explore the way in which people navigate with omnidirectional video (ODV), and its impact on the performance of VR adaptive streaming systems. To reach this goal, we conduced a subjective test across two different European universities (*i.e.*, UCL and TCD) collecting navigation trajectories of 94 participants using three different VR devices (HMD, laptop and tablet). This allowed us to build a dataset with navigation trajectories across different devices that we make publicly available. The collected data have been exhaustively analysed, showing key differences of users' behaviour across device and content category. For instance, users watching contents from the *Movie* category or displaying ODV with HMD will experience a more similar interaction between each other with respect to the case of other devices or other contents. As case study, we apply these findings to the open problem of optimising the storage at the server provider for ODV adaptive streaming systems. A novel *user-centric immersive algorithm* has been proposed to optimise the set of VR representations to be stored at the server, minimizing the total cost and yet maximising the final quality. The key-novelty of our algorithm is to take into consideration users' behaviour beyond the spherical geometry and content information. As result, our optimal representation set ensures the same quality experienced with vendor recommendations but saving up to 70% of coding and storage cost. Moreover, we have shown how the different types of user navigation (e.g., affinity) impact on the optimal set. This opens the gate to a possibility of user-centric studies focused on making the users' behaviour (and user affinity) the driver of VR system designs.

ACKNOWLEDGMENTS

This work has been partially funded by Adobe under Academic Donation scheme. Also, this publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.



(a) Documentary (ID 03): total bitrate

(d) Documentary (ID 03): bitrate level distribution

(b) Action (ID 08): total bitrate

(e) Action (ID 08): bitrate level distribution

(c) Movie (ID 13): total bitrate

(f) Movie (ID 13): bitrate level distribution

Fig. 9. Temporal analysis of optimal tile-representation set for navigation trajectories with HMD across video categories. In the left column, the total stored bitrate over time for each video is presented while in the right column there is the bitrate level distribution of only equatorial area for each selected video. In each plot, the UAI over time is also reported.

REFERENCES

[1] Mathias Almquist, Viktor Almquist, Vengatanathan Krishnamoorthi, Niklas Carlsson, and Derek Eager. 2018. The Prefetch Aggressiveness Tradeoff in 360° Video Streaming. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*.

[2] Amazon. 2019. Cloud Storage pricing. Retrieved 2019 from https://aws.amazon.com/s3/pricing/

[3] Amazon. 2019. Elastic transcoder pricing. Retrieved 2019 from https://aws.amazon.com/elastictranscoder/pricing/

[4] Apple. 2018. HLS Authoring Specification for Apple Devices. Retrieved 2019 from https://developer.apple.com

[5] Samantha W. Bindman, Lisa M. Castaneda, Mike Scanlon, and Anna Cechony. 2018. Am I a Bunny?: The Impact of High and Low Immersion Platforms and Viewers' Perceptions of Role on Presence, Narrative Engagement, and Empathy During an Animated 360° Video. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '18)*.

[6] F. Chao, L. Zhang, W. Hamidouche, and O. Deforges. 2018. Salgan360: Visual Saliency Prediction On 360 Degree Images With Generative Adversarial Networks. In *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*.

[7] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360 Degreee Video Head Movement Dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*.

[8] Xavier Corbillon, Alisa Devlic, Gwendal Simon, and Jacob Chakareski. 2017. Optimal Set of 360-Degree Videos for Viewport-Adaptive Streaming. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*.

[9] Simone Croci, Cagri Ozcinar, Emin Zerman, Julián Cabrera, and Aljosa Smolic. 2019. Voronoi-based Objective Quality Metrics for Omnidirectional Video. In *IEEE 11th International Conference on Quality of Multimedia Experience (QoMEX)*.

[10] Erwan J. David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A Dataset of Head and Eye Movements for 360° Videos. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*.

[11] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. 2017. Look around you: Saliency maps for omnidirectional images in VR applications. In *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX)*.

[12] Francesca De Simone, Jesús Gutiérrez, and Patrick Le Callet. 2019. Complexity measurement and characterization of 360-degree content. *Electronic Imaging* (2019).

[13] Fanyi Duanmu, Yixiang Mao, Shuai Liu, Sumanth Srinivasan, and Yao Wang. 2018. A subjective study of viewer navigation behaviors when watching 360-degree videos on computers. In *IEEE International Conference on Multimedia and Expo (ICME)*.

[14] Ching-Ling Fan, Wen-Chih Lo, Yu-Tung Pai, and Cheng-Hsin Hsu. 2019. A Survey on 360° Video Streaming: Acquisition, Transmission, and Display. *ACM Comput. Surv.* (2019).

[15] Colm O Fearghail, Cagri Ozcinar, Sebastian Knorr, and Aljosa Smolic. 2018. Director's Cut - Analysis of Aspects of Interactive Storytelling for VR Films. In *International Conference for Interactive Digital Storytelling (ICIDS)*.

[16] Colm O Fearghail, Cagri Ozcinar, Sebastian Knorr, and Aljosa Smolic. 2018. Director's Cut-Analysis of VR Film Cuts for Interactive Storytelling. In *IEEE International Conference on 3D Immersion (IC3D)*.

[17] Stephan Fremerey, Ashutosh Singla, Kay Meseberg, and Alexander Raake. 2018. AVtrack360: An Open Dataset and Software Recording People's Head Rotations Watching 360° Videos on an HMD. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*.

[18] M. Graf, C. Timmerer, and C. Mueller. 2017. Towards Bandwidth Efficient Adaptive Streaming of Omnidirectional Video over HTTP: Design, Implementation, and Evaluation. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*.

[19] Jonathan Harth, Alexandra Hofmann, Mike Karst, David Kempf, Annelie Ostertag, Isabell Przemus, and Bernhard Schaefermeyer. 2018. Different Types of Users, Different Types of Immersion: A User Study of Interaction Design and Immersion in Consumer Virtual Reality. *IEEE Consumer Electronics Magazine* (2018).

[20] IBM. 2013. ILOG CPLEX Optimization studio. https://www-01.ibm.com/software/

[21] MulticoreWare Inc. 2018. x265 HEVC Encoder / H.265 Video Codec. http://x265.org/

[22] ITU-T. 2008. Subjective Video Quality Assessment Methods for Multimedia Applications. ITU-T Recom. P.910.

[23] Chakareski Jacob, Aksu Ridvan, Corbillon Xavier, Simon Gwendal, and Swaminathan Viswanathan. 2018. Viewport-Driven Rate-Distortion Optimized 360º Video Streaming. In *IEEE International Conference on Communications (ICC)*.

[24] Sebastian Knorr, Cagri Ozcinar, Colm O Fearghail, and Aljosa Smolic. 2018. Director's Cut - A Combined Dataset for Visual Attention Analysis in Cinematic VR Content. In *The 15th ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*.

[25] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. 2018. Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video: A Large-Scale Dataset and a Deep Learning Model. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*.

[26] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. 360° Video Viewing Dataset in Head-Mounted Virtual Reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*.

[27] Lester C. Loschky, Adam M. Larson, Joseph P. Magliano, and Tim J. Smith. 2015. What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PloS one* (2015).

[28] Anahita Mahzari, Afshin Taghavi Nasrabadi, Aliehsan Samiei, and Ravi Prakash. 2018. FoV-Aware Edge Caching for Adaptive 360° Video Streaming. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*.

[29] Pantelis Maniotis, Eirina Bourtsoulatze, and Nikolaos Thomos. 2019. Tile-Based Joint Caching and Delivery of 360° Videos in Heterogeneous Networks. *IEEE Transactions on Multimedia* (2019).

[30] Kiran Misra, Andrew Segall, Michael Horowitz, Shilin Xu, Arild Fuldseth, and Minhua Zhou. 2013. An Overview of Tiles in HEVC. *IEEE Journal of Selected Topics in Signal Processing* (2013).

[31] Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P. McMahan, Ravi Prakash, Mylène C. Q. Farias, and Marcelo M. Carvalho. 2019. A Taxonomy and Dataset for 360° Videos. In *Proceedings of the 10th ACM Multimedia Systems Conference (MMSys '19)*.

[32] Netflix. 2015. Per-Title Encode Optimization. Retrieved 2019 from https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2

[33] Duc V. Nguyen, Huyen T. T. Tran, Anh T. Pham, and Truong C. Thang. 2019. An Optimal Tile-Based Approach for Viewport-Adaptive 360-Degree Video Streaming. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*

(2019).

[34] Omar A. Niamut, Emmanuel Thomas, Lucia D'Acunto, Cyril Concolato, Franck Denoual, and Seong Yong Lim. 2016. MPEG DASH SRD: Spatial Relationship Description. In *Proceedings of the 7th International Conference on Multimedia Systems (MMSys '16)*.

[35] J.-R. Ohm and G. Sullivan. 2011. *Vision, applications and requirements for high efficiency video coding (HEVC).* Technical Report. ISO/IEC JTC1/SC29/WG11.

[36] Cagri Ozcinar, Julián Cabrera, and Aljosa Smolic. 2019. Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (2019).

[37] Cagri Ozcinar, Ana De Abreu, Sebastian Knorr, and Aljosa Smolic. 2017. Estimation of optimal encoding ladders for tiled 360 VR video in adaptive streaming systems. In *IEEE International Symposium on Multimedia (ISM '17)*.

[38] Cagri Ozcinar, Ana De Abreu, and Aljosa Smolic. 2017. Viewport-aware adaptive 360 video streaming using tiles for virtual reality. In *2017 IEEE International Conference on Image Processing (ICIP '17)*.

[39] Cagri Ozcinar and Aljosa Smolic. 2018. Visual Attention in Omnidirectional Video for Virtual Reality Applications. In *IEEE 10th International Conference on Quality of Multimedia Experience (QoMEX '18)*.

[40] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards Foveated Rendering for Gaze-tracked Virtual Reality. *ACM Trans. Graph.* (2016).

[41] S. Petrangeli, G. Simon, and V. Swaminathan. 2018. Trajectory-Based Viewport Prediction for 360-Degree Virtual Reality Videos. In *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR '18)*.

[42] Michelle M Ramey, Andrew P Yonelinas, and John M Henderson. 2019. Conscious and unconscious memory deferentially impact attention: Eye movements, visual search, and recognition processes. *Cognition* (2019).

[43] Silvia Rossi, Francesca De Simone, Pascal Frossard, and Laura Toni. 2019. Spherical Clustering of Users Navigating 360° Content. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '19)*.

[44] Silvia Rossi and Laura Toni. 2017. Navigation-aware adaptive streaming strategies for omnidirectional video. In *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP '17)*.

[45] Jose Rubio-Tamayo, Manuel Gertrudix Barrio, and Francisco García García. 2017. Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Technologies and Interaction* (2017).

[46] Salient360! 2019. Salient360! - Visual Attention Modeling For 360 Content. Retrieved 2019 from https://salient360.ls2n.fr/grand-challenges/

[47] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics* (2018).

[48] Mel Slater and Maria V Sanchez-Vives. 2016. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI* (2016).

[49] Yule Sun, Ang Lu, and Lu Yu. 2017. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Processing Letters* (2017).

[50] C. Timmerer. 2017. Immersive Media Delivery: Overview of Ongoing Standardization Activities. *IEEE Communications Standards Magazine* (2017).

[51] Laura Toni, Ramon Aparicio-Pardo, Karine Pires, Gwendal Simon, Alberto Blanc, and Pascal Frossard. 2015. Optimal Selection of Adaptive Streaming Representations. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2015).

[52] Audrey Tse, Charlene Jennett, Joanne Moore, Zillah Watson, Jacob Rigby, and Anna L Cox. 2017. Was I there?: impact of platform and headphones on 360 video immersion. In *Proceedings of the ACM conference extended abstracts on human factors in computing systems (CHI EA '17)*.

[53] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. 2017. A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*.

[54] Mengbai Xiao, Chao Zhou, Yao Liu, and Songqing Chen. 2017. OpTile: Toward Optimal Tiling in 360-degree Video Streaming. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*.

[55] M. Xu, C. Li, S. Zhang, and P. Le Callet. 2020. State-of-the-art in 360° Video/Image Processing: Perception, Assessment and Compression. *IEEE Journal of Selected Topics in Signal Processing* (2020).

[56] Matt Yu, Haricharan Lakshman, and Bernd Girod. 2015. A Framework to Evaluate Omnidirectional Video Coding Schemes. In *IEEE International Symposium on Mixed and Augmented Reality*.

[57] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. 2018. Saliency detection in 360 videos. In *Proceedings of the European Conference on Computer Vision (ECCV '18)*.

[58] Junni Zou, Chenglin Li, Chengming Liu, Qin Yang, Hongkai Xiong, and Eckehard Steinbach. 2019. Probabilistic Tile Visibility-Based Server-Side Rate Adaptation for Adaptive 360-Degree Video Streaming. *IEEE Journal of Selected Topics in Signal Processing* (2019).

A COMPARISON USER SIMILARITY INDEX WITH ENTROPY OF SALIENCY MAP

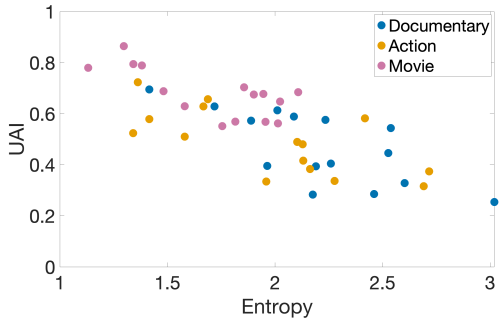


Fig. 10. Comparison of UAI with entropy of saliency maps per each video of the entire dataset.

To check the validity of our proposed metric, we evaluate also the entropy of the saliency map per each video of the entire dataset. This metric is typically use to evaluate model of visual attention, and gives a qualitative idea about the dispersion of users' movements over time. In particular, low value of entropy stands for users focused all on a restricted area (*i.e.*, focused content - high correlation among users); while high value means more exploratory movements (*i.e.*, exploratory content - low correlation among users). Moreover, authors in [12] have applied this metric to omnidirectional images providing its validity also for this kind of content. Fig. 10 shows the correlation between UAI and the entropy of saliency map per each video of the dataset averaged per devices. We can therefore notice a strong correlation between this traditional metric and our UAI. As expected, video characterized by low entropy have also high values of UAI meaning that users move similarly within the content; on the contrary, videos where users navigate more randomly, present high value of entropy.

B FURTHER SIMULATION SETTINGS

B.1 Type of network and available bandwidth

We now complete the information about the user population provided in Section 6.2. Specifically, we clarify the types of networks and available bandwidth.

We consider 3 types of networks with their specific range of throughput, provided in Table 3 in Appendix . We assume that the probability of experiencing a given connectivity is linked to the device, as reported in Table 4. For each connection type, 3 different kinds of users have been considered: *i)* clients with bandwidth BW^u set as the 25-th percentile of the available bandwidth for the selected network, *ii)* users with bandwidth BW^u set as the 75-th percentile of the available bandwidth for the selected network, and *iii)* clients with bandwidth BW^u set to the 50-th percentile of the available bandwidth for the selected network. We assume a probability 1/4 for a user to experience the first two cases and 1/2 to select the third downloading

Table 3. Networks Bandwidth ranges.

Network Type	Minimum Bandwidth (Mbps)	Maximum Bandwidth (Mbps)
4G	4	20
WiFi	2	30
ADSL	5	35

Table 4. Probability associate with each network and device in our simulations.

Network Type	HMD	Tablet	Laptop
4G	0	0.6	0
WiFi	0.8	0.4	0.45
ADSL	0.2	0	0.55

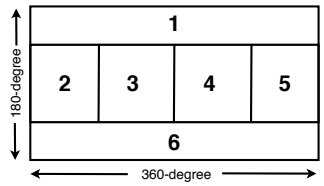


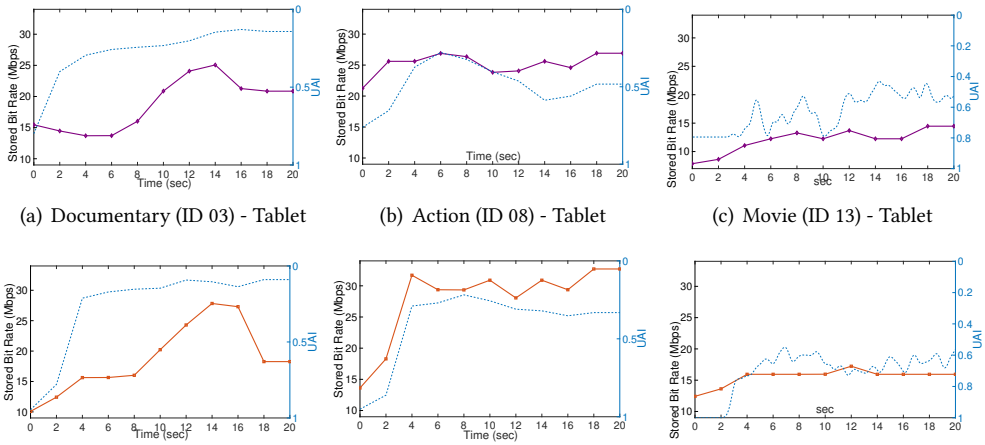
Fig. 11. The used structure for tiling with tile IDs.

B.2 Tiling and Encoding

Each ODV was partitioned into six self-decodable tiles to deliver and render ODVs efficiently. Following the usual assumption of lower importance and low-motion characteristics of the poles and the dominant viewing adjacency of the equator [11, 56], we separate each ODV frame horizontally into three parts: one equator and two poles. The equator represents the middle segment, and the two poles stand for the top and the bottom sections of the frame. The size of equator is the double size of each pole. As the poles occupy the largest regions of the redundant pixels, in those areas, larger tile resolution size was used to compress them efficiently [56]. On the contrary, since the equator region contains the most dominant viewing probability, it is further divided vertically into 4 tiles to efficiently utilised them at both the server and client sides. Fig. 11 illustrates the used structure for partitioning into self-decodable tiles and the tile index order that will be considered in the following.

C ADDITIONAL RESULTS

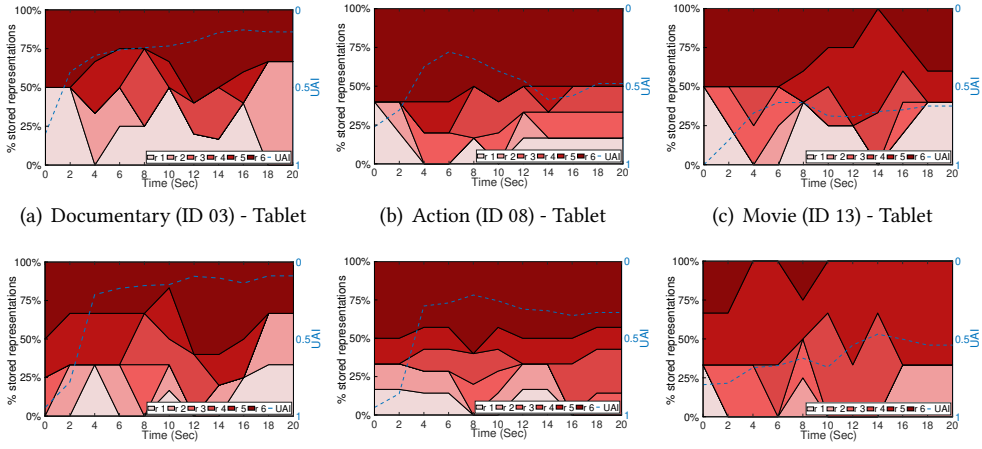
In this section we provide the temporal analysis of optimal tile-representation set for navigation trajectories with tablet and laptop across video categories.



(a) Documentary (ID 03) - Tablet (b) Action (ID 08) - Tablet (c) Movie (ID 13) - Tablet

(d) Documentary (ID 03) - Laptop (e) Action (ID 08) - Laptop (f) Movie (ID 13) - Laptop

Fig. 12. Total stored bitrate over time for each video: on the top line all plots are referred to tablet while on the bottom one to laptop. In each plots, UAI over time is also reported.



(a) Documentary (ID 03) - Tablet (b) Action (ID 08) - Tablet (c) Movie (ID 13) - Tablet

(d) Documentary (ID 03) - Laptop (e) Action (ID 08) - Laptop (f) Movie (ID 13) - Laptop

Fig. 13. Bitrate level distribution of the only equatorial area for each video: on the top line all plots are referred to tablet while on the bottom one to laptop. In each plots, UAI over time is also reported.