VQEG eLetter

Volume 3, Issue 1 • November 2017

Contents[1]

London VQEG meeting, Oct. 2016

Storyboard

Jesús Gutiérrez, Patrick Le Callet, Phil Corriveau, Zhenzhong Chen, Editors

The Video Quality Experts Group (VQEG) was originally grounded in the development and verification of subjective methodologies and objective tools for video quality assessment. However, over the last 20 years from the formation of the VQEG, the multimedia ecosystem has dramatically changed and the VQEG has reacted to this progress by moving from the assessment of visual quality of video to the evaluation of Quality of Experience (QoE). In addition, VQEG has been evolving as multimedia technologies are marching forward, addressing them and supporting their research and development.

As an example of this evolution, the Immersive Media Group (IMG) was formed on March 2016 as the successor of the 3DTV Group to embrace the new emerging immersive technologies, such as virtual reality, augmented reality, omnidirectional content, free viewpoint navigation, and light field content. Thus, the objectives were redefined to baseline the QoE assessment of current immersive systems, such as providing guidelines for QoE evaluation, study appropriate methodologies for subjective assessment (e.g., considering presentation requirements, testing environments, factors to measure, etc.), support the development of objective metrics and visual models, and provide annotated datasets of emerging media content for those purposes.

Therefore, given the recent development of immersive media technologies, this new issue of the VQEG eLetter aims at providing an overview of ideas, developments, and research activities regarding immersive media (e.g., VR, AR, light field, 360 video, multiview technologies, etc.), especially illustrating the need for perceptual tools and assessment.

Issue Overview

This eLetter issue provides a collection of articles covering research activities in relation to QoE evaluation of emerging immersive media technologies, including the current hot topic of virtual reality and omnidirectional content, but also point cloud and light field technologies. We are proud to present seven contributions on the topic coming from leading researchers both from academia and industry.

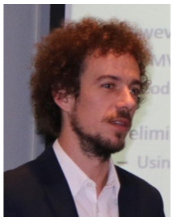
"An overview of developments and standardization activities in immersive media", by Dragorad Milovanovic and Dragan Kukolj, provide a summary of recent and current standardization activities in relation to QoE assessment for immersive media technologies.

"Measuring Virtual Reality Experiences is more than just Video Quality", by Hanan Alnizami, James Scovell, Jacqueline Ong, and Philip Corriveau, provides, from a holistic perspective, an overview of the different aspects influencing virtual reality experiences, including visual performance, ergonomics, audio and other ecosystem variables.

"Omnidirectional video communications: new challenges for the quality assessment community", by Francesca De Simone, Pascal Frossard, Chip Brown, Neil Birkbeck, and Balu Adsumilli, presents an overview of the typical omnidirectional processing chain, identifying the open challenges linked to quality assessment at each step of the chain.

"Anticipate the users' behavior for a deeper immersion", by Laura Toni and Thomas Maugey, shows how the user behavior is exploited in both bit allocation and streaming optimization strategies, and highlights the different interactive models that the two optimization problems require.

"On Streaming Services for Omnidirectional Video and its Subjective Assessment", by Igor D.D. Curcio, provides an introduction to the basic challenges in quality assessment of



__Jesús Gutiérrez__ is a post-doctoral researcher (Marie Curie/Prestige fellow) at the IPI group pf the LS2N of the Université de Nantes. His research interests are in the area of image and video processing, evaluation of multimedia quality of experience, and 3D and immersive media systems.

omnidirectional video streaming services, and proposes a new evaluation metric to measure the degree of watching pattern similarity of the participants in subjective tests.

"Subjective Video Quality Database for Virtual Reality", by Zhenzhong Chen and Yingxue Zhang, presents an annotated database of panoramic videos through a subjective rating test with virtual reality HMD, providing a reliable reference for benchmarking of objective metrics and insights on observers' psychophysical response to the VR contents.

"Quality Assessment Challenges in MPEG's Current and Future Immersive Media Standards", by Sebastian Schwarz and Sébastien Lasserre, describes some of several challenges to related to assessing the quality of point clouds, through the MPEG CfP on point cloud compression technologies.

"Perceptual analysis and characterization of light field content", by Jesús Gutiérrez, Pradip Paudyal, Marco Carli, Federica Battisti, and Patrick Le Callet, provides an overview of the light field processing chain from a perceptual perspective and propses a novel framework for light field content characterization for quality assessment.

***Patrick Le Callet** is a full Professor at Ecole polytechnique de l'Université de Nantes. He led for ten years (2006-16) the Image and Video Communication lab at CNRS IRCCyN. Since January 2017, he is one of the seven members of the steering Board the CNRS LS2N lab (450 researchers), as representative of Polytech Nantes. He is also since 2015 the scientific director of the cluster "Ouest Industries Créatives", a five-year program gathering more than 10 institutions. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing.*

***Philip J. Corriveau** is a Senior Principal Engineer and Director of End to End Competitive UX at Intel. He directs a team of human factors engineers conducting user experience research across Intel technologies, platforms and product lines. Philip is a founding member of and still participates in VQEG*

***Zhenzhong Chen** is a professor at Wuhan University. He leads the Institute of Intelligent Sensing and Computing conducting researches on multimedia technology, computer vision, image processing and understanding, data mining for geoinformatics, etc.*

An overview of developments and standardization activities in immersive media

Dragorad Milovanovic and Dragan Kukolj

Introduction

This article provides the recent and current activities related to the MPEG development and research activities of the emerging immersive media technologies. Especially we outline the need for perceptual tools in MPEG AhG *Immersive Media Quality Evaluation*[1] towards specification the new standard ISO/IEC 23090 *Coded Representation of Immersive Media*. Next, we provide ideas regarding quality assessment of immersive media (VR/AR, LF, 360-video) in QUALINET Task Force *Immersive Media Experiences* (IMEx) and VQEG *Immersive Media Group* (IMG).

Most of these standards activities are currently in early phase. An important aspect, not yet fully addressed is the *Quality of Experience* (QoE) of immersive applications and services.

Up-to-date standardization activities

In June 2016, MPEG started working on MPEG-VR initiative (currently MPEG-I *Collection of standards to digitally represent immersive media*) to develop a roadmap and coordinate the various activities related to VR within MPEG and to liaison also with other SDOs. Other consortia working on innovative products and services in this domain are 3GPP collaboration group of telecommunications associations (TR 26.918 *Virtual*

[1] https://lists.aau.at/mailman/listinfo/immersive-quality

Reality media services over 3GPP), DVB promoted a study mission to official group CM-VR (*Commercial Module on Virtual Reality*), and QUALINET (European network on quality of experience in multimedia systems and services) established Task force IMEx (*Immersive Media Experiences*).

The need for perceptual tools and assessment

Recently, MPEG established an *ad hoc* group *Immersive Media Quality Evaluation* with the goal to document requirements for VR QoE, collect test material, study existing methods for QoE assessment, study VR experience metrics and their measurability in VR services, and develop a test methodology.

In order to optimize the standard ISO/IEC 23090 (Part 2 *Immersive video*, Part 5 *Point cloud compression*) for the intended applications, MPEG-I[2] is calling for video test material to assess algorithm performance for different setups where information is combined from different cameras to generate virtual views scene (*Call for immersive visual test material*, April 2017). Different levels of experience are achieved by the user who may freely move his head around three rotational axes 3DoF (yaw, pitch, roll), and along three translational directions 6DoF (left/right, forward/backward, up/down). Test material should comply to the attributes as follow:

- General considerations. Still image and video sequences from both indoor and outdoor scenes can be submitted, with sufficient complexity to test the limits of the algorithms under study - natural content is highly preferred over computer-generated content. Color components, depth, and metadata are provided separately (particular for the camera parameters). Types of cameras and camera array arrangements (highly dense array of images along a

[2] http://mpeg.chiariglione.org/standards/mpeg-i

predefined track - 2D linear with parallel cameras, 2D linear with convergent cameras, 2D cylindrical surface, 2D spherical surface). Accurate temporal synchronization of multiple cameras is preferred.

- Omnidirectional video with depth data. The content should be captured with an arrangement of cameras that records divergent views, preferably in an arrangement that supports the capture of a full 360-degree field of view. Both the texture and depth data must be provided at the same resolution with an input greater than or equal to 4K, and the same projection - preferably in the equirectangular projection.

- Video material recorded by divergent/convergent camera arrangement with significant overlap preferably in an arrangement that supports the capture of a full 360-degree field of view / volume of visual data. Both the intrinsic and extrinsic camera parameters must also be provided.

- 2D camera array arrangement following a planar, cylindrical or spherical surface. Dense video sequences are particularly sought with a baseline distance between cameras not more than 20cm, and the distance from one end of the array to the other end as wide as possible.

- Plenoptic cameras with density of micro-lenses supposed to be large enough to ensure a good angular sampling of the light field. Resolution of the plenoptic image should be no less than 15 mega-rays.

- Systems of simultaneous multiple acquisitions shall simultaneously acquire the same scene following the specifications defined above.

Currently, the ITU-T started a new work program G.QoE-VR on parametric bitstream-based quality assessment (P.NATS promoted to P.1203). In this context Video Quality Experts Group (VQEG) has an *Immersive Media Group* (IMG) with the mission on quality assessment of immersive media, including virtual reality, augmented reality, stereoscopic 3DTV, and multiview. The initial goals for new established VQEG and QUALINET joint team on immersive media (JQVIM[3]) are collecting and producing open source immersive media content and data set, establishing and recommend best practices and guidelines, collecting and producing open source immersive media tools, and survey of standardization activities.

Dragorad Milovanovic *received the Dipl. Electr. Eng. and M.Sc. degree from Faculty of Electrical Engineering, University of Belgrade, Serbia. He has working as research assistant in DSP, R&D engineer in multimedia communications and ICT lecturer.*

Dr. Dragan Kukolj *is a Professor of computer-based systems with Dept. of Computer Engineering, Faculty of Engineering, University of Novi Sad, Serbia.*

[3] https://www3.informatik.uni-wuerzburg.de/qoewiki/qualinet:imex:jqvim

Measuring Virtual Reality Experiences is more than just Video Quality

Hanan Alnizami, James Scovell, Jacqueline Ong, Philip Corriveau

What is VR

Virtual Reality (VR) allows the user to experience a completely digitized environment while attempting to disconnect the user from her/his real world. In his book, Virtual Reality, Howard Rheingold defines it as an experience in which a person is surrounded by a three-dimensional computer-generated representation, and is able to move around in the virtual world and see it from different angles, to reach into it, grab it, and reshape it [1]. While Virtual Reality has drawn much attention and publicity the past few years, it is not a new concept in technology. VR dates back to the 1960s when Ivan Sutherland pioneered the first head-mounted display at MIT [2], which was then a room-size V.R. machine, with an helmet so heavy that it had to be supported by a mechanical arm suspended from the ceiling [3]. Soon after, HMDs were adopted for military applications [4, 5]. Then on, the US Navy, the US Army, and NASA all invested in VR in hopes of building flight and combat simulators. The US Army deployed the Integrated Helmet and Display Sighting System (IHADSS) on the AH-64 Apache helicopter. Despite the monocular display, the IHADSS greatly contributed to the proliferation of all types of HMDs [6].

While Virtual Reality has drawn much attention and publicity the past few years, it is not a new concept in technology.

Since then, VR has expanded to various applications including automotive, medicine, education, and architecture [7, 8], offering invaluable information-sharing experiences across many applications such as gaming, entertainment, education, and commerce. This new and innovative way of interaction has enabled users to unique experiences such as telepresence [7, 9], and high interactivity [10], especially in virtual commerce experiences from the comfort of one's home.

The explosion of devices available for consumer consumption has been incredible and varies in the quality of implementation for a range of budgets. Regardless of which segment one is aimed at creating immersive experience in VR is not an easy endeavor, and assessing VR interactions holistically is a demanding and complex procedure. VR experiences represent a constellation of engineering metrics which, while can be challenging to simply evaluate independently, they interact together to make or break an experience. It is vital to have a good understanding of Holistic VR experiences and know how to properly assess them. Failure to provide well-tuned virtual content could cause undesired physiological implications on the user.

Virtual Reality Engineering/User Metrics

VR experiences are driven by multiple types of metrics such as visual performance, auditory cues, user interaction, and ergonomics. While a plethora of VR literature has been deployed on enhance computer graphics, display technologies, and input tracking among others, little to no literature has been found that focuses on VR experience usability evaluation and overall ergonomics.

Head Mounted Display Ergonomics

This is an essential pillar for providing an immersive VR experience. One of the significant limitations for these HMDs is their substantial weight, due to the attached computing and display hardware. HMD weight can markedly affect head balance, body posture and locomotion, which in turn can retard voluntary motion and action in response to visual stimuli. A heavier device weight can also increase the mismatch between visual, vestibular and proprioceptive cues, leading to motion sickness symptoms. Not only is weight important, but the distribution of weight around the HMD could play a role in users range of motion and overall experience, see Figure 1.

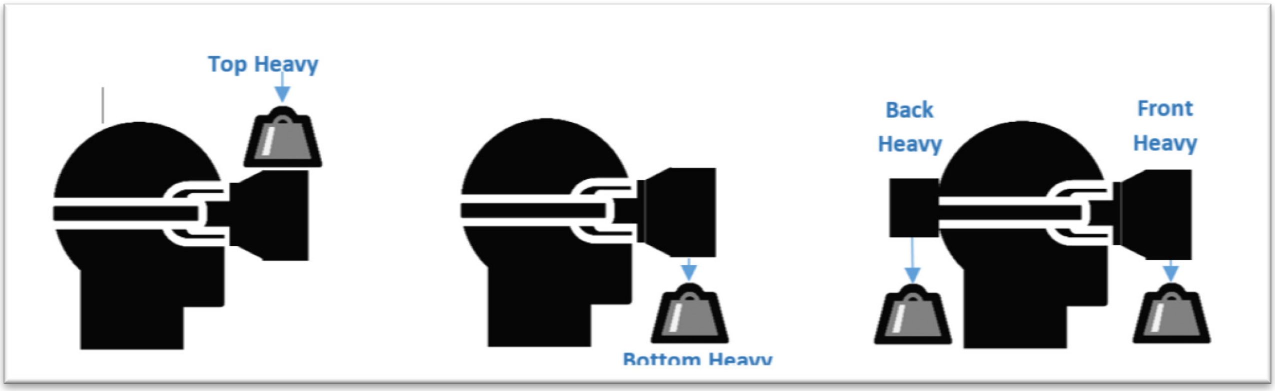


Figure 1. HMD weight and weight distributions can affect overall muscular and physical stress

HMD temperature has become an evident limitation that has been increasingly self-reported by users, especially gamers, who engage in lengthy VR sessions. Typically, HMD manufacturers seek HMD designs that offer good seal around the user's face in order to prevent light leakage that could produce blur and glare on the HMD screen, affecting user's experience. Latter design decisions and poor ventilation solutions within the HMD have caused the relatively small amount of air trapped within the face cup to increase in temperature and humidity.

Whether it is weight, weight balance, pressure, fit and finish, ambient temperature of the HMD's face cup, or overall hygiene, failure to provide good user design will guarantee breaking the VR experience.

Well-designed controller ergonomics are also crucial to allow for comfortable interaction with the virtual world. Controller weight, button design, finish, and hygiene are some of many ergonomic aspects that should be considered.

VR Visual Performance

An area where the Video Quality Experts Group has excelled for many applications and usages, is another critical component in providing a truly immersive experience. Negative experiences such as dropped frames or tearing can disrupt the user interactions breaking the perceived reality. Visual experiences afforded by HMDs are the best when the rendered visual images closely match other sensory inputs, such as motion, balance and proprioceptive feedback. This is especially important for situations where the user is moving around and actively interacting with the visual world.

There are many visual factors that can impact this aspect of VR such as resolution, refresh rate, flicker, field of view, pixels per degree, etc. While each individual variable can be isolated and evaluated to understand acceptable experience thresholds, to truly understand the overall visual experience eventually the variables must be combined to understand potential interactions. Unfortunately, due to the infantile nature of VR we must begin by isolating individual variables for evaluation.

There are some variables like resolution that can be manipulated using the developer settings built-in to some systems. Otherwise a test harness would need to be developed to manipulate each variable. Due to the “black box” nature of various VR systems it can be challenging to do comparisons between systems to isolate a single variable, which is another benefit of developing specific test harnesses that allow researchers to manipulate one variable at a time.

Refresh rate is another variable that can significantly impact the perceived visual performance. Some VR systems utilize a methodology called decoupled refresh rate, where content movement will update at a lower rate than the user head movement allowing for a lower compute cost thus reducing the potential for performance based issues. Though there is little information available on decoupled refresh rate to understand the potential impact to the user.

Common issues caused by impaired system performance include dropped frames or tearing. In an effort to combat this, system developers have developed various methodologies to alleviate these issues. These methodologies include blending of frames, projecting new frames based on user movement and the last rendered frame. These methods will help reduce performance related issues but there is a lack of understanding as to what magnitude of degradation would be too much to handle and what negative impact these methodologies may have. These attempts to combat system performance based issues is a benefit for the end user but an extra challenge for researchers attempting to evaluate various solutions.

While there are many variations, there are generally three ways to evaluating user perception for variables such as frame drops, tearing, refresh rate and others. First is to find natural variation between devices or systems. In the case of refresh rate for example, this is achieved by identifying: 1) various refresh rates for devices to be tested (i.e. 30, 60, and 90Hz); 2) an application that works across devices; 3) representative use cases that would exacerbate the impact of refresh rate. When running the study, participants are tasked with completing tasks and providing subjective experience ratings to understand if any differences exists among conditions. There are some potential issues with this methodology, such as finding natural variation. This methodology also assumes that all other variables are held constant as to prevent confounding factors, which is often unlikely.

The second way is developing a test harness that would allow for test variable manipulation while holding all other variables constant. The trick with developing test harnesses is manipulating variables in a way in which the variable is representative of how the variable would naturally occur. It is important to keep in mind that developing a test harnesses can be a time consuming and costly endeavor.

The third way to evaluating user perception is via expert assessments. Using experts in a study can be especially controversial as the sample size is extremely limited and can often be biased by their experience and knowledge in the space. Having unbiased and industry recognized evaluators becomes crucial if the data is to be accepted.

VR Audio

Audio is a component of VR that can easily be overlooked but is critical to many of the experiences that are of interest to most users. In VR, senses are muted to the real world and as such, the brain relies on stimuli presented through the display and speakers or headsets to accurately and comfortably orient the user in space. 3D positional audio for example has become critical not only for providing an immersive VR experience but for orienting the user in their virtual space. It is no longer good enough to know that something is going outside of your field of view. If a zombie is coming from behind you to your left, you need to have audio that can accurately convey this to the user.

Other Ecosystem Variables

There are other aspects of system performance that can impact the experience including latency and accuracy. These hold true for both the HMD and any controllers that may be used. Latency is a fairly straight forward concept to evaluate but is critical to ensure a positive experience. Excessive latency can cause a variety of issues from minor annoyance to extreme nausea. Accuracy includes a variety of variables that can all impact the perception of perceived reality. It is not as simple as a point and shoot. HMD's and controllers need to be evaluated for directional accuracy, drift, static noise, and scaling errors. Minor variation in any one of these variables may not break the experience but a combination or excess of any one variable can have significant negative effects.

Evaluation Methodologies

Traditional methodologies frequently used by VQEG such as MOS were able to be applied to evaluate many of the variables discussed above such as resolution, refresh rate, thermals, frame drops, tearing and more.

While these base methodologies were applied for some variables, we have been forced to develop adapted methodologies, hardware and software solutions. For example, to understand the thermal impact of the HMD on a user, thermal and humidity sensors were attached to an Arduino board that was fixed to the HMD allowing for continuous real time data collection. This allowed for data to be easily mapped to user rating. The initial research focused on passive experiences but it will be important to also understand the thermal impact to a user during more strenuous experiences. A challenge for thermal testing was to accurately collect relative humidity data in a confined space such as the face cup of the HMD.

When evaluating various bit rate encodings, it is challenging to allow for real time application of the encoding method. Content has to be prerecorded for use in a study. The latter causes an issue if the user moves their head as content will not visually update by their movement. As a result, users are instructed not keep their head in a fixed position to avoid inducing extreme nausea.

Best Practices

When recruiting users to participate it is important to ask the proper screening question to ensure a safe and positive study session. General questions that con preclude users from participating include a history of susceptibility to motion sickness, heart conditions, and eyewear that could interfere with wearing an HMD.

When switching between test harness settings or applications it can be beneficial to ask users to close their eyes until the new content is available as the transition can be disorienting. As the users are immersed in the HMD and cannot see the outside world without removing the HMD every time, asking participants to speak aloud and have a proper training sessions with the rating scale can be extremely beneficial. It was not uncommon for users to forget the rating scale and needed to be reminded, so it is important to ensure the user understands the rating scale each time they give a rating.

If participants do report any eyestrain or nausea, it is important to have water available and ensure they do not immediately drive. Lastly, it is very important to frequently check in with users to ensure they are not suffering from any side effects and that they have the ability to quit at any time.

Conclusion

Visual quality is not enough, VR is much bigger than just the visual experience. The inability to present the user with accurate and consistent stimuli from any of the variables discussed above whether visual, auditory, interactive, or ergonomic, could affect users' neurophysiological responses and physical comfort such as dizziness, nausea, eye strain, and overall muscular and skeletal fatigue. Excessive weight can impede and limit behavioral responses and induce viewing discomfort due to poor HMD fit and will induce undesirable visual, muscular and cognitive symptoms.

Bibliography

[1] H. Rheingold, "Virtual reality: Exploring the brave new technologies", Simon & Schuster Adult Publishing Group, 1991.

[2] I. E. Sutherland, "A head-mounted three dimensional display", *Proc. of the AFIPS Fall Joint Computer Conference, part I*, pp. 757–764, San Francisco, USA, Dec. 1968.

[3] V. Heffernan, "Virtual reality fails its way to success", *The New York Times Magazine,* Nov. 2014. [Online]. Available: https://www.nytimes.com/2014/11/16/magazine/virtual-reality-fails-its-way-to-success.html?_r=0

[4] J. E. Melzer and K. Moffitt, "HMD design--putting the user first", in Head-mounted displays: Designing for the user, McGraw-Hill, pp. 1–16, 1997.

[5] C. E. Rash, "Helmet mounted displays: Design issues for rotary-wing aircraft", vol. 93. SPIE Press, 1999.

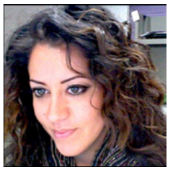
[6] C. E. Rash and J. S. Martin, "The impact of the US Army's AH-64 helmet mounted display on future aviation helmet design", 1988.

[7] F. P. Brooks, "What's real about virtual reality?", *IEEE Computer Graphics and Applications*, vol. 19, no. 6, pp. 16–27, Nov-Dec. 1999.

[8] C. Demiralp, C. D. Jackson, D. B. Karelitz, S. Zhang, and D. H. Laidlaw, "Cave and fishtank virtual-reality displays: A qualitative and quantitative comparison", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 3, pp. 323–330, May-Jun. 2006.

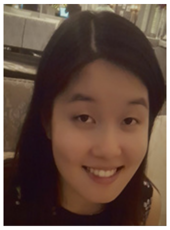
[9] M. Y. Hyun and R. M. O'Keefe, "Virtual destination image: Testing a telepresence model", *Journal of Business Research*, vol. 65, no. 1, pp. 29–35, Jan. 2012.

[10] A. Mollen and H. Wilson, "Engagement, telepresence and interactivity in online consumer experience: Reconciling scholastic and managerial perspectives", *Journal of Business Research*, vol. 63, no. 9–10, pp. 919–925, Sep-Oct. 2010.



__Hanan Alnizami__ is a Human Factors Engineer at Intel. She specializes in setting UX influenced engineering requirements that drive software, hardware, and platform specifications. She received her PhD in Human-Centered Computing from Clemson University in 2015.

__James Scovell__ is a Human Factors Engineer working at Intel. He specializes in experimental design and predictive model development in areas such as system performance, visual quality, and human computer interaction.



__Jacqueline Ong__ has been a user experience software engineer for two years at Intel, driving competitive analysis for virtual reality and machine learning use cases with Intel solutions.

__Philip J. Corriveau__ is a Senior Principal Engineer and Director of End to End Competitive UX at Intel. He directs a team of human factors engineers conducting user experience research across Intel technologies, platforms and product lines. Philip is a founding member of and still participates in VQEG.

Omnidirectional video communications: new challenges for the quality assessment community

Francesca De Simone, Pascal Frossard, Chip Brown, Neil Birkbeck, Balu Adsumilli

Introduction

Fully omnidirectional cameras, able to instantaneously capture the 360° surrounding real world scene, have recently started to appear as commercial products and professional tools. While the popularity of 360° content and applications using such content is rapidly increasing, many technical challenges at different steps of the omnidirectional signal acquisition, processing and distribution chain still remain open. In order to design perceptually-optimised omnidirectional visual communications, the availability of tools to quantify the level of distortion introduced by each processing step, and, ultimately, the overall quality of the processed signal and the 360° experience is critical. With respect to classical image and video signals captured by perspective cameras, the omnidirectional imaging pipeline has some peculiarities, which are related to the *spherical content capture*, the *signal representation*, and the *interactive and immersive nature of content rendering*. A deep understanding of each step of the imaging pipeline is key to design tools able to quantify the quality of 360° signals and the immersive experience. In this letter, we aim at providing an overview of the typical omnidirectional communication chain, identifying the open challenges linked to quality assessment at each step of the chain. A brief review of the existing tools proposed and used in the state of the art to assess the quality of omnidirectional signals, as well as perspective on future research directions, are also presented.

Pipeline and distortions

Content capture

State of the art omnidirectional cameras are mainly multi-dioptric systems, i.e., sets of cameras with fish-eye lenses, and have a global field of view of 360°. Such systems can be modelled as central cameras that project a point in the 3D space to a point on a spherical imaging surface, i.e., the *viewing sphere* [1]. Thus, an omnidirectional image can be considered as a signal lying on a sphere. In practice, the image is the result of a *mosaicking* (i.e., *stitching*) algorithm that merges the signals acquired by the dioptric cameras [2]. Distortions may be introduced by the optics of each dioptric camera (*optical distortions*, example in Figure 1), as well as by the stitching itself (*stitching discontinuities* or *seams*). If the optical distortions are not consistently corrected, they may affect the quality of the stitching [3]. The stitching discontinuities can appear across objects' edges (Figure 2) and as color and brightness discontinuities across different portions of the sphere. For video recordings, an inaccurate synchronization of the dioptric cameras can also result in motion discontinuities [2].

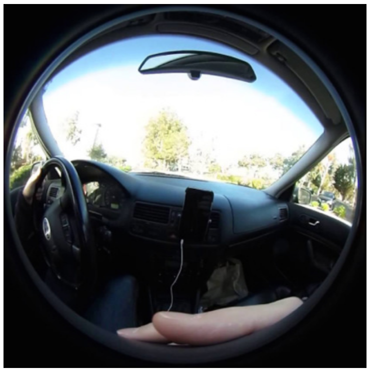


Figure 1. Example of optical distortions: image captured with fish-eye lens.

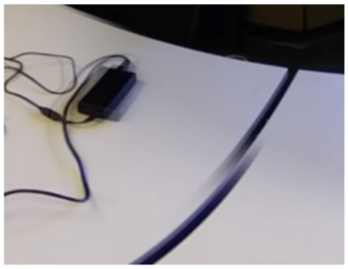


Figure 2. Example of stitching discontinuity on portion of equirectangular image.

Signal representation

An image captured by a 360° camera is usually stored as a rectangular array of samples called a *panoramic image* (i.e., *panorama*). The panorama results from the projection of the sphere to a plane (*map projection* [4] or *spherical parametrization* [5]). This data representation allows re-use of standard file formats and processing pipelines for signals defined on a plane but inevitably modifies the characteristics of the visual signal. Different parametrizations (two examples in Figure 3) correspond to different distortions of lengths, angles, and areas (*warping distortions*) [5] and may introduce *discontinuities*. The panoramic signal, affected by these distortions, is not directly presented to the end-user: the inverse map projection, mapping

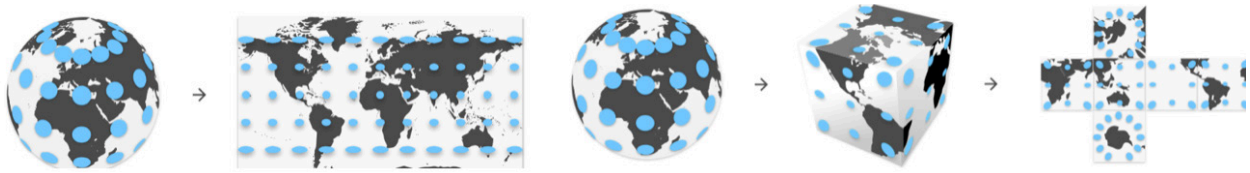


Figure 3. Example of map projections (equirectangular on the left, cube map on the right) and related warping distortions: blue circles on the spherical surface are mapped to ellipses with varying axis lengths on the plane.

the signal from the plane to the sphere, is applied to render the signal, as described in the next subsection. However, the projection of the visual signal to a plane and its inverse projection to the sphere for rendering in the final application, imply signal re-sampling and interpolation. Thus, different map projections may result in *aliasing, blur* and *ringing distortions* in the signal visualized by the end-user [6].

Rendering

When the 360° content is rendered to be viewed by an end-user, for example via a Head Mounted Display (HMD), a portion of the sphere surface is projected to a planar segment tangent to it, called the *viewport* (Figure 4). A viewport is defined by the viewing direction that identifies the point where the viewport is tangent to the sphere, its resolution, and its horizontal and vertical field of view. The image displayed on the viewport, i.e., the display of a HMD, is a regular lattice. The viewport extraction in a HMD is usually performed by OpenGL [7]: the panoramic image is used as a texture for a mesh-based representation of the viewing sphere. The projection of points from the sphere surface to a plane tangent to it at any point is an azimuthal projection, known as *oblique gnomonic projection* [4]. The projection may involve interpolation in order to generate a regular lattice: depending on the resolution of the viewport and the resolution of the spherical image from which the viewport is derived, *aliasing, blur* and *ringing distortions* can occur in the signal visualized by the end-user. Since the spherical image is usually stored in its panoramic representation, the distortions due to the viewport extraction add up to those due to the sphere-plane-sphere projections.

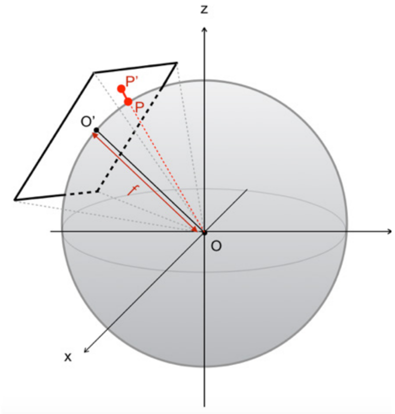


Figure 4. Geometry of viewport: the oblique gnomonic projection projects point *P* on the viewing sphere to point *P'* in the plane at focal length *f* from the center of projection *O*, defined by the viewport.

Processing: Encoding & Streaming

The panoramic representation is used nowadays as an intermediate format for processing 360° images and videos, for example to encode and stream them. New kinds of distortions can occur in the signal presented to the user, due to the fact that the processing is typically done in the planar domain after map projection but the signal is projected back to the sphere and to the viewport, for rendering. We will refer to these distortions as *processing distortions*. As already mentioned, different parametrizations may introduce different warping distortions and discontinuities. Thus, the processing distortions are expected to be dependent on the parametrization.

Figure 5. Example of zoom on radial blocking pattern in a portion of viewport corresponding to a portion of equirectangular frame affected by classical blocking artifacts due to block-based lossy compression.

Lossy compression is a good example of such processing: planar panoramic signals can undergo classical block-based transform coding, as proposed for example in [8-10]. Lossy compression may introduce typical *coding distortions* [11], such as blocking, banding, and ringing artifacts, in the compressed panorama. When the panorama is mapped back to the sphere and the viewport is rendered to the user, these distortions are modified due to warping and interpolation. Figure 5 shows an example of *radial blocking pattern* appearing in a portion of viewport extracted from a compressed equirectangular frame, affected by classical blocking artifacts. Additionally, the presence of discontinuities in the panoramic signal given as input to the encoder can result in visible *seams* in the viewport extracted from the decoded panoramic signal, which might break the sense of immersion (Figure 6). Tile-based coding solutions have also been proposed to encode panoramic frames, dividing it in independently decodable portions [12, 13]: depending on the coding parameters used for each tile, tiling might result in *discontinuities* on the panoramic frame and within the viewports [13].

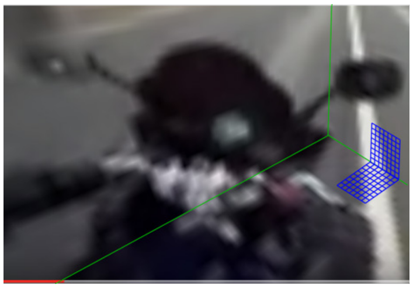


Figure 6. Example of viewport extracted for compressed cube-map panorama: discontinuities between the cube faces in the panoramic arrangement (highlighted by green lines) may cause wireframe cube and underlying image sampling domain (highlighted by blue grids) to be visible.

Due to the high resolution needed to assure a truly immersive experience, streaming omnidirectional content implies new challenges related to optimization of bandwidth consumption and maximization of user's Quality of Experience (QoE). Tile-based encoding solutions can be used to perform *viewport-adaptive streaming* of 360° content where only the portion of the sphere that is most likely to be visualized by the viewer, at a certain instant in time, is transmitted to the client at high quality [14, 15]. If the predicted viewing patterns do not match the user's actual navigation, these streaming strategies may be affected by spatial and temporal *quality fluctuations within the user's viewport*, i.e., the viewport includes tiles encoded at different quality, at a certain instant in time or over time.

Existing tools and open challenges

The availability of quality assessment tools to reliably compare different stitching algorithms, map projections and coding methods, or quantify the overall user's QoE during 360° content navigation, is becoming critical nowadays. With respect to classical visual quality assessment, the spherical geometry of the signal, the sense of immersion and the interactivity, and their user-, application- and content-dependency, represent major novelty factors. Possible differences in perception mechanisms and visual sensitivity, when rendering of visual signals is done using HMDs, are also interesting topics for research.

Objective quality assessment of 360° visual signals

How are these factors taken into account by state of the art algorithmic solutions to assess the quality of 360° signals? Until now, the proposed objective metrics for measuring spherical image quality are simple adaptations of existing full-reference error or quality metrics, in one of a few ways:

- by measuring the pixel error at a discretely sampled set of points on the sphere (Spherical-PSNR [16]);
- by weighting the pixel error by the corresponding pixel area on the spherical surface (Weighted-PSNR [17]);
- by measuring the pixel error on a planar representation of the signal where warping distortions are less prominent, for example obtained via the Craster Parabolic Projection (CPP-PSNR [17]);

- by rendering into viewports and measuring image or video quality in the viewport [18].

These solutions have obvious limitations:

1) They rely on the ability of existing planar image error or quality metrics to correctly detect and quantify distortions that are relatively novel, as discussed in the previous section illustrating the omnidirectional processing chain. Such ability remains to be verified. Additionally, some artifacts have dramatic consequences on the sense of immersion of a 360° navigation experience: none of the adapted objective metrics is designed to discriminate such artifacts.

2) Questions can be raised on the correct parametrization of these metrics and their sensitivity to the way the reference signal has been produced. For Spherical-PSNR, the uniform sampling method of the spherical surface, the interpolation, and the number of samples to be used are not strictly defined and different choices may lead to different results. Warping to a common domain has the limitation that it is biased towards projections "closer" to the chosen common projection domain. In practice, the raw signal from the cameras has been provided in a projection type (like equirectangular or cube map), which has already been re-sampled and stitched during the acquisition phase. So there is an issue with the definition of “ground-truth” used as the reference signal. Of all of these, measuring quality in rendered viewports might appear a more robust solution, as it measures what a user actually sees and this is not biased towards a particular projection type. Nevertheless, this solution provides results that might be difficult to interpret, due to the dependency from the viewing direction at which the viewport is extracted [10].

3) Being full-reference solutions, they cannot be used to automatically compare the performance of stitching algorithms or map projections, when the reference signal is not defined.

4) Last but not least, overall, a formal validation of the proposed solutions with respect to subjective ground-truth data is missing or limited to specific distortions, such as compression artifacts [19].

Conclusions and perspective

In this letter, we have provided a brief overview of the main processing steps that omnidirectional signals undergo till being visualized by the end-user, highlighting the kinds of distortions that may affect the visual quality of the signal and the overall 360° experience. Our goal was to convince the reader about the fact that the perceptual optimization of the omnidirectional communication pipeline opens new exciting research challenges concerning the design of new tools to reliably quantify the quality of omnidirectional signals and of the 360° user experience. We believe that the design of such tools requires a deep understanding of the underlying processing chain, as well as of the subjective implications of the types of artifacts occurring in omnidirectional images and videos. Finally, it is important to mention that we limited our review to monoscopic omnidirectional imaging, but many research challenges for the quality assessment community are also open concerning the stereoscopic acquisition, processing and rendering of omnidirectional images and videos.

References

[1] B. Micusik, "Two View Geometry of Omnidirectional Cameras", PhD Thesis, Center for Machine Perception, Czech Technical University in Prague, 2004.

[2] P. Baheti, "Virtual reality content creation technology", Qualcomm White Paper, 2017.

[3] A. Frich, "The guide to panoramic photography", http://www.panoramic-photo-guide.com/virtual-tour-360-photography/optical-distortions-virtual-tour.html

[4] F. Pearson, "Map Projections: Theory and Applications", CRC Press, 1990.

[5] K. Hormann, B. Lévy, and A. Sheffer, "Mesh parameterization: theory and practice", *ACM SIGGRAPH*, Course notes, Aug. 2007.

[6] C. Brown, "Bringing pixels front and center in VR video", https://www.blog.google/products/google-vr/bringing-pixels-front-and-center-vr-video/

[7] OpenGL, https://www.opengl.org

Francesca De Simone *received the M.Sc. degree in electronics engineering from Università degli Studi Roma Tre, Italy, in 2006, and the Ph.D. degree in computer and information science from the Swiss Federal Institute of Technology (EPFL), Switzerland in 2012. Between 2012 and 2014, she was post-doctoral fellow in the Multimedia Signal Processing Group at Institut Mines Telecom ParisTech, France. In 2015, she worked as senior engineer, video streaming expert, in the cybersecurity department of Kudelski Security, Switzerland. Since November 2015, she is back at EPFL, as post-doctoral fellow in the Signal Processing Laboratory led by Prof. Pascal Frossard. Her research interests include subjective and objective multimedia quality assessment, image and video compression, and multimedia streaming strategies.*

Pascal Frossard *received the M.S. and Ph.D. degrees, both in electrical engineering, from the Swiss Federal Institute of Technology (EPFL), Switzerland, in 1997 and 2000, respectively. Between 2001 and 2003, he was a member of the research staff at the IBM T.J. Watson Research Center, New York. He is now an associate professor at EPFL, where he heads the Signal Processing Laboratory (LTS4). His research interests include image representation and coding, visual information analysis, distributed image processing and communications, and media streaming systems. He is currently a Senior Member of the IEEE, a member of the IEEE MMSP TC, and the past chair of the IEEE IVMSP TC.*

[8] C. Grunheit, A. Smolic, and T. Wiegand, "Efficient representation and interactive streaming of high-resolution panoramic views", *Proc. of International Conference on Image Processing (ICIP)*, Rochester, USA, Sep. 2002.

[9] I. Bauermann, M. Mielke, and E. Steinbach, "H.264 based coding of omnidirectional video", *Proc. of International Conference on Computer Vision and Graphics (ICCVG)*, pp. 209-215, Warsaw, Poland, Sep. 2004.

[10] F. De Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding", *Proc. of IEEE Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016.

[11] H.R. Wu and K.R. Rao, "Digital Video Image Quality and Perceptual Coding", CRC Press, 2005.

[12] Y. Sanchez de la Fuente, R. Skupin, and T. Schierl, "Compressed domain video processing for tile based panoramic streaming using SHVC", *Proc of ACM Int. Workshop on Immersive Media Experiences (ImmersiveMe 2015)*, Brisbane, Australia, Oct. 2015.

[13] M. Yu, H. Lakshman, and B. Girod, "Content adaptive representations of omnidirectional videos for cinematic virtual reality", *Proc of ACM Int. Workshop on Immersive Media Experiences (ImmersiveMe 2015)*, Brisbane, Australia, Oct. 2015.

[14] A. Zare, A. Aminlou, M. Hannuksela, and M. Gabbouj, "HEVC-compliant tile-based streaming of panoramic video for virtual reality applications", *Proc. of the ACM Multimedia Conference (ACM MM)*, Amsterdam, Netherlands, Oct. 2016.

[15] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-Adaptive Navigable 360-Degree Video Delivery", *Proc. of IEEE Conference on Communications*, Paris, France, May 2017.

[16] M. Yu, H. Lakshman and B. Girod, "A Framework to Evaluate Omnidirectional Video Coding Scheme", *Proc. of IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Fukuoka, Japan, Oct. 2015.

[17] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video", *Proc. of SPIE*, vol. 9970, Sep. 2016.

[18] C. Brown, N. Birkbeck, and R. Suderman, "Quantitative Evaluation of Omnidirectional Video Quality", *Proc. of International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, Jun. 2017.

[19] E. Upenik, M. Rerabek and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content *Proc. of International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, Jun. 2017.

__Chip Brown__ received a BS in Physics from the University of Illinois, Champaign-Urbana. He attended the doctoral mathematics program at the University of California, Berkeley before getting pulled a little bit south into Silicon Valley. Chip worked for many years in Adobe's core technology group, focused on graphics rendering for the suite of Adobe products. In 2010 he left Adobe and went on to a stint at Google, Technical Director at Electronic Arts, and CTO of a couple of startups. He rejoined Google in 2015 to work on Omnidirectional video technology.

__Neil Birkbeck__ received his M.Sc and Ph.D degrees from the University of Alberta in 2005 and 2011 respectively. After a position as a research scientist at Siemens Corporate Research, he joined YouTube/Google in 2013 as an engineer working on video processing aspects of 360/VR/Omnidirectional and HDR video.

__Balu Adsumilli__ did his masters in University of Wisconsin in 2002 and his PhD at University of California in 2005, on watermark-based error resilience in video communications. From 2005 to 2011, he was Sr. Research Scientist at Citrix Online, and from 2011-2016, he was Sr. Manager Advanced Software at GoPro, at both places developing algorithms for images/video enhancement, compression, and transmission. He is currently leading the Media Algorithms team at YouTube/Google. He is an active member of IEEE, ACM, SPIE, and VES, and has co-authored more than 80 papers and patents. His fields of research include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas.

Anticipate the users' behavior for a deeper immersion

Laura Toni and Thomas Maugey

Introduction

Immersive media technology aims at endowing any final user with an unprecendent sense of full immersion in virtual (or real-world) environments. This is possible by projecting the user at the center of the 3D scene, which dynamically changes with the user interaction. This interactivity is driven by the headset mounted devices (HMD) in Virtual Reality, or by the user smartphone in Augmented Reality, or by tablet or remote control in Free Viewpoint Television, as depicted in Figure 1.

This user's interaction with the scene has created novel challenges from a coding and transmission perspective [1-3]. While in classical video streaming, the entire scene is encoded, delivered and displayed at the user side, in interactive/immersive systems only a portion of the full 3D scene is actually displayed. For example, in omnidirectional videos, the scene displayed in the HMD (viewport) is only a portion of the acquired spherical scene. Similarly, in multiview video coding, while a great number of views might be acquired,

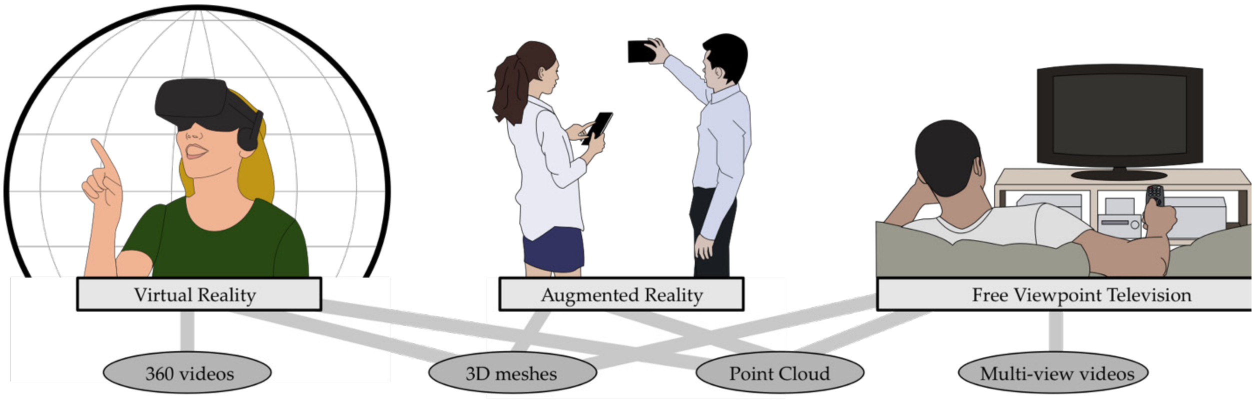


Figure 1. Immersive media technologies and their related data formats.

only a limited subset of them can be displayed at the same time. However, in practice the piece of data displayed by the user is not known *a priori* when both coding and streaming are performed. Therefore, when there is no prediction of the user behavior, the entire multimedia content needs to be coded and prefetched to users. This might lead to a reduction of the overall quality of the content in case of limited network resources. *It is therefore essential to properly predict users behavior to efficiently code and stream interactive content.* In this letter, we show how the user behavior is exploited in both bit allocation and streaming optimization strategies and we highlight the different interactive models that the two optimization problems require.

The Importance of Content Popularity in Bit Allocation Strategies

In data compression techniques, a maximum bit budget is usually available for compressing the data under consideration. The general criteria for an optimal compression is usually to describe with higher bitrates the more important data (and conversely), leading to an unequal bit allocation. In multi-view (MV) systems, for example, different cameras might be encoded at different quality levels [4-6], while in VR settings, different portions of the spherical content can be encoded with different quantization steps [7-9]. It is therefore essential to have proper metrics to reflect the « importance » of the data, *i.e.*, the content popularity. *In interactive services, this popularity reflects the probability for a piece of data to be displayed at the user's side*. In the following, we provide an overview on the optimization that an encoder needs to solve for carrying out the proper bit allocation strategy and we describe the associated challenges for the visual attention community.

Coding Problem formulation

Let us consider an interactive service, in which the video content acquired over time needs to be encoded. The overall goal for the encoder is to optimize the bit allocation strategy such that, on average, users consume high-quality media content while navigating. Decomposing a video content acquired by a camera into multiple portions, we denote by $x_{i,t}$ the i-th portion of the content acquired at time t. This can represent the i-th camera view in multi-view setting, or the i-th tile or portion of spherical content in VR settings. We then identify the full content acquired at time t with $\boldsymbol{x}_t = [x_{1,t}, \dots, x_{i,t}, \dots, x_{N,t}]$.

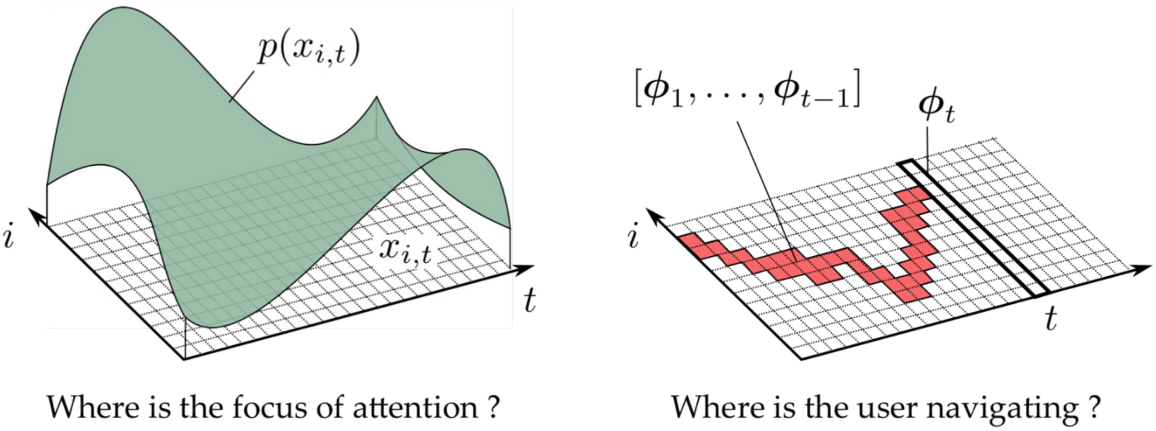
In this framework, the encoder seeks the best bit allocation strategy for each portion of the content. Denoting by $b(x_{i,t})$ the allocation for $x_{i,t}$ (*e.g.*, the QP for each $x_{i,t}$ content [9]), b(**X**) is the allocation strategy for the whole video content acquired in T successive acquisition time, with $\boldsymbol{X} = [\boldsymbol{x_1}, \dots, \boldsymbol{x_t}, \dots, \boldsymbol{x_T}]$. Therefore, the general problem formulation becomes

$$b^*: \arg\min_{b(\boldsymbol{X})} \sum_{t=1}^{T} \sum_{i=1}^{N} p(x_{i,t})\ D[b(x_{i,t})] + \lambda\, R[b(\boldsymbol{X})]$$

where $D[b(x_{i,t})]$ is the distortion of the i-th portion of the content acquired at time t when encoded with $b(x_{i,t})$ allocation strategy and $R[b(\boldsymbol{X})]$ is the total coding cost associated to the allocation strategy b(**X**).

Popularity estimation

The above problem formulation requires the *a priori* knowledge of *(i)* the video content characteristics (to evaluate D and R), *(ii)* the probability *p(***X***)* of a data **X** to be requested by a final user (*content popularity*, *cf.* Figure 2 left). The latter is a new metric needed for interactive services and how to predict this content popularity is still an open question. Therefore, a compelling question that we pose to the visual attention community is: « *how can we predict the content popularity?*» Or analogously, « *how can we estimate the probability p(***X***) of a data* **X** *to be requested by user?*».



Where is the focus of attention ?

Where is the user navigating ?

Figure 2. A priori popularity vs navigation modeling

The Importance of Navigation Paths Prediction in Personalized Streaming

While in the previous problem the goal was to seek the best compression strategy to maximize the quality for a multitude of users, here we rather focus on personalized streaming strategies, properly designed for a specific user or class of users. A possible application of this personalized strategy is the adaptive streaming system, where a video content is encoded in multiple representations (multiple coding rates and resolutions) and stored at the server, and the client selects the representation to download [10]. The intelligence on which representation best fits the need of each client is therefore located at the client side, where the user behavior is known. In the context of interactive strategies, this personalized strategies have been optimized for MV systems [11-13] as well as for omnidirectional content [14-16]. In both scenarios, the personalized strategy optimization is performed knowing the user's past displayed data and predicting the future user's navigation. In the following, we first provide a general overview on the optimization problem to be solved in personalized streaming strategies, and then we describe the challenges on user behavior prediction.

Streaming Problem formulation

Similarly to the bit allocation problem formulation, we consider the whole video content acquired in T successive acquisition times $\boldsymbol{X}$. We denote by $\pi_{i,t} = \pi(x_{i,t})$ the streaming strategy for the content $x_{i,t}$ and by $\boldsymbol{\Pi} = [\boldsymbol{\pi_1}, \dots, \boldsymbol{\pi_t}, \dots, \boldsymbol{\pi_T}]$, $\pi_t = [\pi_{1,t}, \dots, \pi_{i,t}, \dots, \pi_{N,t}]$ the strategy for the whole video $\boldsymbol{X}$. For example, $\pi_{i,t}$ can be a binary variable denoting whether $x_{i,t}$ is scheduled or not [11]. Differently, $\pi_{i,t}$ can specify which representation is sent to the user for $x_{i,t}$. At the client side, the final user displays only a portion of the overall acquired content. We therefore introduce a displaying variable $\phi_{i,t}$ such that $\phi_{i,t} = 1$ if the user displays $x_{i,t}$, $\phi_{i,t} = 0$, otherwise, and we generalize the display vector as $\boldsymbol{\Phi} = [\boldsymbol{\phi_1}, \dots, \boldsymbol{\phi_t}, \dots, \boldsymbol{\phi_T}]$, $\phi_t = [\phi_{1,t}, \dots, \phi_{i,t}, \dots, \phi_{N,t}]$.

Equipped with the above notation, the streaming optimization can be formulated as follows

$$\boldsymbol{\Pi}^*: \arg\min_{\boldsymbol{\Pi}} \sum_{t=1}^{T} \mathcal{D}[p(\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \boldsymbol{\Pi}] + \lambda R[\boldsymbol{\Pi}]$$

where $\mathcal{D}[p(\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \boldsymbol{\Pi}]$ is the objective function reflecting the quality experienced during the navigation or interaction (QoE). In interactive systems, this QoE does not take into account only the distortion of the displayed video, but also other factors such as the smoothness of the quality while navigating. A frequently-adopted metric for the QoE is, for example, the combination of both quality and quality variation over time [17]:

$$\begin{aligned}\mathcal{D}[p(\boldsymbol{\phi}_t | \boldsymbol{\phi}_{t-1}, \dots, \boldsymbol{\phi}_{t-K}), \boldsymbol{\Pi}] \\ &= \sum_{i=1}^{N} p(\phi_{i,t}) \, D(\pi_{i,t}) \\ &+ \mu \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} p(\phi_{i,t} | \phi_{j,t-1}) \Delta D(\pi_{i,t}, \pi_{j,t-1})\end{aligned}$$

where μ is the multiplier that allows to assign the appropriate weight to quality variations in the objective metric, and

$\Delta D(\pi_{i,t}, \pi_{j,t-1})$ is the distortion variation experienced over time. This variation is weighted by the probability $p(\phi_{i,t}|\phi_{j,t-1})$ of displaying the *i*-th portion at time *t*, given that the *j*-th portion was previously displayed. This probability reflects the *navigation path* of the user (*cf*. Figure 2 right).

User navigation modeling

Most of the works focusing on personalized streaming strategies assume to know (or accurately predict) the navigation path, while we actually know that estimating user interactivity is an open challenge. It is worth noting that in this personalized strategies, knowing a global content popularity $p(\boldsymbol{X})$ is not enough. It is additionally required to estimate the behavior of each user *over time*. Solving this problem must take into account both the visual content (as in the popularity estimation) and the user behavior modeling (*e.g.*, highly dynamic vs. static navigation). In other words, the open questions posed to the visual attention community are: «How do we model and categorize users behavior over time?» and «Knowing both the content displayed by one user in the past, and his behavior modeling, can we anticipate the future navigation path?».

References

[1] T. El-Ganainy, and M. Hefeeda. “Streaming Virtual Reality Content”, arXiv:1612.08350 (2016).

[2] M Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes”, *Proc. IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, Fukuoka, Japan, Oct. 2015.

[3] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, “Free-viewpoint TV”, *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67-76, Jan. 2011.

[4] G. Cheung, V. Vladan, and A. Ortega, “On dependent bit allocation for multiview image coding with depth-image-based rendering”, *IEEE Transactions on Image Processing*, vol.20, no. 11, pp. 3179-3194, May. 2011.

[5] J. Chakareski, V. Velisavljevic, and V. Stankovic, "View-popularity-driven joint source and channel coding of view and rate scalable multi-view video", *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no.3, pp. 474-486, Apr. 2015.

[6] A. De Abreu, G. Cheung, P. Frossard, and F. Pereira, "Optimal Lagrange multipliers for dependent rate allocation in video coding", *arXiv:1603.06123*, Mar. 2016.

[7] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo, and J. Wen, "Novel tile segmentation scheme for omnidirectional video", *Proc. of IEEE International Conference on Image Processing (ICIP)*, Phoenix, USA, Sep. 2016.

[8] A. Ghosh, V. Aggarwal, and F. Qian, "A rate adaptation algorithm for tile-based 360-degree video streaming", *arXiv:1704.08215*, Apr. 2017.

[9] F. De Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding", *Proc. of IEEE Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016.

[10] T. Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles", *Proc. ACM Conf. on Multimedia systems (MMSys)*, pp. 133-144, San José, USA, Feb. 2011.

[11] L. Toni, T. Maugey, and P. Frossard, "Optimized packet scheduling in multiview video navigation systems", *IEEE Transactions on Multimedia*, vol. 17, no.9, pp. 1604-1616, Sep. 2015.

[12] A. Hamza, and M. Hefeeda, "A DASH-based free viewpoint video streaming system", *Proc. of ACM Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV)*, pp. 55, Singapore, Singapore, Mar. 2014.

[13] M. Zhao, X. Gong, J. Liang, J. Guo, W. Wang, X. Que, and S. Cheng, "A cloud-assisted DASH-based scalable interactive multiview video streaming framework", *Proc. of Picture Coding Symposium (PCS)*, Cairns, Australia, Jun. 2015.

[14] M. Hosseini, and V. Swaminathan, "Adaptive 360 VR video streaming based on MPEG-DASH SRD", *Proc. IEEE International Symposium on Multimedia (ISM)*, San José, USA, Dec. 2016.

[15] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-Adaptive Navigable 360-Degree Video Delivery", *Proc. of IEEE Conference on Communications*, Paris, France, May 2017.

[16] L. Wang, D. Dai, J. Jiang, T. Yang, X. Jiang, Z. Cai, Y. Li, and X. Li "FISF: Better User Experience using Smaller Bandwidth for Panoramic Virtual Reality Video", *arXiv:1704.06444*, Apr. 2017.

[17] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, "Modeling the time—Varying subjective quality of HTTP video streams with rate adaptations", *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2206-2221, May 2014.

__Laura Toni__ received the M.S. and Ph.D. degrees, both in electrical engineering, from the University of Bologna, Italy, in 2005 and 2009, respectively. Between 2009 and 2011, she worked at the Tele-Robotics and Application (TERA) department at the Italian Institute of Technology (IIT), investigating wireless sensor networks for robotics applications.

In 2012, she was a Post-doctoral fellow in the Electrical and Computer Engineering Department at the UCSD. Between 2013 and 2016, she was a Post-doctoral fellow in the Signal Processing Laboratory (LTS4) led by Prof. Pascal Frossard at the Swiss Federal Institute of Technology (EPFL), Switzerland. In 2016, she has been appointed as Lecturer in the Electronic and Electrical Engineering Institute of University College London (UCL).

__Thomas Maugey__ graduated from Supélec, France in 2007. He received the M.Sc. degree from Supélec and Université Paul Verlaine, Metz, France, in 2007.

He received his Ph.D. degree in Image and Signal Processing at TELECOM ParisTech, Paris, France in 2010. His supervisors were Béatrice Pesquet-Poposecu and Marco Cagnazzo.

From October 2010 to October 2014, he was a postdoctoral researcher at LTS4 of EPFL, headed by Pascal Frossard. Since November 2014, he is a Research Scientist at Inria Rennes-Bretagne-Atlantique. He works in the team SIROCCO headed by Christine Guillemot.

On Streaming Services for Omnidirectional Video and its Subjective Assessment

Igor D.D. Curcio

Introduction

Virtual Reality applications and services that use omnidirectional video are recently making the highlights of news releases related to the most advanced consumer electronics technologies. In particular, streaming of 360-degree video content is one of the most compelling applications in this area. It is technically very challenging, among other reasons, because of the mismatch between the required transmission bandwidth for video and the network bit rates available today for consumers. The Moving Picture Expert Group (MPEG) is currently working on the first standard called Omnidirectional Media Format (OMAF) [1] to be used as common industry platform for encoding, storing, and the delivery of 360-degree video. This letter introduces some of the challenges related to subjective assessment of a streaming system for 360-degree video, and introduces a new metric that could be utilized in the assessment process.

Streaming services for 360-degree video lack of a proper standardized methodology and procedures for subjective assessment. Currently, there are several open issues, which require further research and standardization.

360-Degree Streaming Systems

A streaming system can be depicted in simplified form as in Figure 1. Here, the rendering device is a *Head Mounted Display (HMD)*. Between the server and the HMD there is typically a

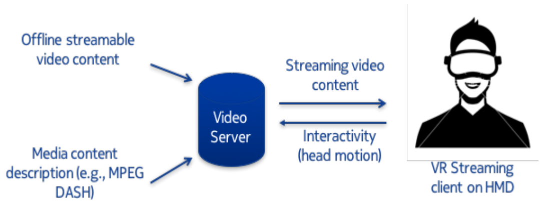

Offline streamable video content
Media content description (e.g., MPEG DASH)
Video Server
Streaming video content
Interactivity (head motion)
VR Streaming client on HMD

Figure 1. Example reference system for streaming omnidirectional video.

network connection which is wired or wireless (e.g., cellular or Wi-Fi) and that is affected by variable latencies.

When dealing with 360-degree video consumed on an HMD, there are, among others, few key parameters that are critical in such a system:

- The HMD *Field Of View (FoV)*.
- The size of the *foreground viewport*, which is the portion of omnidirectional video visible to the viewer in the horizontal and vertical directions. We call the other portions of the video which are not visible at a given time instant as *background viewport* (including part of the top and bottom portions of the 360-degree video).
- *Motion-to-Photon (MTP) Delay*, which is the elapsed time between the head motion to an orientation outside of the foreground viewport, and the subsequent system reaction to render a refreshed high quality viewport on the HMD. This is a factor that heavily impacts system interactivity.

The role of the above parameters in a streaming system for omnidirectional video will be clearer in the following.

Streaming techniques

Given the delivery of omnidirectional video is quite bandwidth hungry, compared to traditional 2D video, one of the main challenges for a successful streaming service is, more than ever, the deliver of the best possible visual quality using the smallest bandwidth.

360-degree video could be transmitted at a uniform quality, without differentiating between foreground and background viewports. This is the case when each rendered bit counts, and no compromises on video quality are permitted. This streaming technique is also referred to as *Viewport Independent Delivery (VID)*. From a bandwidth perspective, VID is quite demanding since the highest video quality is required all the time within the 360-degree space.

A great amount of transmission bandwidth could be saved by making use of the *Viewport Dependent Delivery (VDD)* technique. In this case, the foreground viewport is streamed at higher visual quality, whereas the background viewport(s) is (are) streamed at lower visual quality. This is a reasonable bandwidth saving mechanism, because for a given HMD orientation, the background viewport is not visible (and therefore not rendered).

As the HMD moves outside of the current foreground viewport, the visual quality may degrade for a certain time, until the system provides to stream and render a new foreground viewport corresponding to the new HMD orientation (*viewport switch*). Such time is essentially the MTP delay. The shorter this delay is, the better the user experience of such a system is. The larger (or unpredictable) this delay is, the more unusable becomes this system in terms of interactivity. Also, fatigue and motion sickness may be often experienced by a viewer.

Subjective Quality Assessment

The assumption that subjective assessment procedures for 2D/3D video used for several decades do apply also in the case of omnidirectional video watched with HMDs is too simplistic. In the latter case, the watching conditions and the immersion levels are different and, therefore, further research is needed. Here are some of the main challenges that currently lack of standardized methodology procedures.

Duration of the test video sequences: because of the wider FoV, compared to traditional video viewed on a 2D flat panel display, the length of video clips shall be sufficient such that all 360-degree content is watched and assessed. To perform such task, more time is needed for a test subject in order to explore the content in all directions. It appears logical that different durations for test video sequences may be envisaged in case of 180-degree content, as opposed to 360-degree content. A test case duration may be extended by allowing a looping function for a given video test sequence, and scored by a test subject when the judgment has been comfortably formed by a test subject. However, the duration of each test sequence (or test case) cannot be too long in order 1) not to produce fatigue symptoms on the test subjects, 2) to avoid that test sessions become too long and unmanageable, and 3) to avoid that test subjects forgive the impairments located in an early location of the test sequence (temporal forgiveness).

Full video assessment: for some use cases, such as 180-degree video, partial video subjective assessment may be sufficient. However, in the general streaming case of VDD of omnidirectional video, the subjective evaluation should not be limited to viewing a particular viewport orientation corresponding to a small portion of the whole video. This may happen, for example, if a viewer concentrates to watching only some details of the whole video in a quasi-still orientation, neglecting all other parts. For instance, there should be a way for the test subjects to form an opinion score based on the overall 360-degree video quality assessment, possibly without incurring in side effects, such as motion sickness or nausea.

Fair within-subject and between-subjects assessment: there might be the chance that the parts (and/or time instances) of a 360-degree test sequence viewed and assessed by a subject may differ from the parts viewed by the same subject for another test condition (or by another test subject when assessing the same test sequence). For ensuring a fair evaluation procedure, the test methodology should enable comparable test results for the same subject (or for several subjects) also in the case of not perfectly identical watching patterns for different test cases. In other words, it should be easy to verify whether, for different test cases, the same or different subjects view "the same thing at the same time". When this condition is not met, there should also be a way to measure how far apart are different viewing patterns for different test conditions.

Similarity Ring Metric (SRM)

A new metric for the last of the challenges introduced in the previous section is presented here. Such metric measures essentially the degree of similarity of a set of watching patterns [2]. For simplicity, the remainder of this discussion will focus on Yaw which, by convention, measures the horizontal FoV.

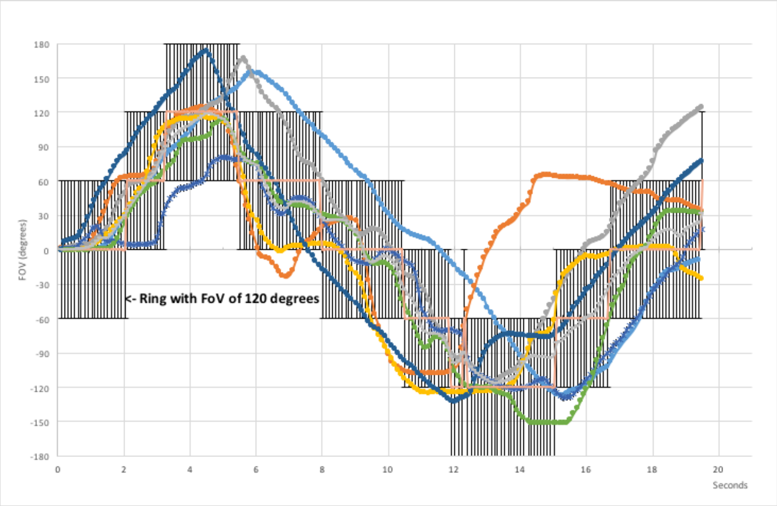


Figure 2. Example Ring with size of 120 degrees. The ring moves through the curves at discrete steps according to the foreground viewport orientation. Here the test subjects were instructed to follow a specific motion pattern without any speed constraints.

A typical plot of Time vs. Yaw could look like in Figure 2. Each curve represents the watching pattern of one (or more) subjects when viewing the same omnidirectional test sequence. It can be clearly seen that, for each time instant, the curves follow the same rough direction, but they are far apart by a certain distance. In practice, it is rare that all curves overlap (i.e., exploit a perfect watching similarity), since each test case carries some elements of variability even within the same test subject. These elements are direction and speed of motion while watching a video with an HMD.

However, it is possible to verify if the aggregate set of curves falls within a certain range. We could ideally think of this range as a "ring" (see Figure 2). The goal is then to check if the ring can travel through all curves from the beginning to the end of a test clip. If this occurs, it means that all clips (i.e., the curves) have been watched with high similarity.

More specifically, if the curves are related to several test cases of the same sequence watched by a test subject, high similarity means that the subject has been watching and assessing the same content at the same time. Differently, if the curves are related to different test subjects that evaluate a particular clip, high similarity means that the test subjects have been watching and assessing the same content at the same time.

The ring size is a critical parameter here, and is determined by the HMD FoV or the content FoV. In Figure 2, the ring size is the size of the foreground viewport, which is 120 degrees.

As it is difficult to achieve a SRM of 100%, it is convenient to define a *Similarity Threshold* ST, e.g., 80%. In this way, a rejection criterion could be established: for example, the results of a particular subjective test set could be rejected if SRM < ST.

Future Developments

The support for multi-dimensional SRM with pitch and roll is one of the development areas. Furthermore, it is worth mentioning that the SRM could also be tailored to tiled streaming supported, for example, by the HEVC video codec. Further research advances are envisaged also in the area of foveated streaming as soon as adequate hardware for gaze tracking will be available in mainstream HMDs.

***Igor D.D. Curcio** is Principal Scientist at Digital Media Laboratory of Nokia Technologies, Tampere, Finland.*

References

[1] "Text of ISO/IEC FDIS 23090-2 Omnidirectional Media Format", ISO/IEC JTC1/SC29/WG11 Doc. N17235, Macau, China, Oct. 2017.

[2] "Similarity Ring Metric (SRM) for Subjective Evaluation of 360-Degree Video", 3GPP TSG SA WG4, Tdoc AHVIC-090, Nokia, Mar. 2017.

Subjective Video Quality Database for Virtual Reality

Zhenzhong Chen and Yingxue Zhang

Introduction

With the development of virtual reality (VR) and related technologies, the establishment of immersion calls for higher quality of video contents. However, the processing such as stitching and compressing on the videos greatly influences the quality. Therefore, quality assessment for panoramic videos attaches much importance in specifying and promoting the quality of experience (QoE).

We establish a subjective video quality database for virtual reality. 48 panoramic video sequences with different levels of compression impairments are viewed and rated through HTC VIVE by 30 non-expert subjects.

A subjective quality database for panoramic videos is established through a subjective rating test with virtual reality HMD, from which we can 1) figure out the observers' psychophysical response to the VR contents, 2) provide reliable reference for evaluating the performance of the objective assessment methods.

Subjective Quality Assessment Test

As shown in Figure 1 and Table 1, 10 panoramic common test sequences released by MPEG [1] are adopted as reference sequences. All the sequences are in the format of equirectangular (ERP), lasting for 10s each.

Coding impairments are introduced to the reference videos to obtain test sequences using HM-16.14 with 360-Lib at 5 QP points, i.e., 22, 27, 32, 37, 42. After processing, a total of 60 sequences on different but relatively stable quality levels are prepared for the experiment, among which reference "AerialCity" and its corresponding impaired sequences are used for training, three sequences from "ChairLift" for stabilizing and the remaining 48 for testing.

The original sequences of different resolutions are sampled to a consistent resolution for presenting on HTC VIVE before coding. Fixed QP values guarantee a consistent quality of different videos on the same compression level.

Table 1. Information of original test sequences in ERP format [1]

Class	Sequence name	Frame count	Resolution@FPS	Bit-depth
8K	Train_le	600	8192x4096@60	8
8K	SkateboardingTrick_le	600	8192x4096@60	8
8K	SkateboardInLot	300	8192x4096@30	10
8K	ChairLift	300	8192x4096@30	10
8K	KiteFlite	300	8192x4096@30	8
8K	Harbor	300	8192x4096@30	8
4K	PoleVault_le	300	3840x1920@30	8
4K	AerialCity	300	3840x1920@30	8
4K	DrivingInCity	300	3840x1920@30	8
4K	DrivingInCountry	300	3840x1920@30	8

The videos are presented one at a time with HTC VIVE and are voted independently. The subjects can view the contents on all directions freely. The reference sequences are also displayed and voted without any special identification, dubbed Hidden Reference [2]. All the test sequences will be presented randomly and only once. Absolute five-grade scale is used to rate the video quality considering the quality range. The final rating scores for the test sequences are defined using Difference Mean Opinion Score (DMOS).

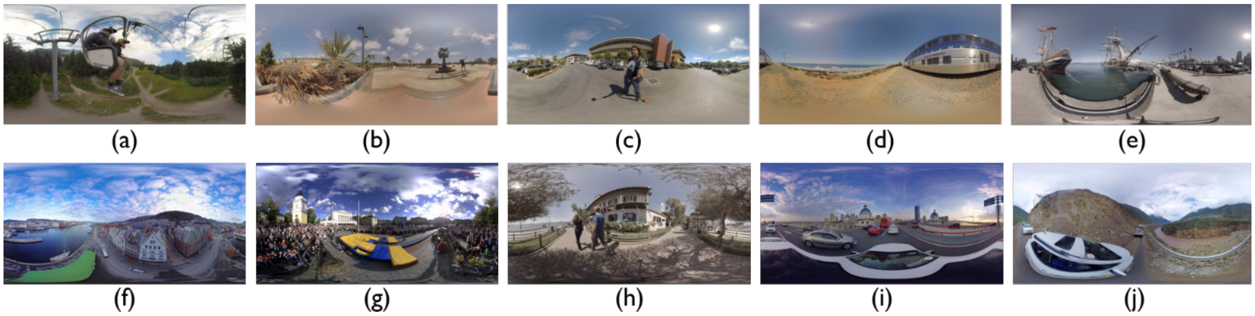


Figure 1. Thumbnails of the ten references used in the test. (a) ChairLift, (b) SkateboardingTrick_le, (c) SkateboardInLot, (d) Train_le, (e) Harbor, (f) AerialCity, (g) PoleVault_le, (h) KiteFlite, (i) DrivingInCity, (j) DrivingInCountry.

As aforementioned, the subjects can view the video freely. Despite of the high consistency on viewing pattern, the free-viewing task will unavoidably lead to some extreme conditions that some subjects may focus on totally different factors from the others. Therefore, the number of subjects for each test is suggested to be more than 15 being recommended for 2D video assessment. A larger number of subjects guarantees the reliability when some extreme data exists.

We recruit 30 subjects to participate in the assessment tests. The subjects are undergraduate and graduate students, including 17 males and 13 females. None of the subjects majors in quality assessment or related areas, nor do they involve in the design or further analysis of the tests. They are asked to evaluate the overall quality of the video.

The Subjective Video Quality Database

After experiment, a set of rating data is obtained for all the sequences. Before calculating DMOS, post-experiment screening is conducted to assess subject reliability and ensure a valid data set. If a subject does not respond according to the instructions, the data has to be discarded. Firstly, a subject's data will be discarded if there is any missed rating [3]. Secondly, the subject with unreliable ratings will also be screened, which is specified in [4].

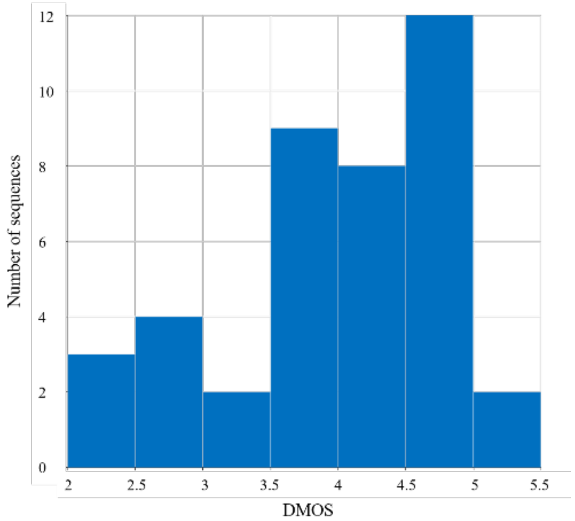


Figure 2. Histogram of the DMOS uniformly spaced between the minimal and maximal values.

In total, the ratings of 3 subjects are discarded by the screening process. DMOS is calculated with the remaining reliable scores on reference and test sequences.

Figure 2 shows the histogram of the DMOS for all the test sequences. The DMOS lies in the range of [2.04, 5.08], corresponding to mean Z-score range of [-0.6, 3], which covers approximately 72% of the area of standard normal distribution.

Conclusion

A subjective quality database for panoramic videos is established through a compact subjective rating test involving 30 subjects and 60 sequences with coding impairments of different levels. The DMOS of the sequences is calculated on the basis of validated subjective ratings and is reasonably distributed on the quality range. Therefore, the database can be promising in further VR applications.

References

[1] J. Boyce, E. Alshina, A. Abbas, Y. Ye, “JVET common test conditions and evaluation procedures for 360° video”, Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D1030, 4th Meeting, Oct. 2016.

[2] ITU-T, “Subjective video quality assessment methods for multimedia applications”, Rec. ITU-T P. 910, Apr. 2008.

[3] VQEG, “Final Report from the video quality experts group on the validation of objective models of video quality assessment”, Mar. 2000.

[4] ITU-R, “Methodology for the subjective assessment of the quality of television pictures”, Rec. ITU-R BT. 500, Jan. 2012.

Zhenzhong Chen *is a professor at Wuhan University. He leads the Institute of Intelligent Sensing and Computing conducting researches on multimedia technology, computer vision, image processing and understanding, data mining for geoinformatics, etc.*

Yingxue Zhang *is a Ph. D. student at Wuhan University. She is a member of the Institute of Intelligent Sensing and Computing. Her research interests include quality assessment and computational vision.*

Quality Assessment Challenges in MPEG's Current and Future Immersive Media Standards

Sebastian Schwarz and Sébastien Lasserre

Introduction

Recent years have shown significant advances in immersive media experiences. Three-dimensional representation formats allow for new forms of entertainment and communication. In this context, point cloud data has emerged as a promising enabler for such experiences. Because efficient enough point cloud compression technologies are still to be found, the Moving Picture Expert Group (MPEG) has just issued a Call for Proposals (CfP) on point cloud compression technologies. This letter will present the MPEG CfP evaluation procedure and try to anticipate some of the many challenges to be faced when assessing point cloud compression performance.

Dynamic point clouds have been identified as a promising format to code immersive worlds allowing free navigation to the user. The geometry-based description of data leads to new challenges, both compression technologies and quality assessment of the compressed immersive world.

MPEG Call for Point Cloud Compression Technology

There is now a huge interest from the Virtual Reality market in being able to represent the world in three dimensions, thus enabling the end-user to freely navigate in this world. MPEG has launched an ambitious roadmap including future coding technologies of 3D scenes. One of these technologies is Point Cloud Compression (PCC) and is expected to be delivered as an ISO standard in 2019/20. MPEG has issued a call for proposals on PCC technology and aims to evaluate submissions in October 2017 [1].

A point cloud is the given of a set of points, each defined by its 3D XYZ location and attributes, e.g. colour, reflectance, opacity. The MPEG call addresses various applications, resulting in several submission categories, i.e. static, dynamic, and dynamically acquired point clouds, and coding conditions, i.e. lossless geometry with lossy attributes and no/lossy geometry with lossy attributes. Testing material varies from huge, high-precision static point clouds, e.g. for map generation, to smaller but dynamic point clouds, thought as an input of a VR system. This article will focus on the lossy compression of the latter, as quality evaluation is considered the most critical for this scenario.

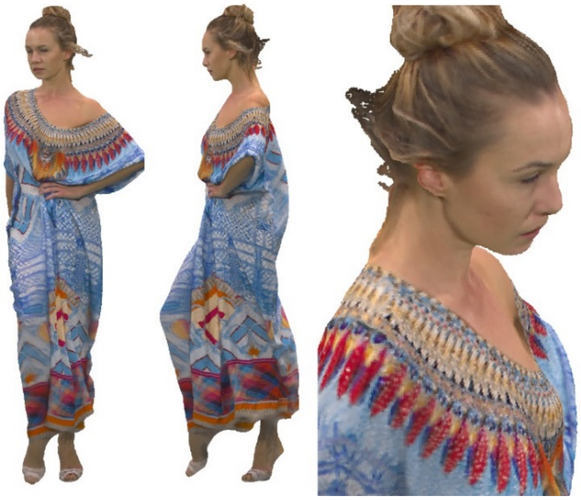


Figure 1. A three-dimensional object represented by a point cloud [3]. By the very nature of the point cloud format, free navigation is possible around the object.

Evaluation Anchors

For the compression of dynamic point cloud data, MPEG requests submissions based on a set of five test sequences, each with roughly one million points per "frame". An example for such a point cloud sequence is shown in Figure 1. Five target bit rates must be achieved for each sequence, ranging from 3 to up to 72 Mbit/s, to cover a wide range of use cases, for a total of 25 test points. At each test point, a proponent's decoded point cloud sequence will be evaluated against the competing submissions, as well as an anchor encoding generated with the provided experimental PCC software.

The anchor software relies on subsampling an octree representation for geometry and JPEG-based colour compression for attributes. The software allows for simple temporal prediction structures (IPIPIP), however, this feature is not used for the CfP. An example for the anchor compression result is shown in Figure 2. Due to the geometry subsampling, the decoded point cloud (middle) has fewer points than the original (left). This effect must be taken account in the objective and subjective quality assessment, for example by increasing the rendered point size for subjective viewing (right).

Objective Quality Assessment

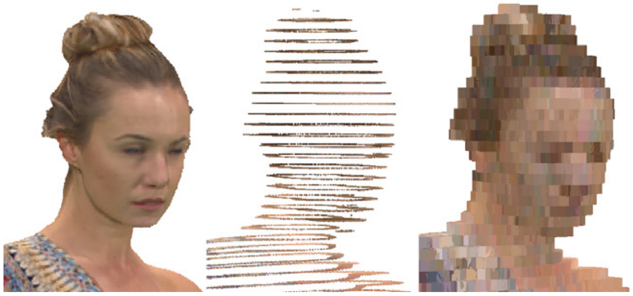


Figure 2. Point cloud compression results using the anchor software [2] at 13 Mbit/s: Original, decoded point cloud (geometry subsampled), decoded point cloud rendered with larger point size.

Classically, encoding performance is assessed in a rate-distortion fashion, comparing the achieved bit rate with the introduced distortions. For 2D video, peak signal-to-noise ratio (PSNR), based on the mean squared error (MSE) between original and decoded pixel, is the most accepted distortion metric. While the PSNR does not necessarily fully represent all effects of the human visual system, it works well for typical 2D video coding artefacts, such as blocking and blurring.

For 3D point clouds, the relation between original and decoded point is not straight forward. As seen in Figure 2, the decoded point cloud might have less (or more) points than the reference point cloud. Furthermore, the decoded point has two kinds of distortion, geometry distortion and colour distortion.

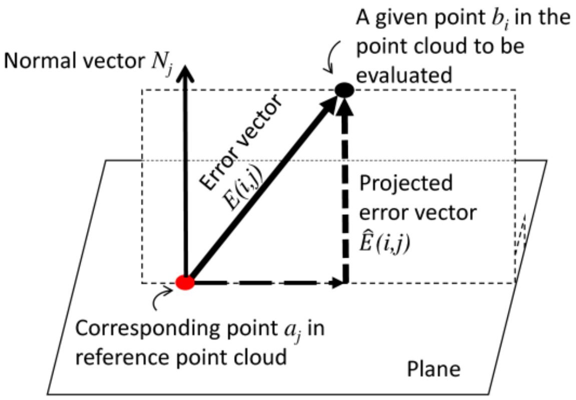

A given point b_i in the point cloud to be evaluated
Normal vector N_j
Error vector $E(i,j)$
Projected error vector $\hat{E}(i,j)$
Corresponding point a_j in reference point cloud
Plane

Figure 3. Illustration of the point-to-point error E (D1) and point-to-plane error Ê (D2).

Therefore, the CfP specifies three distortion metrics. The first metric, D1, calculates the MSE between a position of a point and the position of its closest neighbour in a reference point cloud (point-to-point). D2 calculates the MSE between the position of a point and its projection onto a given reference plane, representing the surface of the point cloud model (point-to-plane). The difference between these two error calculations is illustrated in Figure 3. Finally, colour distortion is calculated in YUV domain, between the current point and its closest reference neighbour (D1).

To take the possibility of a largely varying number of original and decoded points into account, both metrics are run twice.

First, comparing all decoded points to the original, then comparing the original to the decoded points, and symmetric results of both runs are reported.

Subjective Quality Assessment

It is apparent that purely objective quality evaluation for point cloud data suffers from similar problems as 2D video. In addition, with no research on the relation of D1 and D2 to the actual perceived quality, and no knowledge of the different effects and relationship between geometry and colour distortion on subjective quality, subjective quality assessment becomes a key necessity for reliably assessing coding performance. However, there is no standard procedure for assessing visual quality of dynamic point cloud data. In order to establish some kind of standardised procedure, two key aspects need to be addressed: First, the rendering of the points, and second, how to ensure stable viewing between different test subjects and test points.

Regarding the rendering, a point has no size and is not supposed to be visible. Consequently, it must be visualised by something with some shape and size to make an object represented by a point cloud viewable. Given that points are located on a three-dimensional uniform integer-based grid, the minimum shape that fills the space between adjacent points without overlapping is a cube of size unity.

However, due to possible geometry sampling induced by the compression system, the optimal cube size may not be unity. The visual effect of this size is shown in Figure 3 and 4. A powerful rendering scheme would be to allow a local cube size depending on the location of the neighboring points, but this would interfere in some uncontrolled way on the compression scheme and may hide compression artifacts. Since the goal of MPEG is to provide a compression system and not a renderer, it has been decided to impose a uniform shape (cube) for all points and allow the proponents to provide a given point size for their decoded content to be rendered at.

Concerning the methodology for subjective quality evaluation and ensuring stable viewing conditions, it has been decided to not let the participants navigate freely around the object. Instead, 2D video based on a fixed path around the object will be rendered. This path is unknown to the proponents beforehand. In addition to ensuring stable viewing conditions for all test subjects, this approach has the benefit to keep participants focused on the quality evaluation and not distract them with the navigation controls. Video quality can then be evaluated using standardised methods for assessing subjective quality.

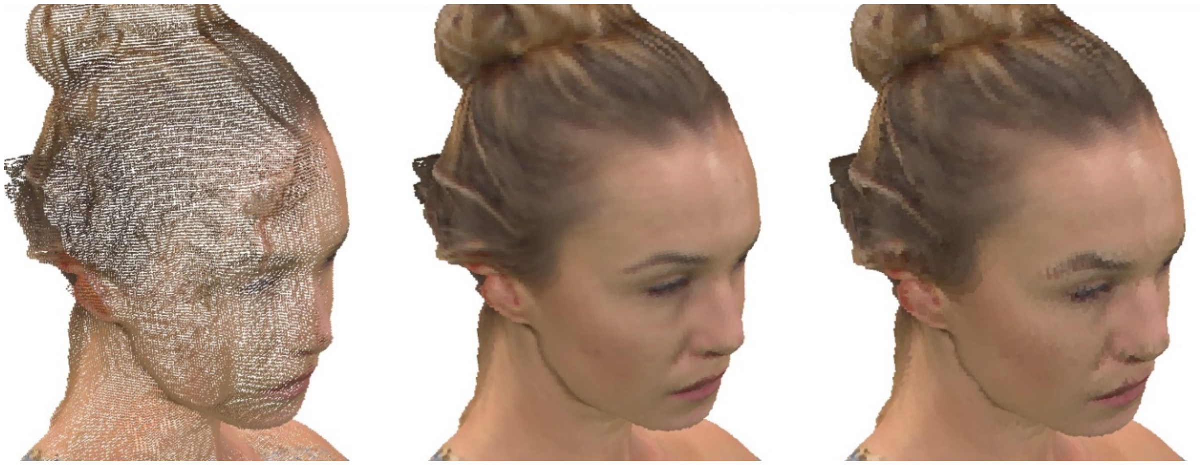


Figure 4. Effect of the rendered point size on the visual quality of the object. A too small size leads to a ghostly rendering (left) but, on the other hand, a too big size masks texture fine details (right). Determining the optimal point size (center) is one of the biggest challenges behind the subjective quality assessment.

Perspective and Future Work

Looking at the presented objective and subjective quality assessment for point cloud compression technology, it becomes apparent that this field is far less researched than 2D video quality assessment. MPEG is aware that the chosen approaches do not necessarily present the final word in assessing point cloud compression quality. Nonetheless, they represent the current state of research at the time of issuing the CfP and should allow for an initial assessment.

Nonetheless, any help on refining point cloud compression quality assessment is more than welcome. Details on the currently chosen approaches are available in [1]. Interested VQEG experts are invited to contribute to this activity. In particular, inputs on the following problems are of high interest to the authors:

- What extensions or improvements to the current distortion metrics could be considered? Are there any other reliable metrics other than D1 and D2?
- How do the metrics for geometry distortion relate to perceived visual distortion?
- What is the relationship between geometry and colour distortion when it comes to visual quality?
- How to reliably assess the effects of geometry sub-/ over-sampling on objective and subjective quality?
- What are the effects of temporal geometry distortions on the visual quality. How to assess them?
- How to standardise visually quality assessment for (dynamic) point cloud data.

As for future work, the MPEG CfP has been issued and proponents are invited to submit their solutions. The above-described objective and subjective quality assessment will be carried out in October 2017 and results should be available by the 120th MPEG meeting. We intend to publish a follow-up article discussing the outcome and faced challenges during this evaluation in a later edition of this VQEG eLetter.

Sebastian Schwarz
Nokia Technologies, Finland

Sébastien Lasserre
Technicolor, France

References

[1] Call for Proposals for Point Cloud Compression V2. ISO/IEC JTC1/SC29/WG11 Doc. N16763, Hobart, Australia, Apr. 2017.

[2] R. Mekuria, K. Blom, P. Cesar, "Design, Implementation and Evaluation of a Point Cloud Codec for Tele-Immersive Video", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828-842, Apr. 2017.

[3] E. d'Eon, B. Harrison, T. Myers, and P. A. Chou, "8i Voxelized Full Bodies – A Voxelized Point Cloud Dataset." ISO/IEC JTC1/SC29/WG11 Doc. M40059, Geneva, CH, Jan. 2017.

Perceptual analysis and characterization of light field content

Jesús Gutiérrez, Pradip Paudyal, Marco Carli, Federica Battisti, Patrick Le Callet

Introduction

The Light Field (LF) may be defined as the set of light-rays at every point in space traveling in every direction. The possibility of capturing this information provides a wide range of applications in various fields, such as surveillance, industrial and medical exploration, and immersive media technologies. In this sense, LF content allows novel ways to explore the captured scenes, like changing the parallax horizontally and vertically, and refocusing the content.

The novel applications provided by light field technologies entail a reconsideration of the methods for assessing Quality of Experience, starting with a proper characterization of the light field content.

This new imaging technology causes new challenges to the information processing system. To guarantee a successful development of the technology, the signal processing chain (coding, processing, delivering, storing) should take into account the peculiarities and the effects of possible impairments on the visual quality. To cope with these challenges, as experienced with previous audiovisual technologies, like 3D video, Quality of Experience (QoE) assessment is an essential factor.

Therefore, this article addresses, on one side, the requirements for properly tackling the perceptual aspects in LF processing, and on the other side, the proper characterization of LF content according to its applications.

LF basics: Perceptual perspective

Adelson and Bergen in [1] defined the plenoptic function to represent the intensity of light seen from any viewpoint, at any time instant, for any wavelength of the visible spectrum. The complexity associated with its high dimensionality can be reduced for practical imaging applications, as shown by the four-dimensional parameterization of the plenoptic function, which represents each light ray by its intersecting points with two parallel planes [2]. Therefore, the LF can be considered as a collection of perspective images of same scene, each one taken from a different viewpoint.

Content acquisition

The previous statement leads to the most intuitive way to acquire LFs, based on camera arrays [3]. However, LFs can be also obtained by using plenoptic cameras, which are based on inserting a microlens array between the camera sensor and the main lens. The main lens creates an image that is re-mapped to the sensor by the microlens array, that provides multiple views of the scene in a single shoot [4]. The differences between both alternatives entail distinct processing of the content and perceptual effects. For instance, on one side, camera arrays provide a set of views with wider baselines and better spatial resolutions. On the other side, plenoptic cameras offer the advantage of being much easier to handle and provide a denser set of views, although they entail a complex decoding process of the raw data (including demosaicing, devigneting, rectification, etc.) that can introduce artifacts and whose perceptual effects should be further explored [5].

Representation formats

Once the raw data is processed, it is possible to use different representations of the LF depending on the application under study. Among these, we can cite viewpoint images (a.k.a. sub-aperture images in plenoptic cameras, representing the scene captured from different viewpoints), the entire plenoptic image captured by the plenoptic camera, microlens images (a.k.a. elemental images or micro-images, captured by each microlens of the plenoptic camera), or epipolar images (containing depth information of the captures scene) [6].

Processing and encoding

The different representation formats of the LF are directly related to the processing that is addressed in the following. One of the main issues regarding LF imaging is the extraction of the 3D information of the captured scene. Nowadays, depth estimation and 3D reconstruction are active research areas [7]. Moreover, increasing the spatial and angular resolution of the acquired content is an important issue to be solved in order to offer improved image quality and 3D perception to the viewers [7]. Finally, given the high redundancy of LF content, many efforts are being devoted to the design of efficient compression techniques [6][8][9].

Rendering

The rendering and display of LF content are also a major issue directly influencing end users' QoE. For example, a simple approach is based on using conventional displays simulating LF applications, like interactive refocusing or viewpoint sweeping. To fully take advantage of the immersivity and interactivity of this content, head-mounted displays (HMDs) may be used if the content has been appropriately captured (e.g., omnidirectional LF content); at the same time, LF displays, which are still under development, will be the best way for visualizing LF content without the need of any other specific equipment (e.g., HMD, glasses, etc.) [3].

Related Works

This section presents an overview of the first efforts towards the QoE evaluation for LF content that have been made lately.

Datasets

Ongoing efforts are devoted to creating LF datasets, especially after the availability of affordable plenoptic cameras (e.g., Lytro and Raytrix). For example, the EPFL LF dataset provides 118 images captured with the Lytro Illum camera and covers a wide range of high-level features [10] and the SMART LF dataset collects 15 LF images designed for image quality assessment [9]. Similarly, Daudt and Guillemot published a Lytro Illum LF dataset containing 43 images for various applications, such as depth estimation, inpainting and compression [11]. Furthermore, it is also worth noticing the existence of datasets generated by different devices, such as the Stanford light field archive [12] captured with a camera array, and synthetic LF dataset generated by computer graphics [7]. However, there is still a lack of further datasets with annotated data from subjective tests to support the research on LF technologies.

Quality assessment

The new possibilities that immersive media technologies offer to the user experience require a revision of traditional methods for QoE evaluation. For example, as the appearance of 3D content entailed the consideration of evaluating visual discomfort and 3D perception in comparison with conventional video content, factors involved in the new immersive experience should currently be addressed, such as full-parallax, adaptive refocusing, interactivity, immersivity, cyber-sickness, etc. In addition, other aspects of QoE evaluation should be further investigated, such as appropriate testing environments, methodologies, and proper test content.

In this sense, some initial works have been proposed for evaluating the quality of LF content. In particular, various studies have been presented dealing with the quality assessment for LF compression algorithms. For example, Viola *et al.* [8] carried out a subjective test to compare different encoding approaches for LF images and analyzed the performance of traditional objective metrics like PSNR and

SSIM on this content. Similarly, Paudyal *et al*. [9] carried out an exhaustive analysis of subjective and objective quality evaluation of compressed LF images, using traditional methodologies and metrics. Apart from these approaches, only few works have addressed the QoE evaluation of LF in relation with other aspects, such as the visualization of LF content in LF displays (which are still under development) [13], and the effect of interactivity when the user is able to change the focus of the image and the viewpoint [14]. Taking this into account, further research on appropriate methodologies for subjective assessment and on reliable objective metrics for LF content is required to correctly evaluate perceptual and technical factors on the QoE.

Characterization of LF content

One of the main issues to deal with when assessing the QoE is the selection of contents to use in the tests under study, which should be based on visual characteristics and on the purpose of the experiment, rather than on personal preference or convenience [15]. This fact emphasizes the need for content characterization to model those aspects. In fact, important efforts have already been made to properly characterize 2D content, usually focused on analyzing spatial, temporal, and color features [16]. Moreover, the advent of 3D content with its new features (e.g., horizontal disparity, depth range, visual discomfort) showed the need for integrating novel features for a complete data characterization [17].

In this sense, the novel characteristics and applications of emerging immersive media technologies require a reconsideration of content characterization. With this aim, we proposed a framework for characterizing and selecting LF content [18], which will be summarized in the following. This framework was especially designed for QoE assessment, considering the new applications that LF technology provides, such as adaptive refocusing and full parallax.

Proposed scheme

The proposed framework is based on the analysis of various indicators representing 2D properties, together with 3D features and refocusing characteristics, given the importance of depth information provided by LFs and the novel possibilities of changing the focused elements of the content. The considered features are described in the following, and some illustrative examples are shown in Figure 1:

- *Spatial and temporal information*: The SI recommended by the ITU is widely used for this purpose, so it was adopted in the proposed framework [16]. Similarly, altough the proposed framework was dedicated to LF images, the TI recommended by the ITU may be used for describing the temporal aspects of video sequences.
- *Colorfulness*: It is an important visual feature having a significant impact on the perceptual quality of a scene. Thus, the proposed framework recomends to use the metric proposed by Hasler *et al.* [19], given its proved performance.
- *Contrast:* This property conveys meaningful perceptual information (e.g., textures, entropy, etc.). In the proposed scheme, the use of the Gray Level Co-occurrence Matrix (GLCM) is adopted for textural and contrast description [20].
- *Depth map and depth histogram*: Different approaches should be used depending on whether the LF data has been acquired by camera arrays (e.g., multi-view methods) or by plenoptic cameras (e.g., especific methods based on multi-view correspondences or occlusions [21]), due to the different acquisition properties (e.g., baseline). In the proposed scheme, for simplicity, the Lytro Desktop software was used to obtain the depth maps and from them, the histograms were computed.
- *Disparity range*: It defines the distance, in terms of pixels, corresponding to the nearest and furthest objects of the

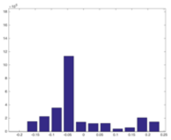
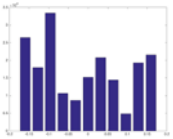
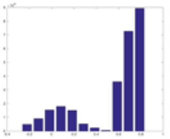
Preview			
Dataset	Own	EPFL	Own
Application	Viewpoint changing	Refocusing	Viewpoint changing & refocusing
Spatial Indicator	34.10	36.14	55.21
Colorfulness	10.12	45.85	38.97
Contrast	0.07	0.14	0.59
Refocusing Range	[-0.4,0.2]	[-0.3, 0.4]	[-1.4, 0.1]
Occluded Pixels	930	813	171
Disparity Range	[-0.16, 0.22]	[-0.16, 0.16]	[-0.22, 0.79]
Depth Distribution			

Figure 1. Characterization examples [18]

scene. To obtain this, the range of the scene in terms of distances to the objects or the camera calibration parameters, are required. Also, pixel disparities may be obtained by using estimation methods, such as the multi-view stereo algorithm [7], used in the proposed scheme.

- *Occlusions*: Although it is one of the most important problems when dealing with 3D content, only few algorithms address occlusions in LF. In the proposed scheme the amount of occluded pixels was computed using the algorithm by Wang *et al.* [21].
- *Refocusing range*: This describes the region from the nearest to the furthest elements of the scene that can be focused. For this purpose, it is possible to analyze the properties of the disparity histogram, which provides information about the depth distribution of the scene (as shown in Figure 1). Also, some objective metrics may be helpful, such as those developed for coping with the blur effect, or some specific approaches for LF content, like the Multi-focal Scene Defocus Quality (MSDQ) metric [22]. Finally, it is possible to use refocusing algorthims (e.g., “shift & sum” proposed by Ng *et al.* [4]) to determine the refocusing range going from the nearest to the furthest object in the scene. In the proposed framework, an implementation of this

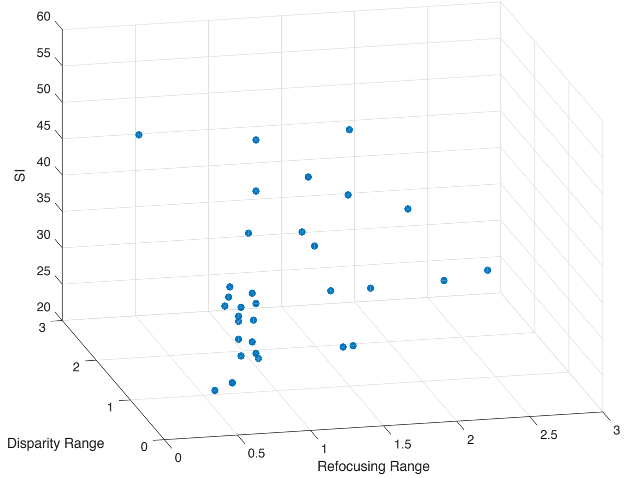


Figure 2. 3D scatterplot with SI, refocusing and disparity ranges [18].

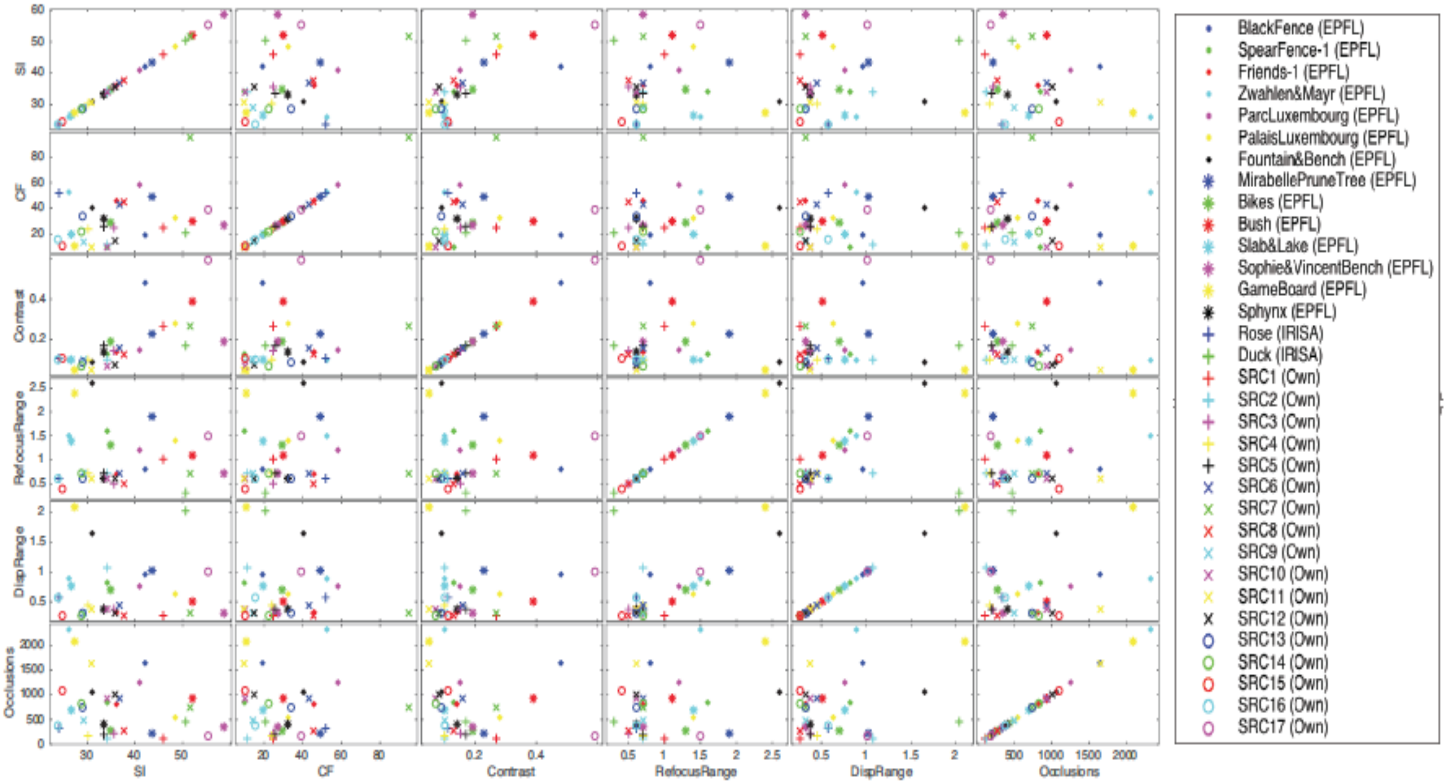


Figure 3. Scatterplot matrix of the main selected features [18].

algorithm was used to define the refocusing range [5].

The set of selected features can be graphically represented by different means based on the specific purpose, such as content selection based on a set of features. As an example, it might help to identify the lack of useful content as presented in the following. Figures 2 and 3 show two possible representations. In this case, images from different datasets where considered. A limited diversity of contents for important LF features, such as refocusing and disparity ranges might be noted. This may emphasize the need for generating and publishing more datasets for which the proposed approach for LF content characterization may be useful.

Conclusions

This article provided an overview of the perceptual aspects related to the processing and QoE assessment of LF content, and highlights the need for a revision of these aspects should be

Jesús Gutiérrez is a post-doctoral researcher (Marie Curie/Prestige fellow) at the IPI group pf the LS2N of the Université de Nantes. He received the Telecommunication Engineering degree in 2008 from the Universidad Politécnica de Valencia (Spain), and the Ph.D. degree in Telecommunication in 2016, from the Universidad Politécnica de Madrid (Spain). His research interests are in the area of image and video processing, evaluation of multimedia quality of experience, and 3D and immersive media systems.

addressed, given the new possibilities and applications provided by emerging immersive technologies. In this context, we also provided some insights on proper LF content characterization as a first step towards further research on QoE assessment.

Pradip Paudyal *is Assistant Director at Nepal Telecommunications Authority (NTA), Kathmandu, Nepal. He received the M.Sc. degree in Information and Communication Engineering from Tribhuvan University, Kathmandu, Nepal, in 2010 and the Ph.D. degree from Department of Engineering at the Università degli Studi Roma TRE, Roma, Italy, in 2017. He is also serving as a Visiting Lecturer in different universities in Nepal. His research interests are in the area of multimedia signal processing, light field image processing, and perceptual quality assessment.*

Marco Carli *is Assistant Professor with the Department of Engineering at the Università degli Studi 'Roma TRE', Roma, Italy. He received the Laurea degree in Telecommunication Engineering from the Università degli Studi di Roma 'La Sapienza', Roma, Italy and the Ph.D. degree from Tampere University of Technology, Tampere, Finland. He was a Visiting Researcher with the Image Processing Laboratory, UCSB, USA (2000-2004). His research interests are in the area of digital signal and image processing with applications to multimedia communications. He is Area Editor of Elsevier Signal Processing: Image Communication.*

References

[1] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision", in "Computational models of visual processing", M. Landy and J. A. Movshon, Eds. MIT Press, pp. 3–20, 1991.

[2] M. Levoy, "Light fields and computational imaging", *IEEE Computer*, vol. 39, no. 8, pp. 46–55, Aug. 2006.

[3] K. Akeley, "Envisioning a light field ecosystem", *SID Symposium Digest of Technical Papers*, vol. 43, no. 1, pp. 459–462, Jun. 2012.

[4] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera", *Stanford Tech Report CTSR*, Apr. 2005.

[5] D. G. Dansereau, O. Pizarro, and S. B.Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras", *Proc. Of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1027–1034, Portland, USA, Jun. 2013.

[6] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Full parallax 3D video content compression", in "Novel 3D media technologies", A. Kondoz and T. Dagiuklas, Eds., Springer, 2015.

[7] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[8] I. Viola, M. Řeřábek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and Subjective Evaluation of Light Field Image Compression Algorithms", *Proc. of Picture Coding Symposium*, Nuremberg, Germany, Dec. 2016.

[9] P. Paudyal, F. Battisti, M. Sjostrom, R. Olsson, and M. Carli, "Toward the perceptual quality evaluation of compressed light field images," *IEEE Transactions on Broadcasting*, vol. 63, no. 3, pp. 507–522, Sep. 2017.

[10] M. Řeřábek and T. Ebrahimi, "New light field image dataset", *Proc. Of the International Conference on Quality of Multimedia Experience*, Lisbon, Portugal, Jun. 2016.

__Federica Battisti__ received the Laurea Degree (Master) in Electronic Engineering and the PhD degree from Roma Tre University in 2006 and 2010 respectively. Her research interests include signal and image processing with focus on subjective quality analysis of visual contents. She is currently assistant professor at the Department of Engineering at Roma Tre University.

__Patrick Le Callet__ received both an M.Sc. and a PhD degree in image processing from Ecole polytechnique de l'Université de Nantes. Since 2003, he teaches at Ecole polytechnique de l'Université de Nantes where is now a Full Professor. His current centers of interest are Quality of Experience assessment, Visual Attention modeling and applications, Perceptual Video Coding and Immersive Media Processing. He is co-author of more than 250 publications and communications and co-inventor of 16 international patents on these topics. He serves or has served as associate editor or guest editor for several Journals such as IEEE TIP, IEEE STSP, IEEE TCSVT, SPRINGER EURASIP Journal on Image and Video Processing, and SPIE JEI. He is serving in IEEE IVMSP-TC (2015- to present) and IEEE MMSP-TC (2015-to present) and is one the founding member of EURASIP SAT (Special Areas Team) on Image and Video Processing.

[11] R. Daudt and C. Guillemot, "Lytro Illum light field dataset", 2016. Online: https://www.irisa.fr/temics/demos/IllumDatasetLF/index.html

[12] V. Vaish and A. Adams, "The (new) Stanford light field archive," 2008. Online: http://lightfield.stanford.edu/

[13] P. A. Kara, M. G. Martini, P. T. Kovacs, S. Imre, A. Barsi, K. Lackner, and T. Balogh, "Perceived quality of angular resolution for light field displays and the validy of subjective assessment", *Proc. Int. Conf. on 3D Imaging*, pp. 1–7, Liege, Belgium, Dec. 2016.

[14] I. Viola, M. Řeřábek, and T. Ebrahimi, "Impact of interactivity on the assessment of quality of experience for light field content", *Proc. of the International Conference on Quality of Multimedia Experience*, Erfurt, Germany, Jun. 2017.

[15] M. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests", *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 50-61, Aug. 2013.

[16] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Rec. ITU-T P.910*, 2008.

[17] M. Urvoy, J. Gutiérrez, M. Barkowsky, R. Cousseau, Y. Koudota, N. García, V. Ricordel, and P. Le Callet, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences", *Int. Workshop on Quality of Multimedia Experience*, pp. 109–114, Yarra Valley, Australia, Jul. 2012.

[18] P. Paudyal, J. Gutiérrez, P. Le Callet, M. Carli, and F. Battisti, "Characterization and selection of light field content for perceptual assessment", *Proc. of International Conference on Quality of Multimedia Experience*, Erfurt, Germany, Jun. 2017.

[19] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images", *Proc. of the SPIE*, vol. 5007, pp. 87–95, Jun. 2003.

[20] R. M. Haralick, K. Shanmugam, I. Dinstein, "Textural features for image classification", *IEEE Trans. on systems, man, and cybernetics*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[21] T.-C. Wang, A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2170–2181, Nov. 2016.

[22] W. Wu, P. Llull, I. Tosic, N. Bedard, K. Berkner, and N. Balram, "Content-adaptive focus configuration for near-eye multi-focal displays", *Proc. of the IEEE International Conference on Multimedia and Expo*, Seattle, USA, pp. 1–6, Jul. 2016.