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].

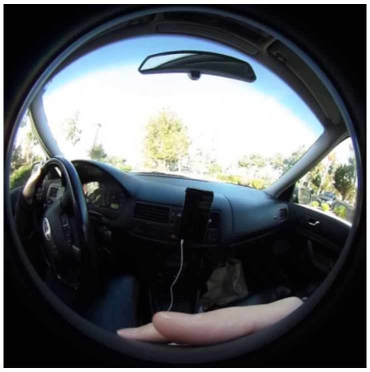

Figure 1. Example of optical distortions: image captured with fish-eye lens.

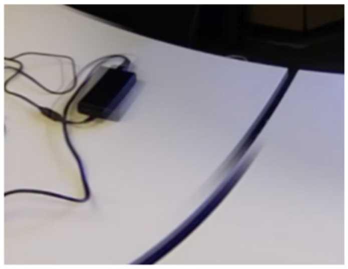

Figure 2. Example of stitching discontinuity on portion of equirectangular image.

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

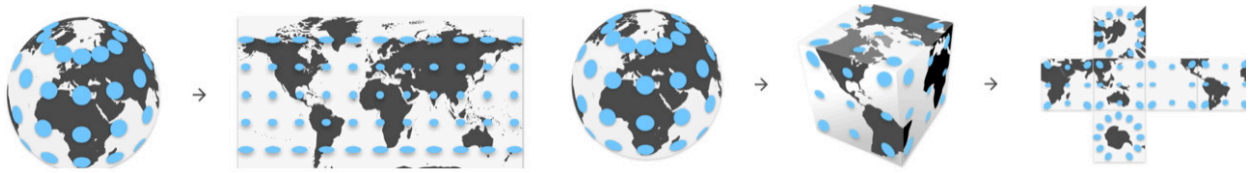

Figure 3. Example of map projections (equirectangular on the left, cube map on the right) and related warping distortions: blue circles on the spherical surface are mapped to ellipses with varying axis lengths on the plane.

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.

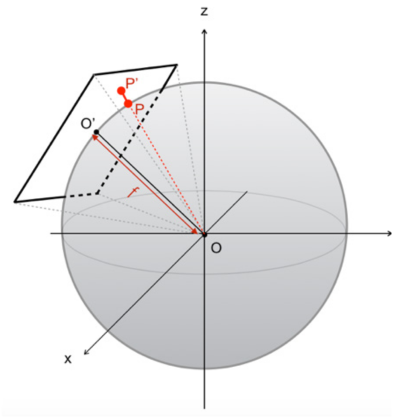

Figure 4. Geometry of viewport: the oblique gnomonic projection projects point *P* on the viewing sphere to point *P'* in the plane at focal length *f* from the center of projection *O*, defined by the viewport.

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].

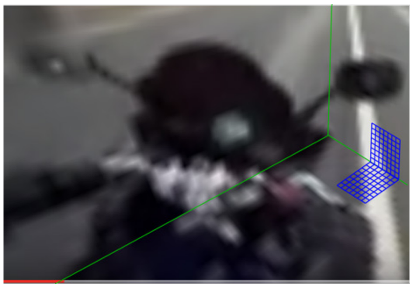

Figure 6. Example of viewport extracted for compressed cube-map panorama: discontinuities between the cube faces in the panoramic arrangement (highlighted by green lines) may cause wireframe cube and underlying image sampling domain (highlighted by blue grids) to be visible.

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.*

__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.

[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.