Quality of Experience for 3-D Immersive Media Streaming

Alexandros Doumanoglou, David Griffin, Javier Serrano, Nikolaos Zioulis, Truong Khoa Phan, David Jiménez, Dimitrios Zarpalas, Federico Alvarez, *Member, IEEE*, Miguel Rio, *Member, IEEE*, and Petros Daras, *Senior Member, IEEE*

***Abstract*—Recent advances in media capture and processing technologies have enabled new forms of true 3-D media content that increase the degree of user immersion. The demand for more engaging forms of entertainment means that content distributors and broadcasters need to fine-tune their delivery mechanisms over the Internet as well as develop new models for quantifying and predicting user experience of these new forms of content. In the work described in this paper, we undertake one of the first studies into the quality of experience (QoE) of real-time 3-D media content streamed to virtual reality (VR) headsets for entertainment purposes, in the context of game spectating. Our focus is on tele-immersive media that embed real users within virtual environments of interactive games. A key feature of engaging and realistic experiences in full 3-D media environments, is allowing users unrestricted viewpoints. However, this comes at the cost of increased network bandwidth and the need of limiting network effects in order to transmit a realistic, real-time representation of the participants. The visual quality of 3-D media is affected by geometry and texture parameters while the temporal aspects of smooth movement and synchronization are affected by lag introduced by network transmission effects. In this paper, we investigate varying network conditions for a set of tele-immersive media sessions produced in a range of visual quality levels. Further, we investigate user navigation issues that inhibit free viewpoint VR spectating of live 3-D media. After reporting on a study with multiple users we analyze the results and assess the overall QoE with respect to a range of visual quality and latency parameters. We propose a neural network QoE prediction model for 3-D media, constructed from a combination of visual and network parameters.**

***Index Terms*—Quality of experience (QoE), virtual reality (VR), immersive media, 3D content transmission, tele-immersion (TI), real-time 3D reconstruction, 3D streaming, free viewpoint video (FVV).**

Manuscript received December 27, 2017; revised March 26, 2018; accepted March 31, 2018. Date of current version June 5, 2018. This work was supported in part by EC Project 5G-MEDIA (www.5gmedia.eu), and in part by the European Union Horizon 2020 Research and Innovation Programme under Grant 761699. *(Corresponding author: Alexandros Doumanoglou.)*

A. Doumanoglou, N. Zioulis, D. Zarpalas, and P. Daras are with the Information Technologies Institute, Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece (e-mail: aldoum@iti.gr; nzioulis@iti.gr; zarpalas@iti.gr; daras@iti.gr).

D. Griffin, T. K. Phan, and M. Rio are with the Electronic and Electrical Engineering Department, University College London, London WC1E 7JE, U.K. (e-mail: d.griffin@ucl.ac.uk; t.phan@ucl.ac.uk; miguel.rio@ucl.ac.uk).

J. Serrano, D. Jiménez, and F. Alvarez are with GATV—ETSIT, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: jsr@gatv.ssr.upm.es; djb@gatv.ssr.upm.es; federico.alvarez@upm.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2018.2823909

I. Introduction

The advent of low-cost depth sensors, such as Microsoft Kinect v1 and Asus Xtion, in late 2010 made good quality 3D scanning technology widely available to the public. These low-cost depth sensors had low resolution depth-maps but they operated at high frame rates, reaching levels of up to 30 frames per second (fps). By relying on the depth maps provided by those sensors, accurate human skeleton tracking algorithms were developed [1], while their high frame rates allowed novel applications in human-computer-interaction interfaces. Moreover, in contrast to traditional 3D laser scanners which, at that time, were slow and costly, the high frame rates of these sensors allowed researchers to advance the state-of-the-art in real-time human 3D reconstruction with applications in 3D Tele-Immersion (3D-TI) [2], [3]. The full 3D reconstructed mesh of the human body produced by such algorithms enables free viewpoint content viewing which was not possible in standard 2D video or stereoscopic 3D video. Furthermore, the full 3D reconstructed human mesh is also possible to be virtually embedded inside a static or dynamic (i.e., time-varying) virtual environment. Depending on the 3D-TI application, the 3D reconstructed human mesh can interact with the elements of the virtual environment according to the physics rules imposed by the specific application. This embedding of the "real" human avatar (also referred as "3D-TI Content") inside the "virtual" fictional environment is often called *augmented virtuality* [4], [5].

Nowadays, while Microsoft has discontinued the production of Kinect, other manufacturers, such as ASUS and Intel, still continue to provide improved depth sensors to the market. Current depth sensors have higher frame rates, higher depth resolution, higher depth fidelity but at a lower cost. The amount of user generated 3D content is expected to vastly increase in the next few years, especially as the first smart-phones with integrated depth sensors are becoming more popular. On top of this, low-cost virtual reality (VR) headsets are becoming available on the market that bring exciting new ways for viewing and interacting with the increasingly available 3D content. (Here, by "3D content" we refer to any data that can be perceived as full 3D when visualized in a proper display technology being either a standard 2D display, a VR headset, or other). Thus, since technology has already offered easy ways to capture and consume live 3D content, we've reached a point where broadcasting such content is closer to mainstream consumption.

In order to enable free viewpoint remote consumption of live 3D content, it is necessary for the broadcaster to transmit the content in a 3D media format as opposed to standard 2D or stereoscopic video. For the content of a 3D-TI session (which is a special case of live 3D content), the 3D media format typically consists of a 3D geometry mesh of the captured human, plus rendering material information typically in the form of multiple 2D-textures. The main alternative to 3D media streaming would be for the broadcaster to stream pre-rendered views of the 3D content in the form of 2D-Video. The computational load to support free viewpoint viewing with pre-rendered views grows linearly for the broadcaster, as the number of subscribers increase. This is because a subjective view needs to be rendered for each spectator to enable true free view-point spectating. A variation of this approach, which is also a 3D media variant, transmits instead of the pre-rendered views, the captured color textures along with their corresponding depth maps, eventually offloading their fusion processing to the viewing clients. The drawback of the latter approach is that the clients of the streamed content require sufficient processing power to undertake the fusion task. Hence, it is preferential to transmit tele-immersion content in the form of 3D media (i.e., geometry mesh plus 2D textures). Moreover, as an additional advantage, the aforementioned 3D media format also makes the 3D-TI content easier to be incorporated in augmented virtuality applications.

In this paper, a quality of experience (QoE) study for spectating live 3D-TI augmented virtuality sessions in a full free viewpoint VR setting is presented. The cases covered concern applications where the live broadcaster is willing to interact with their subscribers in real-time. This kind of application imposes a low latency requirement in the transmitted stream in order to realize real-time interactions with the spectators, compared to typical unsynchronised content streaming. Moreover, this low latency requirement also prohibits the use of client-side buffering which deteriorates real-time interactions. The parametric space which is considered to influence the QoE of the participants in this study is divided in two groups. The first group of parameters influence the visual quality of the transmitted 3D media while the second group of parameters influence the perceived temporal consistency of the 3D media with the fictional virtual environment. Essentially, this means that the first group of parameters affects the quality of the 3D reconstructions while the second group of parameters corresponds to different network conditions and protocols that affect the perceived lag. While this study considers an application in next-generation immersive gaming (i.e., the augmented virtuality application is a next generation 3D-TI video game), the concepts and the ideas presented could also be applied in applications featuring tele-presence, tele-medicine, design collaboration, webinars and others.

The contribution of this paper is twofold. First, it is one of a very few quantitative QoE evaluations for TI systems in general. Moreover, other existing works on quantifying QoE in TI systems ([6], [7]) do not study 3D-TI with full-body 3D reconstructions of immersed participants. The work that we find mostly related to the present paper is that from [8]. However, in [8], the focus is mainly on the 3D-TI platform specifics and only a qualitative and not a quantitative study is undertaken for the QoE of the platform's users. The second contribution of this work is the fact that it studies 3D-TI QoE from the perspective of a spectator. To this aspect, the most relevant previous work is that in [9]. However, the platform studied in [9] only allows fixed viewpoint spectating in contrast to the complete unrestricted free viewpoint spectating that is offered by the platform studied in this paper. Moreover, to the best of our knowledge, the present work is the first to study QoE of 3D-TI spectators in a virtual reality setting using head mounted displays.

The rest of the paper is structured as follows: in Section II related work relevant to the subject of study is presented. In Section III-A a detailed explanation of the considered 3D media is given while in Section III-B we give a thorough presentation of the augmented virtuality tele-immersive video game which has become the subject of this study. In Sections IV and V the experimental setup and the results of the study are illustrated, while in Section VI an introductory QoE prediction model for 3D-TI immersive media streaming in presented. Finally, Section VII concludes the paper with a summarizing discussion.

II. Related Work

In this section of the paper, we enumerate related work found in the literature that is connected to our study. In Section II-A 3D-Media formats for TI is presented. Section II-B describes related QoE studies in VR applications and immersive experiences while in Section II-C relevant work to the network aspects of 3D-Media transmission are discussed. While QoE for 3D-Video streaming is also another related area to the present study, with exemplary works being [10] and [11], an extensive list of 3D-Video related work is excluded in this paper mainly because 3D-Video cannot exactly enable augmented virtuality applications such as the one presented in this paper.

A. 3D Media Formats for TI

In a typical TI application architecture, the 3D data corresponding to the appearance of the participants are captured in specialized TI capturing rooms (TI stations) equipped with multi-camera setups. In most cases, the 3D data acquired by the cameras are locally fused into a textured 3D mesh in the TI station by dedicated hardware (PCs). Subsequently, this 3D mesh is streamed to the subscribed viewers of the TI application for free viewpoint spectating. This type of TI 3D media (i.e., the textured 3D mesh) has been the mainstream approach for the previous works in [3] and [8]. The same concept has also been adopted in [12] but instead of using one TI capturing station per individual, the TI capturing site served a group of participants. In the work of [9], the second type of 3D media format is utilized, i.e., the color plus depth. As already discussed in the introduction, this 3D media format imposes certain limitations. However, the authors consider only the case where the viewers of the TI content have fixed locations inside the virtual space and thus leveraging this fact to only stream a single color plus depth pair over the network. The drawback in the latter case is the lack of offering free view-point spectating to the client viewers which, nevertheless, in that case, is a decision by design. On the other hand, this approach has the advantage of utilizing less network bandwidth than the textured 3D mesh and is able to serve more clients.

B. QoE in VR Applications and Immersive Experiences

Burdea and Coiffet [13] describe VR as a computer simulation where computer graphics are utilized in order to generate virtual worlds with whom the users of the application can interact in real-time. The characteristic that makes VR what it is, is the feeling of immersion that it transpires to the users in conjunction with real-time interactivity. Two display technologies for VR are the most common: CAVE Environments [14] and Head Mounted Displays (HMDs) [15]. Nowadays, HMDs are the most affordable means to experience VR. They are low-cost and they offer a high degree of immersion [16]. However, in certain situations, they may cause motion sickness [17]. A comparative study of different VR display technologies can be found in [18]. The low cost of HMDs is a critical factor that makes them the default choice for VR display, while their downsides are mitigated by careful application design.

With the latest technology advancements in VR HMDs and 360° video capturing, it has become quite common to stream 360° video content to VR headsets. However, this type of content does not offer a true VR experience as it neither offers a true 3D virtual world nor does it allow for real-time interactions with it. However, QoE studies in streaming 360° video to HMDs justify the immersiveness of the medium [19]. The same generic aforementioned QoE principles about immersiveness have also been studied in VR gaming [20]. The new medium (HMDs) was found to increase the engagement levels of the immersed users. Nonetheless, in the latter study it was once again witnessed that HMDs may cause effects of nausea after wearing the goggles. However, the results of this study lay more in the qualitative side of the use of HMDs on the topics of perceived presence, perceived usability and emotions, and less on the technical parameters.

Keighrey *et al.* [21] perform a QoE study of an interactive and immersive speech and language assessment application implemented both in VR and Augmented Reality (AR). Their findings demonstrate similar QoE ratings for both VR and AR, with users being acclimatized to AR more quickly than VR. While that work is relevant to the current one in terms of utilizing the same display medium (VR), the application studied does not cover any networking aspects as the current work does. Other relevant QoE studies, like [22], defined technical parameters affecting the QoE, especially for the visual and user comfort aspects, but with the focus on stereoscopic and not pure 3D content.

In [23], QoE prediction models are introduced that predict the user-perceived QoE of a TI conferencing application. While it certainly offers a valuable contribution, it does not apply to the same context of 3D-TI augmented virtuality such us the present work because in their case the TI content is not full watertight 3D meshes.

In 3D-TI, VR has found limited applicability compared to other areas like gaming, maybe due to the high complexity of deploying such an application. In [24], the idea of sharing the same virtual collaborative space by remote participants, which is the core concept of 3D-TI, is exploited in order to realize a VR environment for Taichi learning. The work of [24] shows that students of Taichi can present increased learning efficiency in a VR environment even only when their representation in the virtual world is constituted of avatars instead of real 3D reconstructions of the teacher and themselves. In [8], an initial qualitative study of the 3D-TI platform is conducted. However, this is mostly a preliminary work based on user's comments and not a quantitative QoE evaluation.

C. Network Transmission of 3D Media

Delay is an important factor in QoE of any interactive content. Even with over-provisioned networks and devices, delay is always lower-bounded by the propagation component [25]. Although current data center distribution allows for low service delay [26] this does not help services where several users, often randomly distributed around the world, need a consistent view of the virtual environment.

Transport Protocols: Transport protocols play a crucial role in delivering data reliably and at the right speed. Unfortunately neither Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) is appropriate for this kind of content. TCP's full reliability comes at a cost of delay. UDP does not provide reliability or a way of controlling congestion. New protocols like Quick UDP Internet Connections (QUIC [27]), allow for faster start-up times than TCP and alternative congestion control algorithms such as TCP Friendly Rate Control (TFRC [28]) can be deployed to deliver more stable transmission rates compared to TCP, however their adoption is still limited.

Application Layer Rate Limiting: In order to deal with the limitations of the transport layer, developers have been increasingly adopting application level strategies like MPEG DASH [29]. However, these are more appropriate for "pre-recorded" content that is not transmitted in real-time.

In this work we examine the impact of the network on the QoE experienced by users of 3D media considering, in particular, the trade-off between latency introduced by a reliable transport protocol versus frame loss rate. Higher quality 3D media streams require a greater quantity of data to be transmitted which also increases transmission time and therefore latency, especially when using a reliable transport protocol. The results of our study into how 3D media quality levels should be traded-off with network parameters will identify the requirements for the development of new (or the adaptation of existing) transport protocols and associated application-level dynamic quality adaptation mechanisms for 3D media.

III. Streaming Live 3D Tele-Immersion Sessions

While there are a couple of ways to immerse real users during virtual experiences as detailed in Section II, in this work we focus on the most demanding case in terms of bandwidth, but also the most satisfying in terms of the resulting

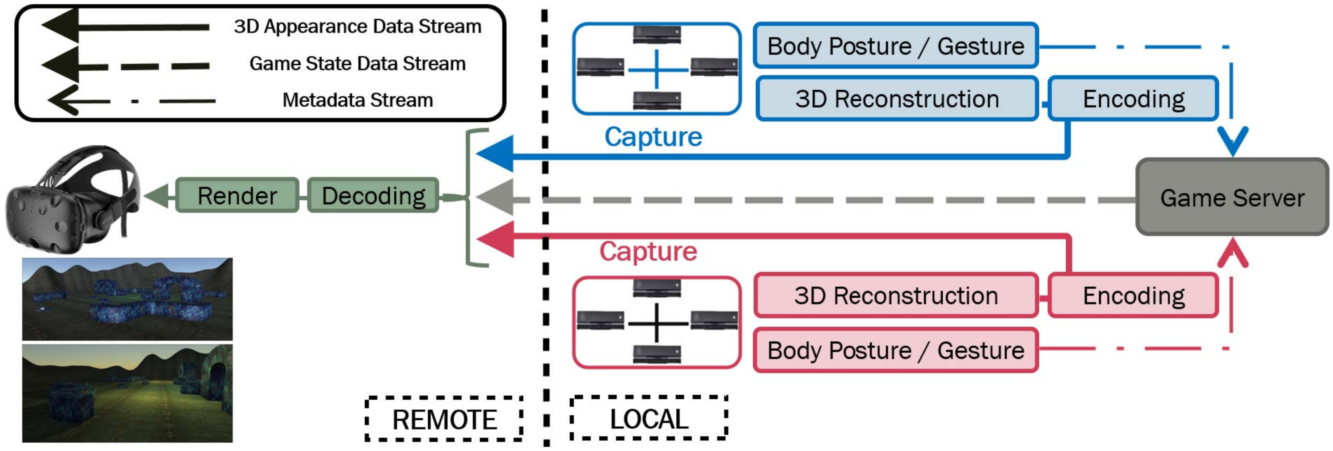

Fig. 1. The end-to-end 3D media tele-immersive pipeline used in the survey. Each local 3D capturing station captures the 3D appearance of a single user, as well as interaction and navigation metadata. The latter are synced by a game server to produce a consistent and synchronised game state which is transmitted along the 3D media to a remote VR spectator. The consuming spectator is able to watch the virtual environment which is augmented with the playing users' realistic 3D representations from any viewpoint.

experience due to unrestricted viewing and ease of developing augmented virtuality applications. The aim is to reconstruct the TI users in real time and embed their 3D appearance in a common virtual environment by streaming the full 3D content in the form of a textured 3D mesh. In this way, complete unrestricted viewing experiences are possible. Furthermore, numerous advantages related to the full three-dimensional information can be exploited like the inherent non-linearity of the content in addition to multi-view productions, enhancement with 3D visual effects, an elevation of the sense of presence due to collisions as well as the real-world scaling of the content and various others. In the rest of the section, a detailed description of the 3D media produced by the utilized 3D-TI pipeline of this study is presented. The section ends by illustrating a novel use-case of this 3D-TI pipeline in next generation immersive gaming for which the QoE study was performed.

A. Immersive 3D Media

Each user participating in a live 3D tele-immersion session is effectively a 3D media producer that streams her own 3D appearance and is supported by a local TI station. The local TI station is responsible for sensor data acquisition, sensor data fusion to a textured 3D mesh (3D reconstruction), textured-mesh encoding and finally data stream transmission as presented in Figure 1. In order to produce a full 3D reconstruction, 4 distinct viewpoints are used, with each viewpoint using an RGB-D sensor to grab synchronized color and depth frames. The depth information is fused into a watertight 3D mesh comprised of vertices, normals and triangles. The geometry is then textured using the corresponding color images. It should be noted though that the geometry and its connectivity is not consistent across frames and thus, the resulting mesh is not dynamic but time-varying, as the generated topology is different for each new frame [30]. More details regarding the spatial alignment of the sensors, the 3D reconstruction process used in this paper and the final texturing can be found in [3], with exemplary screen shots presented in Figure 2.

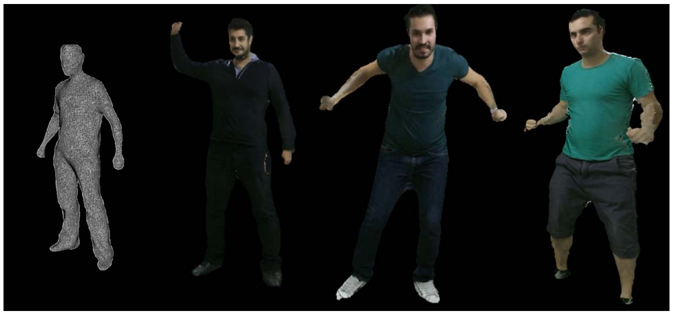

Fig. 2. 3D reconstructions produced by the studied 3D media tele-immersive pipeline. On the leftmost the pure geometry output is shown, while on its right the remaining three 3D reconstructions depict fully textured outputs from viewpoints other than those used to produce them.

The 3D data stream to be transmitted over the network consists of both the 3D geometry, representing the user's shape, as well as the 4 textures representing the user's visual appearance. This creates a high bandwidth scenario due to a number of reasons. As a general rule, the majority of the codecs available for both geometry and texture, require a trade-off between compression efficiency and processing time. Due to the real-time nature of TI, this means that those codecs will probably operate at a suboptimal level of compression efficiency resulting in higher payload sizes to be streamed over the network. Moreover, for geometry compression, there is currently a lack of efficient time-varying mesh codecs that exploit the temporal redundancy between adjacent frames and run in real-time. Given that, our current TI platform [3], utilizes a custom modification of the OpenCTM [31] static mesh compression library that uses LZ4 entropy compression instead of standard LZMA for faster performance. For the texture part of the 3D data stream standard JPEG compression was used instead of using slower but more efficient video codecs such as AVC or HEVC [32].

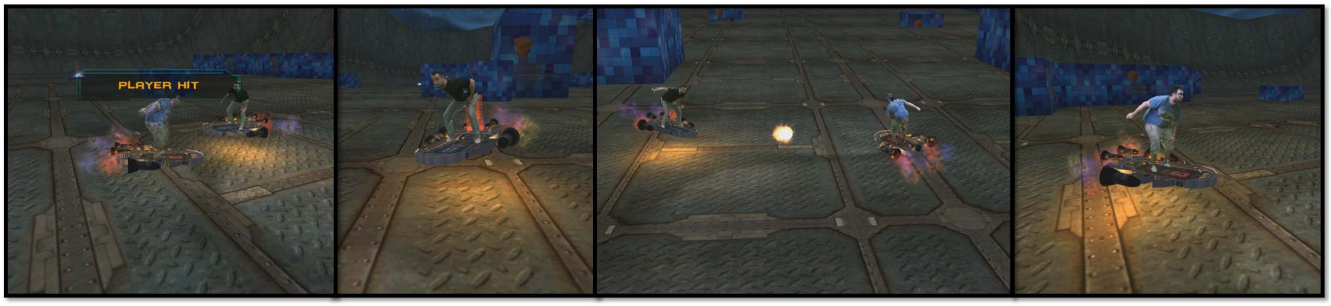

Fig. 3. Screen captures from the 3D-TI game studied in this work. Two players compete against in other inside a virtual arena within which they are embedded with their realistic 3D virtual replicas. Given that the transmitted content is fully three dimensional (3D), the action can be viewed from any angle and position. The middle right screen capture also showcases the projectile that each playing user throws against her/his opponent using a specific gesture. Further visual information about the spectator application and the navigation within the environment can be found in the supplementary video.

For the two distinct data types (3D geometry and 2D textures) included in each immersive 3D media payload, there are a number of parameters affecting the resulting visual quality and payload size. Those parameters can be grouped in two main categories: production parameters and compression parameters. Production parameters include geometry and texture resolution. We mention them as production parameters because they can be explicitly set to arbitrary values when setting-up the 3D reconstruction pipeline. In particular, the 3D reconstruction algorithm with large values for geometry resolution results in an output mesh with an increased number of vertices and triangles. Higher values of geometry resolution leads to better approximation of the 3D human's silhouette. Secondly, when rendering the 3D reconstructed human mesh on a display device, the rendering algorithm is going to use the texture images captured by the RGB-D sensors during the frame acquisition phase of the 3D reconstruction pipeline. Obviously, higher texture and geometry resolutions results in better visual quality but also higher payload size. Especially for the geometry part, higher resolution also means higher processing times. On the other hand, compression parameters include geometry precision and texture bit-rate. The higher the precision and the bit-rate, the better the visual quality but, at the same time, the higher the payload size.

B. Application: Immersive 3D Media Live Broadcasting; Spectating a Live TI Game

Aiming to evaluate the overall experience of spectating a live stream of immersive 3D media, a TI session between 2 participants was employed in the context of an interactive game[1,2] [33], with exemplary screen shots illustrated in Figure 3. The playing users - '*players*' - are immersed into the virtual environment via their realistic appearance through local TI capturing stations. Within the virtual environment, they navigate and interact with each other using gestures and their body postures [1] while competing in a capture-the-flag setting, a highly interactive and fast paced gaming concept, where you need to outmaneuver your opponent and anticipate her/his actions.

A second type of user is also considered - the '*spectator*' that is watching the live session through a client application and can freely navigate within the virtual environment. While it is possible for the content to be displayed in numerous consuming devices (e.g., typical display of a desktop PC, mobile or tablet screen), in this work the focus lies on the resulting experience of the spectators when utilizing a VR headset. This case greatly capitalizes on the free view-point spectating of 3D content and achieves a higher degree of immersion for the spectator [18]. In the developed VR spectating application, special care about preventing motion sickness was taken. Extensive tests in our lab revealed that motion sickness was mainly caused in the cases where the subject is continuously moving inside the virtual environment, while standing still in the physical one. We prevent this from happening by only allowing the VR spectator to navigate inside the virtual world by instant teleportation to the desired location of the virtual environment. The location of the teleportation was controlled by ray casting using one of the VR controllers. It is worth mentioning that, during the experiments, none of our surveyed users mentioned experiencing motion sickness.

There are two types of content that are presented to each user. The first is the static virtual environment that is locally stored at each *player's* and *spectator's* game client. The second is the players' full 3D appearance that is streamed and embedded in the virtual environment by the game clients in real-time.

Given that the underlying application is an interactive game, the *players* and *spectators* are supported by a game server that is responsible for synchronizing the state of the virtual environment. More specifically, the game server receives the gesture and body posture data stream of each player (i.e., the interaction stream) and depending on that input it produces a synchronized game state for both players. This synchronized game state is then streamed to all game clients, both players and spectators. The players' 3D appearance is not explicitly synchronized with the game state. Instead, it is considered as a separate stream that only affects visualization and not game-state and thus it is separately streamed to each game client. This restriction is mainly imposed by the fact that the 3D reconstruction algorithm used to reconstruct the players appearance runs at a much lower rate than the rate at which

[1] https://www.youtube.com/watch?v=nK7pC41YjZY
[2] https://www.youtube.com/watch?v=J3zJmMNxV0k

the game operates. Exact syncing of 3D appearance with the game state would eventually result in low-frame rate game updates leading to considerable amount of game-interaction delay experienced by the players. Separating the interaction stream from the 3D appearance stream allows for almost zero-latency real-time interactions with the game environment while still benefiting from the immersive nature of the embedding of the 3D reconstructed appearance of the players inside the virtual environment.

In summary, Figure 1 shows three data flows for the presented next-generation TI media application, the two heavyweight *players'* 3D appearance data streams and the lightweight global game state data stream, as well as three end-points which consist of the two *players* and the *spectator*.

IV. Experimental Setup

We present the experimental setup of our study in two parts. In Section IV-A we give the exact scope of our study and the aspects that we take into account in greater detail, while in Section IV-B we elaborate on the survey methodology.

A. Scope and Details of the Study

In this study we are aiming to quantify the VR Spectators' QoE in a live 3D-TI gaming session of the application presented in Section III-B. In particular, we examine the case of two participating players plus one VR spectator. The VR spectators are allowed to freely navigate inside the virtual environment of the game and arbitrarily choose their position and orientation by using the VR headset and its controllers. This allows completely unrestricted free viewpoint spectating of the game session. Whilst unrestricted spectating may seem to introduce unfairness in direct comparison of the opinions between different subjects (as each subject may choose to spectate the game from a different viewpoint perspective), nevertheless it captures a realistic scenario. A similar unrestricted viewpoint QoE evaluation has also been conducted before in [34] for 360° video in VR.

While there are numerous network conditions that could be evaluated, we decided to narrow down our study to the following scenario: the two players along with the game server are considered to be co-located in a LAN network environment, while the spectator is assumed to be located at a remote location.

We expect that the QoE of the spectator is affected by two main factors: a) the visual quality of the players' 3D appearances and b) any time inconsistencies between the game-state and the players' visual appearance that are caused by network parameters. While (a) maybe easily understood, for (b) a detailed explanation is given subsequently.

As already discussed in Section III-B, the player's 3D appearance data stream is separated from her interaction data stream. The interaction data stream is refreshed at a high frame-rate and has a very small payload size, allowing the player to interact in real-time with the game environment. Due to its small size, the game-state data stream is delivered at low latency over the network. In a simplified version, the only factor affecting the game-state stream transmission is the network line's latency. On the other hand, the players' 3D appearance lags behind the interaction stream by the amount of time that is imposed by the TI pipeline: i.e., the time needed to reconstruct, compress, transmit and decode it at the receiver side. In a LAN setting the transmission of the player's appearance can be considered to be almost instantaneous, as the network is of almost zero latency and of high bandwidth. However, for a remote receiver (i.e., a distant spectator in this case) the 3D appearance stream is further delayed by the non negligible, time needed to transmit the appearance data over the network. This is affected by the network latency, the throughput, the payload size, the packet loss probability and the network protocol used (i.e., UDP vs TCP).

To summarize, the studied perceived factors that affect the QoE of the VR spectator of the game are

- The players' 3D reconstruction geometry resolution.
- The players' 3D reconstruction texture resolution.
- The players' 3D reconstruction's lag with respect to the game state.

The compression method and parameters used to compress 3D reconstructions, as described in detail in Section III-A, were fixed for all experiments. For geometry, the precision parameters discussed in [35] were chosen, while for textures we used JPEG quality 20% which we experimentally found to be a reasonable compromise between visual quality (measured by Peak Signal-to-Noise Ratio) and payload size.

In order to conduct a valid comparative study over multiple test subjects it is necessary that all surveyed subjects experience the same content. This is not feasible while viewing live 3D-TI gaming sessions, since each game is unique in its own way. For this purpose, a 3D-TI gaming replay system was developed. Initially, the 3D-TI game session is recorded in a LAN setting. During recording, three timestamped streams of data are captured: the first player's 3D appearance stream, the second player's 3D appearance stream and the stream of the game-state. For the purposes of the experiment, two live gaming sessions were recorded. During the recordings, each session was set up using different 3D media production parameters. For the first session, high quality geometry resolution ($r = 6$, [35]) was used while for the second session we used low quality geometry resolution ($r = 5$, [35]). Further, and again during the recordings, for the first session we also used full high definition texture resolution while for the second session the texture resolution was set to half of this. From each of the recorded sessions we artificially produced data corresponding to another session of inferior quality by further downscaling the texture resolution by a factor of two or three. In the production of the artificially produced data we kept the geometry untouched (i.e., the geometry was not altered in any way compared to the original recording). The slightly increased processing time needed to produce downscaled textures compared to the original ones was expected to be mitigated by the reduced time needed to compress the lower resolution textures and thus we conducted the experiments with the assumption that this transformation does not have impact on the corresponding timestamps of the data streams. Overall, two gaming sessions were used to generate the four different sequences that we used in this QoE study.

TABLE I
VISUAL QUALITY PARAMETERS OF THE DIFFERENT SEQUENCES USED IN THIS QoE STUDY

Sequence	**Session**	**Geometry Resolution**	**Texture Resolution**	**Visual Quality**	**Stream Rate (MBit/s)**	**Frame Rate (fps)**	**Duration**
1	2	High	1920 × 1080	a	47.5	13	52s
2	2	High	640 × 360	b	24		
3	1	Low	960 × 540	c	44	17	2min 27s
4	1	Low	480 × 270	d	8.5		

TABLE II
SEQUENCE PARAMETERS

					Mean frame transmission latency (ms)				**Frame loss rate**		
Sequence ID	**Quality level**	**Duration (mm:ss)**	**Protocol**	**RTT (ms)**	**Player 1**	**Player 2**	**Game state**	**Lag (ms)**	**Player 1**	**Player 2**	**Game state**
1	b	00:52	TCP	50	181.87	221.56	25.17	196.39	0.00%	0.00%	0.00%
2	c	02:27	TCP	50	261.42	281.21	25.17	256.03	0.00%	0.00%	0.00%
3	d	02:27	TCP	100	137.14	152.01	50.35	101.66	0.00%	0.00%	0.00%
4	b	00:52	TCP	100	363.74	443.12	50.33	392.79	0.00%	0.00%	0.00%
5	c	02:27	TCP	100	522.84	562.41	50.35	512.07	0.00%	0.00%	0.00%
6	b	00:52	UDP	50	41.38	45.52	25.02	20.51	7.05%	7.63%	0.00%
7	c	02:27	UDP	100	74.70	76.75	50.02	26.74	11.24%	13.05%	0.01%
8	a	00:52	TCP	100	661.90	780.91	50.33	730.58	0.00%	0.00%	0.00%
9	a	00:52	TCP	50	330.95	390.45	25.17	365.29	0.00%	0.00%	0.00%
10	a	00:52	UDP	100	81.96	88.17	50.02	38.15	12.17%	16.77%	0.00%
11	d	02:27	TCP	50	68.57	76.00	25.17	50.83	0.00%	0.00%	0.00%
12	d	02:27	UDP	50	29.55	30.33	25.02	5.31	1.68%	1.69%	0.01%

The selected sequences were chosen in such a way that the coverage of the parametric space of the perceived visual quality is maximized. We label the sequences' visual quality from (a) to (d). The visual quality levels along with their parameters are presented in Table I. Note that the reason for two different session durations is that these were sequences of actual gameplay recorded from live games involving real people.

During the playback of the replay, all streams' timestamps undergo a network simulation transformation depending on the studied network parameters. During the VR spectator survey, all the gaming sessions presented are pre-recorded and played back locally on the test laboratory equipment. The network affect on the traffic streams, i.e., the latency per frame and frame losses, were modeled in the playback software by using modified timestamps of the frames in the pre-recorded streams. At this point it is important to remind the reader that the only wide-area network being studied in this work is that between the remote spectator and the LAN hosting the players and the game server. This means that artificial modification of the spectator's network conditions does not alter the gameplay from the point of view of the two players and hence the use of the same prerecorded player appearance and game state streams with modified timing and loss, is an accurate representation from the viewpoint of the spectator.

In order to study the effect of the network on the data streams delivered to remote spectators we simulated the network latency and loss on the data streams generated by each of the players as well as the game state data in the two recorded gameplay sessions. We considered four different network scenarios: the spectator located at 50ms and 100ms round-trip times (RTT) away from the players and game server; and with the game data being delivered by UDP and TCP transport layer protocols. 50ms RTT corresponds to a geographical distance of approximately 2750km [25], modeling the spectators being in the same continent as the players; 100ms RTT corresponds to a distance of 5500km modeling the spectators being located in a different continent. We assumed that the network path between the players and the spectator had a bottleneck link of capacity 100Mbit/s corresponding to the speed of a typical high-capacity residential broadband connection. Network throughput for TCP traffic was modeled using the Mathis equation relating RTT and packet loss probability to mean transmission rates [36]. UDP throughput was constrained to the maximum rate of the bottleneck link, which we assumed was uncongested in our tests.

Based on the payload size and the generation timestamp of each data frame, we calculated its arrival time at the spectator's equipment for both UDP and TCP protocols and at both RTT latencies. In addition, for the UDP transmission protocol we simulated the effect of packet losses on frame losses. We assumed that a single packet loss would mean that the frame could not be reconstructed. Hence the larger the frame size the greater the frame loss rate, even with identical packet loss probabilities. Frame losses resulted in the player appearance or game state not being updated in the spectator's replay equipment, until the following successful frame was received. It was assumed that no packet and frame losses would occur with TCP traffic as it is a reliable transport protocol and the retransmissions result in overall lower throughput, as calculated by the Mathis equation. The network parameters generated by the simulation of each set of network conditions (protocol and RTT) with each of the two gameplay sessions, each at two quality levels are summarized in Table II. It can be seen that there is a trade-off between frame latency and frame loss when selecting between UDP and TCP protocols. It should be noted that absolute latency is not especially important in non-interactive scenarios such as the currently presented application. Latency translates to a start-up delay at the start of the gameplay session, which it is assumed is not noticeable by the spectators in our experiments. However the relative latency

TABLE III
IMPACT OF VARYING PACKET LOSS RATE ON LAG

Visual Quality	Packet Loss rate	Lag (ms)
a	0.0001	163.4
a	0.001	516.6
b	0.0001	87.8
b	0.001	277.7
c	0.0001	114.5
c	0.001	362.1
d	0.0001	22.7
d	0.001	71.9

TABLE IV
DEMOGRAPHICS OF SURVEYED SUBJECTS

Total subjects	males	females
43	34	9
age		
<30	15	5
30 - 40	13	3
>40	6	1

difference between the players and the game-state data is more important, this is shown as *lag* in Table II, and is calculated as the maximum time difference between the frame arrivals of player appearance data and game-state data. When the lag is large, spectators will notice that the players body movements are unsynchronized with their environment - in particular the hover-board will change direction before the player has shifted their body weight, or a projectile has been released before the player has been seen to throw it.

The sequence parameters shown in Table II were calculated with a fixed packet loss rate at 0.05%, as typical in practical systems where measurements show a loss probability between 10^{-3} and 10^{-4} [37]. Note that lag is dependent on a combination of network latency and packet loss rate and a range of lag values were investigated. Table III shows the range of lag values when the RTT is fixed at 50ms and the loss rate is varied. Comparing the lag ranges in Table II and Table III, we can see that we have examined the full range of lag values, hence there was no need to introduce another variable for loss rate and increase the number of sequences evaluated by our test subjects. To allow a comparison of how users perceive different quantities of lag we did not alter the loss rate, and hence lag, during a sequence run. Thus the dynamic temporal variation of loss is out of the scope of this study and the degree to which variations in lag can impact quality assessment is a potential topic for future investigation.

To sum up, two live recorded game sessions were augmented by further parameterizing among texture resolution to produce four sequences with four different visual qualities labeled from (a) to (d) (Table I). In addition, those four sequences were further transformed by undergoing a network simulation of two different RTT latencies (50ms and 100ms) and two different network protocols (TCP/UDP). Among all the sixteen possible sequences that may be produced by all of those parameters, twelve were shuffled and chosen to be presented to real spectating users for QoE study (with their parameters being depicted in Table II). The four sequences omitted from the evaluation were the ones that gave similar performance with the rest of the included sequences and they were chosen to be omitted in order to be able to limit the survey session to one hour per user.

B. Survey Methodology

To conduct the hereby presented QoE 3D-TI spectator study, in total 43 subjects were surveyed with their demographic distribution presented in Table IV. While the ages of our subjects cover a wide range of values, the gender distribution is more biased towards males. About 79% of our subjects were males while 21% of them were females. Out of all these subjects, twelve remarked that they had previous experience with immersive VR systems. As already mentioned in the last paragraph of Section IV-A, each QoE survey session lasted approximately one hour. For each subject, the survey time was split into four parts; a training part, two sequence evaluation parts and, finally, a questionnaire filling part.

Initially, a training sequence of high visual quality and no network transformation was presented in order for the subject to familiarize with the VR headset and get accustomed to navigating inside the virtual environment. After a 5-minute break, the first six sequences (Sequence IDs 1-6, Table II) were presented to the subject one-by-one. At the end of each sequence's playback, the subject was asked to assess her/his overall experience by giving an opinion score in the scale from "1" (worst) to "5" (best), taking into account the quality of the 3D reconstructions and the perceived lag. A short five minute break followed and then the subjects repeated the same assessment procedure for sequences 7-12. Finally, at the end of the survey, the subjects filled in a short questionnaire containing four quantitative questions about the overall experience with two fields for overall comments.

The questions included in the questionnaire are listed below:

Q1: How would you judge the appearance of the players?
Q2: Did you find the navigation within the virtual environment easy?
Q3: Did you feel comfortable during the spectating sessions?
Q4: Was the movement and position of the players consistent with how you would imagine such a game being played in the real world?

V. EXPERIMENTAL RESULTS

Table V presents the average subject's quality assessment for each individual sequence in the form of a Mean Opinion Score (MOS). The MOS is calculated over two groups of subjects: a) among all the subjects participating in the trials (column "All") and b) among all subjects having previous experience with immersive VR applications (column "Experienced"). MOS was used as being the method proposed in the international standardized subjective video quality assessment methodologies in ITU-T Recommendations P.910 [38], P.913 [39] and BT.500 [40] which include detailed guidelines on how to set up and conduct video quality experiment, allowing a comparison of qualities in the sequences selected. Once the MOS is collected, and in order to refine

TABLE V
MOS RATINGS FOR EACH INDIVIDUAL SEQUENCE

Sequence	**MOS**		**Sequence**	**MOS**	
	Subjects			***Subjects***	
	All	***Experienced***		***All***	***Experienced***
1	3.378	3.167	7	3.486	3.667
2	2.946	3.000	8	3.243	3.000
3	2.946	3.083	9	3.054	2.917
4	3.405	3.417	10	3.405	3.417
5	2.892	3.000	11	3.000	2.750
6	3.459	3.333	12	2.946	2.667

TABLE VI
MOS ON SURVEY'S QUANTITATIVE QUESTIONS. THE RESULTS FOR EACH QUESTION HAVE BEEN SCALED TO THE SAME SCALE (1-5)

Question	**MOS**	
	Subjects	
	All	***Experienced***
Q1	3.110	3.334
Q2	4.302	4.722
Q3	4.256	4.667
Q4	3.698	4.000

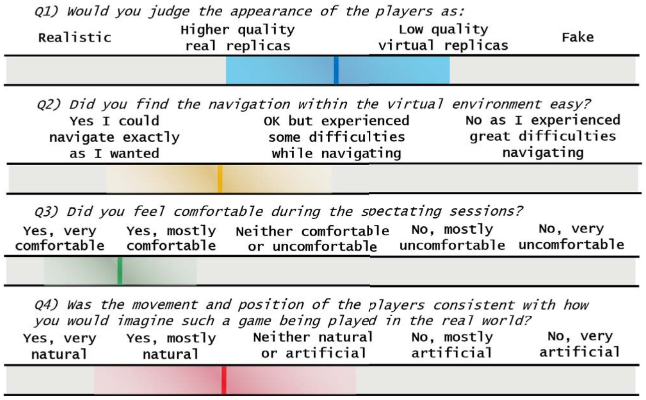

Fig. 4. Questions included in the survey's questionnaire. Mean Opinion Score among all participants is denoted with a bold vertical line while gradient color indicates standard deviation.

TABLE VII
MOS FOR EACH VISUAL QUALITY FOR ALL THE SUBJECTS AS WELL AS THE EXPERIENCED ONES

Subjects	**MOS**			
	Quality			
	a	***b***	***c***	***d***
All	3.234	3.414	3.108	2.964
Experienced	3.111	3.306	3.222	2.833

TABLE VIII
AVERAGE MOS FOR EACH GEOMETRY RESOLUTION FOR ALL THE SUBJECTS AS WELL AS THE EXPERIENCED ONES

Subjects	**MOS**	
	Geometry Resolution	
	Low	***High***
All	3.04	3.32
Experienced	3.03	3.21

TABLE IX
MOS FOR EACH NETWORK PROTOCOL FOR ALL THE SUBJECTS AS WELL AS THE EXPERIENCED ONES

Subjects	**MOS**	
	Protocol	
	TCP	***UDP***
All	3.108	3.324
Experienced	3.042	3.271

the analysis, some results were filtered by removing the outlying subjects. For removing these outliers the average MOS for each subject was set up and a threshold was chosen at two times the absolute average deviation, removing a total of six subjects. Moreover, in Table VI the MOS of the answers on the quantitative questions Q1-Q4 that were introduced in the end of Section IV-B are also presented, with a visual representation of the acquired scores, their standard deviation, as well as the answers to the questions, illustrated in Figure 4. The answers to the questions have been normalized to the scale 1 (negative) to 5 (positive). In the rest of the section we set target questions that we aim to answer and showcase semantic notions obtained from the analysis of the survey results.

Which resolution parameter influences the resulting QoE more, geometry or texture resolution?

In order to answer this question, we compute the subjects' MOS against all the sequences of the same visual quality and eventually obtaining an average score that characterizes the sequence that is independent of any network conditions. The calculated MOS scores for each individual visual quality are illustrated in Table VII, while Table VIII further groups and averages the MOS scores across Geometry resolution levels. The highest scores are obtained for Visual Qualities (b) and (a) which both of them correspond to High Geometry Resolution, essentially answering our target question. In addition, Visual Quality (b) scores higher than (a) giving a hint that the reduced lag benefit obtained by the usage of the downscaled texture matters more than the extra fidelity provided by the higher texture resolution.

Which protocol is more suitable for spectating 3D media, TCP or UDP ?

The UDP protocol is well known for its improved latency performance over TCP but at the cost of unreliable transmission. In order to obtain an indication of which protocol is more suitable for spectating 3D media, we average the scores given to each sequence utilizing TCP in separation from the sequences utilizing UDP obtaining two MOS scores, once for each individual case and independent of any visual quality parameters. The resulting numbers are depicted in Table IX. From the table it is deduced that the average subject, being either experienced with VR or not, scored UDP higher than TCP. Eventually, from a QoE perspective, this means that the reduced perceived lag obtained by the usage of the unreliable transport layer protocol is preferred at the cost of some frame drops. Finally, we split all the sequences in two other categories: the ones with high lag (above 250ms) and the ones with low lag (below 250ms). This threshold was empirically selected since its purpose is only to show whether the MOS is different for extreme values of lag, in a more analytical way, the exact value of lag will be considered. The intuition behind its selection was based on typical values for casual gaming latency, the average human reaction time - given that the users are using their own bodies to navigate and interact - and the rate at which the user's body posture is captured. We

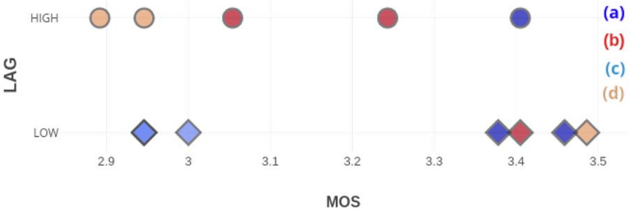

Fig. 5. MOS scores for each individual sequence. Visual Quality levels are color encoded, while shapes and vertical positions of the markers denote lag conditions.

TABLE X
MOS FOR SEQUENCES OF HIGH AND LOW LAG FOR ALL THE SUBJECTS AS WELL AS THE EXPERIENCED ONES

Subjects	**MOS**	
	Lag	
	Low	***High***
All	3.232	3.108
Experienced	3.155	3.067

compute the MOS scores over those groups of sequences in order to understand whether the average subject was able to distinguish the two cases. In Table X, it is shown that the average subject, scored the sequences of low lag higher than the sequences of high lag, as expected. In further detail, Figure 5, depicts the MOS score for each individual sequence. In that Figure, the visual quality levels are color encoded and the lag conditions (high/low) are encoded in the marker's shape and position.

What is the most efficient way of navigating within a virtual 3D environment using VR headsets?

While developing the VR spectating application for this QoE study, we experimented with various alternatives for the navigation of the spectators inside the virtual world. As already discussed in Section II-B, in the literature it is well documented that HMDs are prone to causing effects of nausea or motion sickness to the participants. In either case, for this study to be successful and valid, we wanted to give the subjects an easy and comfortable method to navigate around the virtual environment. Our internal tests showed that motion sickness and discomfort are mostly caused in the cases where the subjects continuously move inside the virtual environment while standing still in the physical one. This means that spectators should not continuously follow the players' movements inside the game world. However, an efficient way to navigate and spectate the game action was necessary. To overcome the issue while still allowing free viewpoint spectating and full freedom of navigation inside the virtual environment we employed a teleportation paradigm. By utilizing the VR headset's controllers, the spectator casts a ray inside the virtual world and selects a point on the game terrain where she/he would like to teleport. Teleportation is instant and no virtual movement is conducted essentially removing any chances of causing nausea. Although we did not conduct a dedicated quantitative study for all the various navigation alternatives that we developed, the described way was assessed to be the best after in-house testing. Further, during the survey, none of our subjects complained about effects of motion sickness or nausea. On the contrary, the subjects found the navigation system easy and comfortable to use, something that is also confirmed by the results of the survey. As illustrated in Table VI, the MOS scores for questions Q2 and Q3 which are relevant to the navigation system and the overall VR experience are strongly positive.

What was the overall perceived quality regarding the immersive 3D media?

After the end of the experiment, each subject was asked to offer their opinion on the realism of the virtual replicas (the *players*' 3D reconstructions) - Question Q1 on the survey's questionnaire. As presented in Table VI the MOS score is approximately 3 out of 5. Moreover, from the individual statistics of the results we have deduced that the subjects split evenly between high and low judgment of visual quality, while none selected the extremes of Fake or Realistic.

VI. QOE PREDICTION MODEL

In this section, we present a preliminary study in constructing a model that will be able to predict the VR Spectator's subjective QoE MOS score given the parameters used in production of the TI content as well as the networking conditions. In order to know the importance of each parameter, and decide which to include in our final model, we performed a multiple regression by using all potential inputs: Geometry Resolution, Frame Rate, Frame Loss, Lag, Network Protocol, Texture Resolution, RTT and Stream Rate.

Nonetheless, the frame rate input variable has no influence at all into the model. This is explained by the relation of the geometry resolution to the processing time to produce each frame, and thus, the overall frame rate, as already mentioned in Section III-A. This can be confirmed in Table I, where it is shown how the frame rate is linked to the geometry resolution. Taking that into account, frame rate was removed from the input variables of the model. We then obtained the p-values [41] for the rest of inputs. With these p-values, we found that the frame loss input variable, with a p-value of 0.66, has a very low influence in the model. Consequently, the frame loss variable was also removed from the model.

Qualitatively, only the geometry resolution, texture resolution, network protocol and RTT are full independent variables that may affect the final QoE of the subjects. However, analytically, the influence of some of these independent variables may be modeled by lag and stream rate. A simple model is not able to find this dependency, thus, a more complex model is required when using only independent parameters. To that end, we opted for a neural network model.

The proposed neural network prediction model has four inputs (each one corresponding to the independent production parameters discussed above) and one hidden layer of 10 cells, each performing a logistic regression. Finally, as shown in Figure 6, the output layer, consisting of only a single cell, performs a linear regression to predict the final MOS score. All the neural network's input parameters are encoded in floating point values in the interval [0,1], with the discretization avoiding extreme values near 0 and 1.

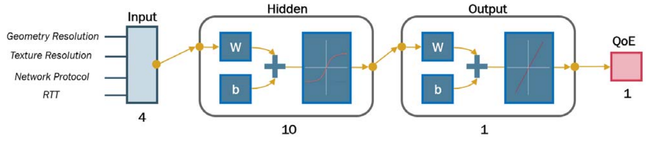

Fig. 6. The neural network architecture used for predicting the QoE of the TI system used in this survey. The numbers below the neural networks components denote the amount of input parameters, hidden and output cells used to predict the final QoE value.

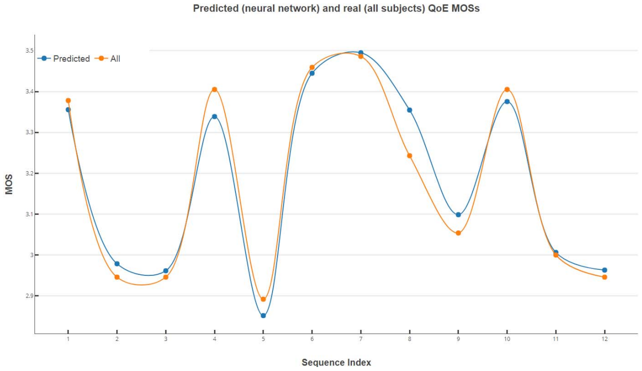

Fig. 7. Neural network QoE prediction evaluation. Orange curve: MOS scores given to each individual sequence from all subjects. Blue curve: predicted MOS Scores obtained from the trained neural network.

The network is trained using the Levenberg-Marquardt back-propagation algorithm [42], [43]. The set of 12 sequences among with their average MOS values were split to 70% for training, 15% for validation and 15% for testing (eventually leading to a training set of 8 samples and validation and test sets with 2 samples). The validation set was used in order to stop the training phase at the correct epoch to avoid overfitting.

Once the neural network was trained, we obtained a comparison between the actual values for each sequence given by all subjects and the predicted values from the neural network. These values are presented in Figure 7.

The Pearson and Spearman coefficients for the correlation between real and predicted data are 0.98 and 0.96 respectively. In order to identify which input variables have the most impact on the neural network output, and gaining further insights on the parameters affecting the QoE, the Garson algorithm [44] was used to calculate the relative importance percentage of each input variable. Table XI shows that all input variables have similar importance. This maybe means that no input variable dominates the resulting QoE, revealing the complexity of the problem.

By training the same model with the 4 independent input parameters plus the lag and the stream rate, no significant changes on the results were noticed. Thus, the neural network is able to find out the independent variables, and it can also model the QoE from them.

VII. Conclusion and Future Work

In this work we conducted a subjective QoE study on spectating a two-player 3D-TI game using VR headsets. The parameters affecting QoE that were taken into account were related to both visual quality of the "3D replicas" of the players, as well as the physical conditions of the players-to-spectator network paths. To conduct the study, two actual gaming sessions were pre-recorded and transcoded off-line to different quality levels and replayed to the subjects by taking into account the simulation of player-to-spectator network latency and loss degradations. The subjects participating in the study were spectating the recorded game sessions in VR and were able to freely navigate within the game's virtual environment.

TABLE XI
Relative Importance of Each Input Parameter of the Neural Network

Input variable	Relative importance
Geometry resolution	23.5346
Texture resolution	27.9650
Network Protocol	21.5971
RTT	26.9033

At the end of each spectating session the subjects were asked to rank their overall experience with a score from 1 (worst) to 5 (best). We performed a statistical analysis of the subjects' MOSs and qualitative comments. We found that the navigation in the VR setting was satisfying and that QoE was influenced by both visual quality and network lag. This is contrast to traditional video where QoE can be predicted mainly by its bitrate. In 3D media, higher visual quality requires higher bandwidth streams which means higher lag between game-state and visual appearance when using a reliable transport layer protocol. In case we employed a buffering mechanism when using the reliable transmission protocol, many of the lag issues experienced by the spectators would potentially be eliminated. However, buffering is not an option in applications where the spectators would like to interact with the players in the live game. While this exact case is not studied in the present work, the findings of our study very well applies to this future scenario.

While the complex relationship between geometry resolution, texture resolution and network lag is difficult to model analytically, we have developed a neural network model that is able to predict user QoE scores from input visual quality and network parameters. This indicates that QoE for the transmission of 3D media streams over the Internet is a complex combination of multiple parameters. Thus, in the absence of exact mathematical models, QoE can be modeled by machine learning mechanisms and is a potential method for the prediction of user satisfaction.

This is one of the first studies of QoE in the area of 3D-TI media and VR, aiming to stimulate further future work and experimentation. Future studies are needed into developing an exact mathematical model for QoE prediction, rather than using a neural network as presented in this paper. Additionally, we intent to embed QoE prediction models algorithms in an overall practical system implementation for managing the deployment and delivery of interactive and immersive 3D media between players, and between players and spectators, distributed around the globe.

References

[1] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, Jun. 2011, pp. 1297–1304. [Online]. Available: https://www.microsoft.com/en-us/research/publication/real-time-human-pose-recognition-in-parts-from-a-single-depth-image/

[2] G. Kurillo and R. Bajcsy, "3D teleimmersion for collaboration and interaction of geographically distributed users," *Virtual Reality*, vol. 17, no. 1, pp. 29–43, Mar. 2013, doi: 10.1007/s10055-012-0217-2.

[3] D. S. Alexiadis *et al.*, "An integrated platform for live 3D human reconstruction and motion capturing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 798–813, Apr. 2017.

[4] *Augmented Virtuality.* Accessed: Apr. 12, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Mixed_reality#Augmented_virtuality

[5] A. Karakottas *et al.*, "Augmented VR," in *Proc. IEEE Virtual Reality*, Mar. 2018. [Online]. Available: https://www.youtube.com/watch?v=7O_TrhtmP5Q

[6] W. Wu *et al.*, "'I'm the jedi!'—A case study of user experience in 3D tele-immersive gaming," in *Proc. IEEE Int. Symp. Multimedia*, Taichung, Taiwan, Dec. 2010, pp. 220–227.

[7] K. Venkatraman *et al.*, "Quantifying and improving user quality of experience in immersive tele-rehabilitation," in *Proc. IEEE Int. Symp. Multimedia*, Taichung, Taiwan, Dec. 2014, pp. 207–214.

[8] S. Orts-Escolano *et al.*, "Holoportation: Virtual 3D teleportation in real-time," in *Proc. 29th ACM Annu. Symp. User Interface Softw. Technol. (UIST)*, Tokyo, Japan, 2016, pp. 741–754. [Online]. Available: http://doi.acm.org/10.1145/2984511.2984517

[9] S. Chen, K. Nahrstedt, and I. Gupta, "3DTI amphitheater: A manageable 3DTI environment with hierarchical stream prioritization," in *Proc. 5th ACM Multimedia Syst. Conf. (MMSys)*, Singapore, 2014, pp. 70–80. [Online]. Available: http://doi.acm.org/10.1145/2557642.2557654

[10] C. T. E. R. Hewage and M. G. Martini, "Quality of experience for 3D video streaming," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 101–107, May 2013.

[11] Y. Liu, S. Ci, H. Tang, Y. Ye, and J. Liu, "QoE-oriented 3D video transcoding for mobile streaming," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3s, pp. 1–20, Oct. 2012. [Online]. Available: http://doi.acm.org/10.1145/2348816.2348821

[12] S. Beck, A. Kunert, A. Kulik, and B. Froehlich, "Immersive group-to-group telepresence," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 4, pp. 616–625, Apr. 2013.

[13] G. C. Burdea and P. Coiffet, *Virtual Reality Technology*, 2nd ed. New York, NY, USA: Wiley, 2003.

[14] S. Manjrekar *et al.*, "CAVE: An emerging immersive technology—A review," in *Proc. UKSim-AMSS 16th Int. Conf. Comput. Model. Simulat.*, Cambridge, U.K., Mar. 2014, pp. 131–136.

[15] B. S. Santos *et al.*, "Head-mounted display versus desktop for 3D navigation in virtual reality: A user study," *Multimedia Tools Appl.*, vol. 41, no. 1, pp. 161–181, Jan. 2009, doi: 10.1007/s11042-008-0223-2.

[16] F. Weidner, A. Hoesch, S. Poeschl, and W. Broll, "Comparing VR and non-VR driving simulations: An experimental user study," in *Proc. IEEE Virtual Reality (VR)*, Los Angeles, CA, USA, Mar. 2017, pp. 281–282.

[17] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays," in *Proc. 9th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, Erfurt, Germany, May/Jun. 2017, pp. 1–6.

[18] K. Kim, M. Z. Rosenthal, D. Zielinski, and R. Brady, "Comparison of desktop, head mounted display, and six wall fully immersive systems using a stressful task," in *Proc. IEEE Virtual Reality Workshops (VRW)*, Costa Mesa, CA, USA, Mar. 2012, pp. 143–144.

[19] R. Schatz, A. Sackl, C. Timmerer, and B. Gardlo, "Towards subjective quality of experience assessment for omnidirectional video streaming," in *Proc. 9th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, Erfurt, Germany, May 2017, pp. 1–6.

[20] I. Hupont, J. Gracia, L. Sanagustíon, and M. A. Gracia, "How do new visual immersive systems influence gaming QoE? A use case of serious gaming with oculus Rift," in *Proc. 7th Int. Workshop Qual. Multimedia Exp. (QoMEX)*, Pylos-Nestor, Greece, May 2015, pp. 1–6.

[21] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A QoE evaluation of immersive augmented and virtual reality speech & language assessment applications," in *Proc. 9th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, Erfurt, Germany, May 2017, pp. 1–6.

[22] J. P. López, J. A. Rodrigo, D. Jiménez, and J. M. Menéndez, "Stereoscopic 3D video quality assessment based on depth maps and video motion," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, p. 62, Dec. 2013, doi: 10.1186/1687-5281-2013-62.

[23] N. R. Veeraragavan, H. Meling, and R. Vitenberg, "QoE estimation models for tele-immersive applications," in *Proc. Eurocon*, Zagreb, Croatia, Jul. 2013, pp. 154–161.

[24] T. He *et al.*, "Immersive and collaborative Taichi motion learning in various VR environments," in *Proc. IEEE Virtual Reality (VR)*, Los Angeles, CA, USA, Mar. 2017, pp. 307–308.

[25] R. Landa *et al.*, "The large-scale geography of Internet round trip times," in *Proc. IFIP Netw. Conf.*, Brooklyn, NY, USA, May 2013, pp. 1–9.

[26] P. Simoens *et al.*, "Service-centric networking for distributed heterogeneous clouds," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 208–215, May 2017, doi: 10.1109/MCOM.2017.1600412.

[27] A. Langley *et al.*, "The QUIC transport protocol: Design and Internet-scale deployment," in *Proc. Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*, Los Angeles, CA, USA, 2017, pp. 183–196. [Online]. Available: http://doi.acm.org/10.1145/3098822.3098842

[28] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-based congestion control for unicast applications," in *Proc. Conf. ACM Appl. Technol. Architect. Protocols Comput. Commun. (SIGCOMM)*, Stockholm, Sweden, 2000, pp. 43–56. [Online]. Available: http://doi.acm.org/10.1145/347059.347397

[29] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the Internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011, doi: 10.1109/MMUL.2011.71.

[30] S.-R. Han, T. Yamasaki, and K. Aizawa, "Time-varying mesh compression using an extended block matching algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1506–1518, Nov. 2007.

[31] *OpenCTM*. Accessed: Apr. 12, 2018. [Online]. Available: http://openctm.sourceforge.net/

[32] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," in *Proc. Picture Coding Symp. (PCS)*, San Jose, CA, USA, Dec. 2013, pp. 394–397.

[33] N. Zioulis *et al.*, "3D tele-immersion platform for interactive immersive experiences between remote users," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 365–369.

[34] H. T. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A subjective study on QoE of 360 video for VR communication," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Luton, U.K., Oct. 2017, pp. 1–6.

[35] D. Alexiadis, A. Doumanoglou, D. Zarpalas, and P. Daras, "A case study for tele-immersion communication applications: From 3D capturing to rendering," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Valletta, Malta, Dec. 2014, pp. 278–281.

[36] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," *SIGCOMM Comput. Commun. Rev.*, vol. 27, no. 3, pp. 67–82, Jul. 1997. [Online]. Available: http://doi.acm.org/10.1145/263932.264023

[37] D. Zhang and D. Ionescu, "Reactive estimation of packet loss probability for IP-based video services," *IEEE Trans. Broadcast.*, vol. 55, no. 2, pp. 375–385, Jun. 2009.

[38] "Subjective video quality assessment methods for multimedia applications," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation P. 910, 1999.

[39] "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation P. 913, 2016.

[40] "Methodology for the subjective assessment of the quality of television pictures," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation I. REC Bt. 500-12, 2009.

[41] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: Context, process, and purpose," *Amer. Stat.*, vol. 70, no. 2, pp. 129–133, 2016.

[42] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Appl. Math.*, vol. 11, no. 2, pp. 431–441, Jun. 1963, doi: 10.1137/0111030.

[43] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.

[44] G. D. Garson, "Interpreting neural-network connection weights," vol. 6, no. 4, pp. 46–51, 1991.

Alexandros Doumanoglou received the graduation degree from the Electrical and Computer Engineering Department, Aristotle University of Thessaloniki in 2009. He has been an Electrical and Computer Engineer with the Informatics and Telematics Institute since 2012. His research interests include computer vision, pattern recognition, machine learning, 3-D reconstruction, 3-D Graphics, and GPGPU computing.

David Griffin received the B.Sc. degree in electronic and electrical engineering from Loughborough University and the Ph.D. degree in electronic and electrical engineering from University College London (UCL), where he is a Principal Research Associate with the Department of Electronic and Electrical Engineering. His research interests include planning, management and dynamic control for providing QoS in multiservice networks and novel routing paradigms for the future Internet.

Javier Serrano received the graduated degree in sound and image engineering from the Universidad Politécnica de Madrid (UPM). He has also acquired a Research Master on signal, image, speech, and telecommunications with INP Grenoble, France. He is currently with Grupo de Aplicación de Telecomunicaciones Visuales, UPM focusing on testing and validation of new UHD and immersive content delivery models in next-generation mobile networks.

Nikolaos Zioulis has been an Electrical and Computer Engineer with the Aristotle University of Thessaloniki since 2012. He has been with the Information Technologies Institute, Centre for Research and Technology Hellas since 2013. His interests include 3-D processing and graphics, particularly performance oriented real-time computer vision, 3-D reconstruction and graphics, multi modal acquisition, and tele-immersion technology.

Truong Khoa Phan received the Ph.D. degree from INRIA/I3S, Sophia, France. He is currently a Research Associate with the Department of Electronic and Electrical Engineering, University College London. His research interests include network optimization, cloud computing, multicast, and P2P.

David Jiménez received the Telecom Engineer and Telecom Ph.D. degrees from the Universidad Politécnica de Madrid (UPM) in 2004 and 2012, respectively, where he is currently a Lecturer. His research interests include image processing, digital video broadcasting, video compression, and HDTV.

Dimitrios Zarpalas received the M.Sc. degree in computer vision from Pennsylvania State University, USA, in 2006, and the Ph.D. degree in medical informatics from Medical School, Aristotle University of Thessaloniki in 2014. He is an Electrical and Computer Engineer with the Aristotle University of Thessaloniki since 2003. In 2007, he joined the Information Technologies Institute, Centre for Research and Technology Hellas. His current research interests include real time tele-immersion applications, 3-D computer vision, AR technologies, 3-D object recognition, 3-D motion capturing and evaluation, while in the past has also worked in indexing, search and retrieval and classification of 3-D objects and 3-D model watermarking.

Federico Alvarez (M'07) received the Telecom Engineer degree (Hons.) and Ph.D. degree (*cum laude*) from the Universidad Politécnica de Madrid in 2003 and 2009, respectively, where he is currently a Assistant Professor. He has co-authored over 60 papers, books, book chapters, and patents in the field of ICT networks and multimedia technologies. He coordinated six EU projects in the last eight years.

Miguel Rio is a Professor with the Department of Electronic and Electrical Engineering, University College London, where he researches and lectures on Internet technologies. His research interests include on real-time overlay streaming, network support for interactive applications, quality of service routing, and network monitoring and measurement.

Petros Daras received the Diploma degree in electrical and computer engineering, the M.Sc. degree in medical informatics and the Ph.D. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is a Principal Researcher Grade A', with the Information Technologies Institute, Centre for Research and Technology Hellas. He is the Head Researcher of the Visual Computing Laboratory coordinating the research efforts of over 35 scientists. He has co-authored over 160 papers in refereed journals and international conferences, and has been involved in over 30 national and international research projects. His research interests include 3-D media processing and compression, multimedia indexing, classification and retrieval, annotation propagation and relevance feedback, bioinformatics and medical image processing.