Human-centric Quality Management of Immersive Multimedia Applications

Conference Paper · April 2020

CITATIONS
0

READS
55

3 authors:

Sam Van Damme
Ghent University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Maria Torres Vega
Ghent University
38 PUBLICATIONS **266** CITATIONS
SEE PROFILE

Filip De Turck
Ghent University
731 PUBLICATIONS **8,874** CITATIONS
SEE PROFILE

Human-centric Quality Management of Immersive Multimedia Applications

Sam Van Damme, Maria Torres Vega and Filip De Turck
IDLab, Department of Information Technology, Ghent University - imec
E-mail: sam.vandamme@ugent.be

***Abstract*—Augmented Reality (AR) and Virtual Reality (VR) multimodal systems are the latest trend within the field of multimedia. As they emulate the senses by means of omni-directional visuals, 360° sound, motion tracking and touch simulation, they are able to create a strong feeling of presence and interaction with the virtual environment. These experiences can be applied for virtual training (Industry 4.0), tele-surgery (healthcare) or remote learning (education). However, given the strong time and task sensitiveness of these applications, it is of great importance to sustain the end-user quality, *i.e.* the Quality-of-Experience (QoE), at all times. Lack of synchronization and quality degradation need to be reduced to a minimum to avoid feelings of cybersickness or loss of immersiveness and concentration. This means that there is a need to shift the quality management from system-centered performance metrics towards a more human, QoE-centered approach. However, this requires for novel techniques in the three areas of the QoE-management loop (monitoring, modelling and control). This position paper identifies open areas of research to fully enable human-centric driven management of immersive multimedia. To this extent, four main dimensions are put forward: (1) Task and well-being driven subjective assessment; (2) Real-time QoE modelling; (3) Accurate viewport prediction; (4) Machine Learning (ML)-based quality optimization and content recreation. This paper discusses the state-of-the-art, and provides with possible solutions to tackle the open challenges.**

***Index Terms*—Quality-of-Experience (QoE) Management, Immersive Media, Virtual Reality (VR), Haptics**

I. Introduction

Augmented Reality (AR) and Virtual Reality (VR) multimodal experiences are the latest revolution within multimedia applications [1]. By emulating (certain) senses as accurately as possible, they create a realistic and interactive virtual or augmented environment. Omni-directional video or holograms (point clouds or light fields), combined with 360° sound, motion tracking and touch simulation (gloves or suit) make the users feel that they truly are "in" the environment. These truly immersive multimodal systems are already being adopted in the entertainment sector, e.g. for gaming and video content such as PlayStation VR [1] or Netflix VR[2]. In addition, they open good opportunities for sectors with more societal and economic impact. Future applications in industry (Industry 4.0), healthcare (tele-surgery) and education (remote learning), just to name a few, will highly benefit from this multi-sensor systems. However, they impose stringent conditions (high bandwidth, ultra-low latency) that could negatively affect the user. For instance, small desynchronization of inputs could lead to dizziness [2]. Furthermore, quality degradation could induce lack of immersiveness or concentration. AR/VR multimodal experiences are built from multiple input and feedback channels to improve the user's feeling of immersion and interactivity. In order to make this feel as a natural experience, however, the system's Quality-of-Service (QoS) parameters need to be kept within the boundaries of acceptance of human perception. Delay, jitter and data loss, for example, are perceived rather differently depending on the sensorial type (Table I). When requirements are not fulfilled, the user experience will feel less authentic and the Quality-of-Experience (QoE) decreases, possibly even inducing cybersickness in the most severe cases.

Table I
Overview of acceptable QoS parameters for audio, video, VR and haptics [3], [4]

QoS Param.	Audio	Video	Graphics	VR	Haptics
Delay [ms]	≤ 150	≤ 400	$[100-300]$	15	$[3-60]$
Jitter [ms]	≤ 30	≤ 30	≤ 30	?	$[1-10]$
Packet loss [%]	≤ 1	≤ 1	≤ 10	?	$[0.01-10]$

Currently, most research is limited to delay while jitter and data loss are barely researched [1], [5]. In addition, most test setups include only one or two modalities, while studies on truly multi-modal systems are rather scarce. As a result, the synchronization and prioritization of the simultaneously transmitted signals is barely investigated within the light of QoE maximalization. Thus, it is not hard to imagine that (a lack of) synchronization between the multiple feeds will have a highly influence on the user experience. As most of the current network infrastructure consists of reliable, high-speed connections, synchronization mistakes typically arise within the local network. Especially when using wireless connections (which is preferred to enable maximal freedom of movement), the above QoS parameters might become more stringent. Managing these applications will require to shift the focus from the network (QoS) to the human. Therefore, the expectations of the users, i.e. the QoE, will drive the application and network decisions.

[1]https://www.playstation.com/nl-be/explore/playstation-vr/
[2]https://play.google.com/store/apps/details?id=com.netflix.android_vr&hl=nl

Pre-printed version. Please cite as: S. Van Damme, M. Torres Vega and F. De Turck. "Human-centric Quality Management of Immersive Multimedia Applications" *4th International Workshop on Quality of Experience Management.* 2020.

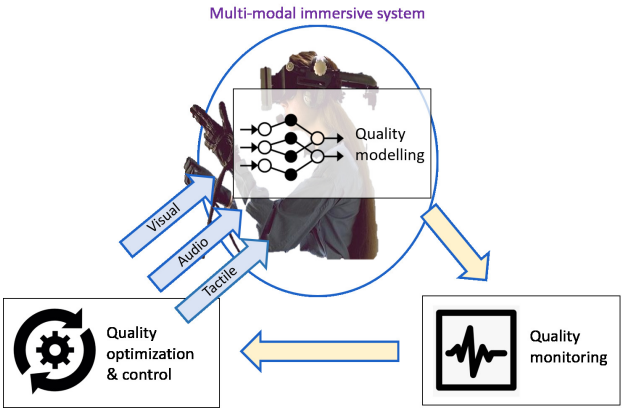

Figure 1. QoE management loop for immersive multimedia systems.

The purpose of this position paper is to present the challenges and possible solutions to enable QoE-driven management of immersive multimodal applications. Therefore, four open areas of research have been identified. For each of them, an overview of the state-of-the-art as well as discussion on possible research directions to tackle the challenges are provided.

The remainder of this paper is distributed as follows. section II provides a short overview of the different elements of the QoE management loop, pointing at how it has been traditionally done and presenting the open challenges for immersive multimodal media. Based on this analysis, four areas of research have been identified. Sections III to VI provide with state of the art, challenges and possible solutions for each of the four areas. Finally section VII concludes this paper.

II. Enabling the QoE-management loop for Immersive Multimedia Systems

QoE is defined as *the degree of delight or annoyance of the user of an application or service, based on both objective and subjective psychological measurements* [6]. In order to enable the end-user to experience the given application in the most optimal way, QoE-management mechanisms need to be incorporated. Three elements conform the QoE management loop (Figure 1): quality modelling, quality monitoring and quality optimization and control. Quality modelling aims at providing an accurate estimate of the client-side quality as perceived by the end-user. The quality monitoring component focuses on analyzing the services and gaining understanding of the different factors that influence the quality of the application. Optimally, this needs to happen in a (near) real-time fashion such that the system parameters can be adapted accordingly to optimize the QoE over time. This is done in the quality optimization and control part, within the limits of the available resources.

For the case of traditional 2D or omnidirectional 3D video delivery, quality is assessed by means of subjective studies. These are then used as a ground truth for the creation of objective, mathematical metrics that describe visual quality that reflect human perception (quality modelling). Finally, quality optimization and control is often realized in terms of adaptive streaming, in which the content is provided as multiple streams, each established using different encoding parameters [6]. While this loop provides satisfactory results for (audio)visual feeds, it becomes insufficient for truly immersive multimodal applications. Therefore, enabling the QoE management loop for these systems requires novel approaches in all loop elements. In particular, four open challenges are put forward:

1) **No standardized methodologies exist for subjective testing of multimodal experiences.** Furthermore, research needs to be conducted towards the influence of each of the feeds on the end-user perception, the presence of cybersickness and their ability to complete the tasks at hand.
2) **The time-critical characteristics of non-entertainment related immersive experiences make realtime end-user quality assessment essential.** As such, there can be directly intervened whenever cybersickness or an inability to perform the task at hand tend to occur. Therefore, objective, light-weight real-time quality models are needed as well as user-independent benchmarks to evaluate them.
3) In order to sustain the end-user perception it is of great importance to know at which area of the VR or AR he/she is looking at as fast as possible. Therefore, **accurate and real-time viewport prediction** is fundamental for quality control of the omnidirectional visuals.
4) **Proactive quality optimization and content enhancement algorithms (in case of late delivery) should be implemented to optimize user perception.** Therefore, challenging problems arise such as decision-making on if and when to recreate content upon late arrival.

The remainder of this paper provides an overview of the current state-of-the-art of these four challenges. Furthermore, for each of them, future research directions are discussed.

III. Subjective evaluations: From Rating visual quality to Assessing performance

The most straight-forward manner to understand the effects of multimedia feeds on the user's perception has traditionally been subjective evaluations [7]. Subjective studies and evaluations of multimedia services are typically performed by means of experimental setups in laboratory environments. Typical tests include a few dozen people with varying demographic backgrounds and limited knowledge about signal processing and encoding. For certain, more traditional applications such as 2D video, testing conditions are standardized by the International Telecommunication Union (ITU) [7]. Tests are performed by offering multiple sequences with different degrees of impairment to the subjects in either a double stimulus approach (the subject experiences both the original and the impaired sequence at the same time) or a single stimulus approach (the subjects only experience one sequence at a time). For each sequence, users grade the quality on a

certain scale (usually between 1 and 5). The average score over all subjects is called the Mean Opinion Score (MOS) [7].

As the degree of immersiveness of the application rises, however, assessing the QoE of the end-user becomes more difficult [8]. In 360° videos, for example, one must take into account that a user only sees a limited part of the 360° hemisphere (the viewport) at each instant [9], [10]. Therefore, users might watch different portions of the video during playback, which makes it rather difficult to compare quality scores among users and to combine them to a single average score per video. At this point in time, no standardized assessing methodology exists to solve this issue.

It becomes even more complex when additional sensory inputs, and tactile feedback in particular, are added to the experience. Due to the complex combination of sensorial data types that influence the user, MOS becomes rather infeasible to define the quality. Therefore, it is beneficial to define the effectiveness of the interactive system in terms of the ability to perform certain tasks, e.g. the ability to pick up an object, localizing an object etc. [2] and the feeling (or rather the absent) of cybersickness. Research towards the actual design of such performance tests is currently scarce, however.

As such, we propose to enhance subjective quality assessment by interpreting the end-user QoE of AR/VR multimodal applications as the combination of *performance* and *well-being*. We express *performance* as *the ability to perform a certain task within a reasonable amount of time*, e.g. picking up an object, or locating a certain place, in addition to the numerical quality scoring [2]. Observe that we interpret performance within the context of the *user*, rather than the immersive system itself. The goal is to measure this by means of a fully data-driven approach, to avoid intrusion of the application. Different metrics will be used to assess performance, such as the time needed to complete the task, the number of attempts before success, the average accuracy (e.g. hitting a target etc.) [11]. A possible experiment could take place in a remote presence setting, for example, in which the user is asked to localize an object (e.g. a ball) based on visual and audio feedback. Afterwards, a remote robot arm should be controlled by means of tactile feedback to grab the object and put it in a basket.

We define *well-being* as *the opposite of cybersickness, with cybersickness being the physical discomfort resulting from use of the immersive system, obstructing the user from accurately performing the task at hand.* Note that our definition of well-being slightly differs from the one common in literature, as ours focuses on the effects on the user *during* playback of the immersive experience while the term in literature mostly refers to the long-term effects of repeated exposure to these kinds of experiences [12]. Well-being could be measured by means of questionnaires prior and subsequent to the experiment, in which different aspects of (cyber)sickness such as dizziness, blurred vision, decreased concentration, headache etc. will be assessed [13]. By taking the difference between both, the influence of the immersive experience and its underlying parameters on the appearance of cybersickness can be assessed. In addition, the sensors embedded within the Head-Mounted Device (HMD) can be exploited to measure indications of cybersickness such as eye tracking (drop of attention) and tactile sensors (drop in pointing accuracy). Furthermore, a correlation analysis between the performance metrics and the occurrence of cybersickness can be performed [13].

Table II
OVERVIEW OF EXISTING , OBJECTIVE QUALITY METRICS AND MODELS

Authors	**Metric/Model**	**Sensory channel**
Widely used	PSNR, SSIM, VMAF	Visual (2D video)
Alexiou et al. [14]	PQR	Visual (point clouds)
Tran et al. [15]	WS-PSNR	Visual (360°video)
Van der Hooft et al. [16]	Gaze-driven model	Visual (360°video)
Narbutt et al. [17]	ViSQOL Audio	Audio (traditional)
Narbutt et al. [17]	ViSQOL Speech	Audio (speech)
Narbutt et al. [17]	AMBIQUAL	Audio (ambisonics)
Sakr et al. [18]	HPWPSNR	Haptics
Hassen et al. [19]	HSSIM	Haptics

IV. OBJECTIVE MODELLING: QOS-QOE MAPPING, BENCHMARKS AND REAL-TIME ASSESSMENT

Most studies within existing, scientific literature are limited to subjective studies on limited groups of users. Although such studies provide an accurate view on the end-user quality perception, they are rather limited in scalability and inefficient in terms of time and money [7]. In addition, multimedia systems benefit from real-time quality assessment to allow for dynamic adaptation of the system parameters to optimize end-user experience. Therefore, objective metrics are more tailored for this task [2]. The amount of research concerning overarching objective QoE metrics for multimodal experiences is limited, however.

More research has been conducted on each of the individual feeds of the system. Especially for traditional, 2D video applications a wide variety of objective metrics exist. Some of them are a pure mathematical comparison of signals (e.g. Peak Signal-to-Noise Ratio (PSNR)), while others take the Human Visual System (HVS) into account (e.g. Video Multimethod Assessment Fusion (VMAF) [20]). The latter are often based on Machine Learning (ML) techniques such as Support Vector Machines (SVMs) to combine multiple metrics. Limited attempts exist to expand this towards 360° and holographic content, such as the Point cloud Quality Rating (PQR) for point clouds of Alexiou et al. [14] or the Weighted to Spherically uniform PSNR (WS-PSNR) of Tran et al. [15] and the gaze driven model for adaptive tile-based streaming of van der Hooft et al. [16] for standard omnidirectional video content.

For auditory feeds, there also exist a handful of metrics such as ViSQOL Audio, ViSQOL Speech and AMBIQUAL by Narbutt et al. [17] for traditional audio, speech and ambisonics (i.e. a full sphere audio surrounding technique) respectively.

The haptic feed is the least explored path of the three senses. The limited amount of haptic-related, objective metrics

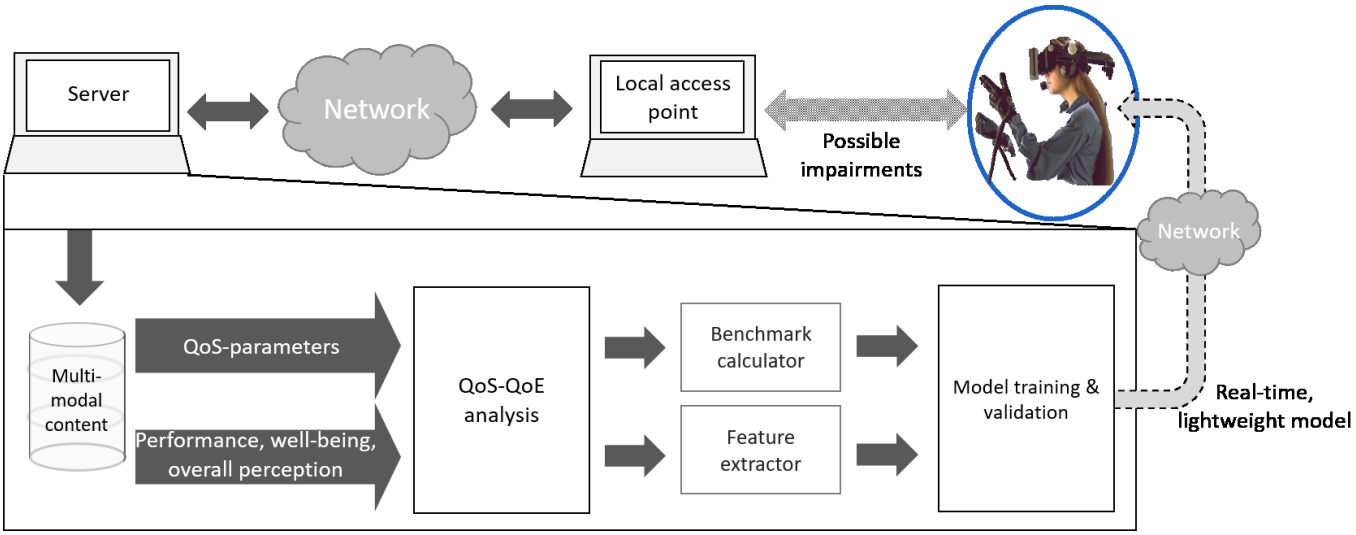

Figure 2. Schematic overview of the objective modelling approach

is based on generic metrics for evaluation of signal quality in general, e.g. Mean Squared Error (MSE) or Signal-to-Noise Ratio (SNR). Sakr et al.'s HPWPSNR [18] is probably the most known example of a haptic objective metric. Another, more recent one is the HSSIM by Hassen et al. [19], which is an adaptation of the classic Structural Similarity Index (SSIM) for video quality estimation to the specific case of haptic feedback signals.

An overview of these metrics and models is provided in Table II. These approaches are rather limited in amount and, more importantly, stay within the limits of one particular sensor channel. The influence of multiple feeds on each other remains unexplored [1]. Therefore, it is of great interest to investigate how deviations in synchronization between the signals affect the end-user's ability to perform a certain task. In addition, it should be researched to what extent the combination of quality degradation of the sensory channels enforces the sensibility to cybersickness. ML algorithms could provide a valuable tool to model these, probably complex, relationships [6]. To this extent, we propose to objectively model the user's *performance* and *well-being*, as well as an *Overall Perception Index* that combines the former two, optionally enhanced with subjective MOS as well. Starting from our previous research [21] in which we created both an objective benchmark and a lightweight, real-time quality model for passive Game Video Streaming (GVS), we envision three subsequent phases to this extent: QoS-QoE correlation analysis, objective benchmark creation for immersive applications and the design of real-time client-side quality models. Figure 2 illustrates the envisioned methodology.

The server-side correlation analysis will reveal the QoS-parameters of highest influence on both performance and well-being. In our previous work [21], Pearson Correlation Coefficient (PCC) and MSE were used to measure the relation with MOS. Additional metrics such as the Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC) and Mean Absolute Error (MAE) can of course be added to gain more insight in the data. Hereby, the data-driven and questionnaire-based perception assessment presented in section III will act as the benchmark in addition to (or even replacing) subjective quality scoring.

A second phase will aim at the establishment of highly accurate, objective metrics that mimic the subjective end-user perception without the actual intervention of a test subject. It is expected that a combination of highly complex per-feed quality metrics will reach the highest accuracy. Whether to apply early or late fusion to combine these metrics is a question open for research. ML modelling and analysis can provide valuable tools to this extent.

The output of these two phases will result in a set of lightweight features and objective benchmarks describing performance, well-being and overall perception. As such, a subsequent third phase will include the creation of lightweight quality models that relate the former to the latter. These trained models can then be included within the client device for real-time quality assessment. Once again, ML-algorithms such as Artificial Neural Networks (ANNs) and SVMs are put forward to fulfill this task.

V. Accurate and Real-time Viewport Prediction: Knowing the Users' Location Before They Move

In order to sustain the end-user perception it is of great importance to control the delivered quality and optimize it whenever possible. Within the light of immersive experiences, an important supporting mechanism for quality control is viewport prediction for omnidirectional visuals. Hereby, the goal is to proactively predict in which direction the user will be looking. As such,the major part of the available resources can be allocated to this particular part of the video hemisphere, resulting in the perceived quality being optimized. Two major types of viewport prediction can be identified: content-based [22] and content-agnostic [23] pre-

Table III
OVERVIEW OF EXISTING VIEWPORT PREDICTION APPROACHES

Authors	Type	Method
Heyse et al. [24]	Content-agnostic	Contextual Bandits
Petrangeli et al. [25]	Content-agnostic	Trajectory-based + clustering
Van der Hooft et al. [26]	Content-agnostic	Further refinement of [25]
Ban et al. [9]	Content-agnostic	LR + probability voting
Fan et al. [27]	Hybrid	Saliency analysis & motion detection + CNNs/LSTMs on historical viewport orientations
Jeong et al. [28]	Content-agnostic	360°audio location as predictor for next viewport

diction. Content-based prediction aims at characterizing the given content, independent from the user, in terms of saliency maps, motion detection, Regions-Of-Interest (ROIs) etc. Based on this information, density functions can be created that estimate the probability of a generic user looking at a particular part of the hemisphere at a given time instant [22]. Content-agnostic approaches, on the other hand, do not take the content into consideration, but try to predict the user's future fixation point based on this and other user's historical movement. These methods differ from rather straightforward approaches, e.g. static prediction (the user's orientation does not change between two consecutive samples) or 3D linear interpolation (the user's movement between the current and the next sample is the same as between the previous and the current) to more complex ML-based methods based on ANNs or SVMs [23]. An overview of these approaches is provided in Table III.

Heyse et al. [24], for example, present a content-agnostic Reinforcement Learning (RL) approach based on Contextual Bandits (CBs). To this extent, a set of previous HMD orientations is fed to two learners per spatial dimension (polar angle and azimuth). For each dimension, one learner estimates whether or not the user will actually move while the other predicts the actual movement. The results show significant improvements compared to the 3D linear interpolation approach.

Petrangeli et al. [25] propose a content-agnostic, trajectory-based prediction method. They exploit cross-user behaviour by clustering past user trajectories with similar movement. Per cluster, a single trend trajectory is computed that acts as a predictor for future user trajectories that fall within the same cluster. There results show significant improvements in comparison with static prediction and a LR method.

Van der Hooft et al. [26] further improve this solution by re-interpreting user movement as a trajectory on the sphere rather than on the equirectangular projection. In addition, the unidirectional extension of the current movement path of the user is limited to a fraction rather than the full trajectory to anticipate on volatile user movement. There results show a significant reduction of the prediction error.

Ban et al. [9] present a content-agnostic approach in which a LR is applied to predict the user's future fixation based on his/her past trajectory. Afterwards, this prediction is compared to the viewports of other users at the same time instant. A probability voting mechanism is implemented to adapt the prediction based on these observations. There approach shows an absolute improvement of about 20% in terms of viewport deviation in comparison with standard LR.

Fan et al. [27] developed a hybrid approach by combining saliency analysis and motion detection of the video content with the historical HMD viewport orientations. A combination of CNNs and LSTMs (which is a type of Recurrent Neural Networks (RNNs)) are applied to this extent. There results show similar prediction accuracy compared to other approaches in literature, but with the advantage of a reduction in both bandwidth consumption and buffer time. They further propose the inclusion of eye tracking sensors for more accurate performance.

Jeong et al. [28], at last, introduce an interesting research direction by utilizing the location information of the 360° audio feed as an indicator for the user's next viewport. Although their experiments are rather limited in nature, they tend to indicate that the addition of sound in the viewport prediction algorithm could prove to be an added value.

Within immersive applications, especially for non-entertainment purposes, there is a high level of entanglement between the particular content and the task to be fulfilled. In addition, preliminary results of own research show certain types of videos can be identified in which behavior is rather similar over all users and highly entangled to the content [10]. In some cases, however, the content itself seems to give little to no indication on the user's viewing behaviour [10]. In these cases, the viewport trajectory mainly depends on the type of user. As such, a hybrid (content-aware + content agnostic) approach is proposed to establish accurate viewport prediction. Furthermore, it is important to keep in mind that although model training can be allowed a certain level of complexity, its evaluation should be realizable in real-time fashion. This is a rather challenging task, given the rather varying behavior of users, especially when predictions need to be made a couple of seconds upfront to allow the server to prepare the content. Starting from this state of the art, we propose a hybrid solution to incorporate both the nature of the task at hand and the characteristics of the user performing it. The content-aware part of the proposed solution is envisioned as a combination of saliency detection combined with foreground/background extraction for motion identification. In addition, both the audio and haptic signals will be analyzed as triggers for user movement. As such, multimodal ROI can be defined overarching the three input feeds to define a general fixation probability distribution over the 360° sphere. Conditioned on these functions, users can be clustered depending on the ROIs that grab their attention and, more importantly, how they navigate between ROIs over time. Within each cluster, predictions can be further personalized using a content-agnostic approach on the viewport trajectories. CNNs on sliding trajectory windows

are proposed to this extent in order to exploit the temporal correlations in the data.

VI. QoE Optimization: From Quality Adaptation to Content Generation

Traditionally, quality optimization techniques are primarily incorporated for traditional 2D video. One of the most well-know and widely implemented mechanisms is *adaptive streaming*. Adaptive video streaming services deliver the content as a set of streams with different encoding parameters and thus, different qualities. The client device can dynamically change the representation depending on the perceived availability of the network. As such, stalling and rebuffering of videos is avoided and user QoE is maximized within the bandwidth limitations [6]. The current, internationally adopted standard for this purpose is Dynamic Adaptive Streaming over HTTP (DASH) [29].

Similar ideas are adopted for 360° video, although combined with accurate viewport prediction as discussed in the previous section. This often leads to the use of *tile-based* streaming mechanisms. Hereby, the 360° video is divided into temporal segments and spatial tiles. The client-device can request each tile at a different quality. As such, more resources can be allocated to the center of the viewport, *i.e.* the current looking direction of the user, such that QoE is maximized [26]. Note that this approach brings additional challenges to sustain QoE. Frequent quality switching of the video tiles in both the spatial and temporal dimension will heavily affect the user's feeling of immersion and can even lead to cybersickness. In addition, the quality adjustment should be realized according to the user's movement speed. Otherwise, the user will constantly spectate the quality change within his/her viewport, which is of course not beneficial for his/her perception of the content. At last, the eyes of the users are only rarely fixating on the exact center of the viewport [30]. As such, eye tracking is needed to identify the most critical tiles that should hold the highest quality [26].

While DASH provides a stable solution for presentational/unidirectional 2D video, it is less suited for the more interactive 360° tiled video streaming. The reason for this is that DASH is implemented on top of TCP, which causes additional delay due to its inherent data delivery guarantees. As such the application of DASH for 360° content delivery might result in sub-optimal streaming as requested tiles might arrive too late [31]. Therefore, 360° video streaming is often implemented on top of the *QUIC* transport protocol [32] designed by Google. It is an encrypted, multiplexed and low-latency transport protocol on top of UDP. The main advantage of QUIC over DASH is that it allows for urgent requests such that tiles with high priority can be quickly send to the client. As such, negative effects on the QoE by means of missing tiles can be mitigated. Therefore, it allows for a higher level of *interactivity* while more or less sustaining the inherent simplicity of DASH [31].

Next to the video signal, both the audio and haptic feed need to be delivered over the network as well. As these signals

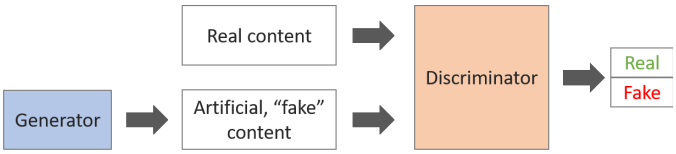

Figure 3. Schematic illustration of a GAN.

are less complex than the 360° video, however, most of the current, high-speed network infrastructures are reliable enough for almost perfect transmission. Most of the delay and packet loss typically arise within the local network, especially when wireless connections are used (which is often the case to enable maximal freedom of movement). To counteract this behaviour, similar quality optimization mechanisms as for the video case can be adopted.

If, however, situations still arise in which one of the sensory feed packages arrives late in comparison with the others, a backup mechanism is needed in order to control and minimize the influence on the QoE. This can be done by creating the expected content on the fly. Complex, but rather promising ML techniques exist in this direction. A rather recent, promising technique for this purpose are GANs [33]. GANs are CNNs consisting of a *generator* and a *discriminator*. The discriminative network's task is to learn the difference between "original" and artificially generated content (video frames or audio/haptic samples) or, more simply put, to distinguish between "real" and "fake". The generative network's task, on the other hand, is to artificially generate content such that it becomes indistinguishable from its real counterpart. In other words, the generator should learn to "fool" the discriminator. This will enforce the latter again to further increase its performance. As a result, the ongoing competition between both has the potential to deliver high quality results [33]. This is schematically illustrated in Figure 3. The generative properties of GANs can be applied to multimodal systems, as they enable to create multiple future samples of the sensorial feeds, conditioned on their current values, within the next few seconds of the experience. For the visual feed, this approach can further be optimized in combination with an accurate, long-term viewport prediction algorithm as it can predict upfront which part of the 360° sphere should actually be generated while the other tiles can be ignored [34].

As an extension to this, an automatic reasoning algorithm is needed to let the end-user system (e.g. the HMD) decide whether it is feasible to wait for the delayed content or that recreation should be applied. This algorithm should take current and historical measurements of the local network into account in order to calculate per-feed probabilities of the next sample being delayed [35]. Note that RL or supervised ML algorithms are one of the techniques that could be exploited to this extent [6].

VII. Conclusion

Given the stringent requirements of multimodal AR/VR systems, it is not enough to manage them in terms of QoS

parameters. Desynchronization and quality degradation need to be reduced to a minimum to avoid feelings of cybersickness or loss of immersiveness and concentration. Therefore, the quality management needs to shift from the system-centered performance metrics towards a more human, QoE-centered approach. However, this requires for novel techniques in the three areas of the QoE-management loop (monitoring, modelling and control). This position paper has pinpointed open areas of research to fully enable human-centric driven management of immersive multimedia. To this extent, four main dimensions have been identified, providing current state-of-the-art and future solutions. We believe this work provides the means to open new opportunities for research not only within the challenging AR/VR QoE arena, but also in the field of management and control of multimedia applications.

ACKNOWLEDGMENTS

Maria Torres Vega is funded by the Research Foundation Flanders (FWO), grant number 12W4819N.

This research is part of a collaborative project between Huawei and Ghent University, funded by Huawei Technologies, China.

REFERENCES

[1] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal of Selected Areas of Communication*, vol. 34, no. 3, pp. 460–473, 2016.

[2] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K. T. Chen, "A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 2, 2018.

[3] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, 2018.

[4] M. Eid, J. Cha, and A. El Saddik, "Admux: An adaptive multiplexer for haptic–audio–visual data communication," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 1, pp. 21–31, 2011.

[5] H. Yu, M. K. Afzal, Y. B. Zikria, A. Rachedi, and F. H. Fitzek, "Tactile internet: Technologies, test platforms, trials, and applications," *Future Generation Computer Systems*, vol. 106, pp. 685 – 688, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X20304192

[6] M. Torres Vega, C. Perra, F. De Turck, and A. Liotta, "A Review of Predictive Quality of Experience Management in Video Streaming Services," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, 2018.

[7] ITU-T, "Recommendation P.910 (09/99) ITU-T RECOMMENDATION P.910: Subjective video quality assessment methods for multimedia applications," 1999.

[8] C. D. Pham, H. N. T. Phan, and P. J. From, "Evaluation of Subjective and Objective Performance Metrics for Haptically Controlled Robotic Systems," *Modeling, Identification and Control*, vol. 35, no. 3, pp. 147–157, 2014.

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

[10] S. Van Damme, J. Heyse, F. De Backere, M. Torres Vega, and F. De Turck, "A Trajectory-based Analysis for Viewport Prediction in 360-Degree VR Video," 2020, Submitted to the *11th ACM Multimedia Systems Conference (MMSys) 2020*.

[11] N. S. and M. Minaus, *New Perspectives on Game-Based Assessment with Process Data and Physiological Signals*. Springer, 2019.

[12] K. P. Kruzan and A. S. Won, "Embodied well-being through two media technologies: Virtual reality and social media," *New Media & Society*, vol. 21, no. 8, pp. 1734–1749, 2019. [Online]. Available: https://doi.org/10.1177/1461444819829873

[13] V. Secinc and M. Berkman, "Psychometric evaluation of Simulator Sickness Questionnaire and its variants as a measure of cybersickness in consumer virtual environments," *Appl. Ergon.*, 2020.

[14] E. Alexiou and T. Ebrahimi, "On subjective and objective quality evaluation of point cloud geometry," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–3.

[15] H. T. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, and T. C. Thang, "An evaluation of quality metrics for 360 videos," in *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2017, pp. 7–11.

[16] J. van der Hooft, M. Torres Vega, S. Petrangeli, T. Wauters, and F. De Turck, "Quality assessment for adaptive virtual reality video streaming: A probabilistic approach on the user's gaze," in *2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, Feb 2019, pp. 19–24.

[17] M. Narbutt, A. Allen, J. Skoglund, M. Chinen, and A. Hines, "Ambiqual - a full reference objective quality metric for ambisonic spatial audio," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6.

[18] N. Sakr, N. D. Georganas, and J. Zhao, "A perceptual quality metric for haptic signals," in *2007 IEEE International Workshop on Haptic, Audio and Visual Environments and Games*, Oct 2007, pp. 27–32.

[19] R. Hassen and E. Steinbach, "Hssim: An objective haptic quality assessment measure for force-feedback signals," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6.

[20] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C. Wu, and C. . J. Kuo, "Challenges in cloud based ingest and encoding for high quality streaming media," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1732–1736.

[21] S. Van Damme, M. Torres Vega, J. Heyse, F. De Backere, and F. De Turck, "A Low-Complexity Psychometric Curve-fitting Approach for the Objective Quality Assessment of Streamed Game Videos," 2020, To appear in *Signal Processing: Image Communication*.

[22] C. Li, W. Zhang, Y. Liu, and Y. Wang, "Very long term field of view prediction for 360-degree video streaming," in *Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019*, ser. Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019. Institute of Electrical and Electronics Engineers Inc., 4 2019, pp. 297–302.

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

[24] J. Heyse, M. T. Vega, F. de Backere, and F. de Turck, "Contextual bandit learning-based viewport prediction for 360 video," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, March 2019, pp. 972–973.

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

[26] J. van der Hooft, M. Torres Vega, S. Petrangeli, T. Wauters, and F. de Turck, "Optimizing Adaptive Tile-Based Virtual Reality Video Streaming," in *Proceedings of the IFIP/IEEE International Symposium on Integrated Network Management 2019*, 2019.

[27] C. Fan, J. Lee, W. Lo, C. Huang, K. Chen, and C. Hsu, "Fixation Prediction for 360-Degree; Video Streaming in Head-Mounted Virtual Reality," in *Proceedings of the Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2017, pp. 67–72.

[28] E. Jeong, D. You, C. Hyun, B. Seo, N. Kim, D. H. Kim, and Y. H. Lee, "Viewport prediction method of 360 vr video using sound localization information," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2018, pp. 679–681.

[29] ISO/IEC 23009-1, "Information technology — Dynamic adaptive streaming over HTTP (DASH)," ISO/IEC, Tech. Rep., 2014.

[30] Y. Rai, P. Le Callet, and P. Guillotel, "Which Saliency Weighting for Omni Directional Image Quality Assessment?" in *Proceedings of the International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

[31] S.-C. Yen, C.-L. Fan, and C.-H. Hsu, "Streaming 360° videos to head-mounted virtual reality using dash over quic transport protocol," in *Proceedings of the 24th ACM Workshop on Packet Video*. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3304114.3325616

[32] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, I. Swett, J. Iyengar, and et al., "The quic transport protocol: Design and internet-scale deployment," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 183–196. [Online]. Available: https://doi.org/10.1145/3098822.3098842

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[34] M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3539–3548.

[35] B. Sliwa, T. Liebig, R. Falkenberg, J. Pillmann, and C. Wietfeld, "Efficient machine-type communication using multi-metric context-awareness for cars used as mobile sensors in upcoming 5g networks," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018, pp. 1–6.