This paper has been presented at **SPIE Optical Engineering + Applications (OP18O), Applications of Digital Image Processing XLI Conference, San Diego, CA, USA, August 2018.**

Subjective and Objective Quality Assessment of Omnidirectional Video

Francisco Lopes, João Ascenso*, António Rodrigues, Maria Paula Queluz
Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais, 1049-001, Lisboa, Portugal

ABSTRACT

Omnidirectional video, also known as 360º video, is becoming quite popular since it provides a more immersive and natural representation of the real world. However, to fulfill the expectation of an high quality-of-experience (QoE), the video content delivered to the end users must also have high quality. To automatically evaluate the video quality, objective quality assessment metrics are then required. This paper starts by presenting the results of a subjective assessment campaign that was conducted to evaluate the impact, on quality, of HEVC compression and/or spatial/temporal subsampling, when the videos are displayed in a head mounted device (HDM). The subjective assessment results are then used as ground-truth to evaluate conventional quality assessment metrics developed for 2D video, as well as some of the recently proposed metrics for omnidirectional video, namely, spherical peak-signal to noise ratio (S-PSNR), weighted to spherically uniform PSNR (WS-PSNR), and viewport PSNR (VP-PSNR); in the context of this study, the adaptation of two SSIM based metrics, to omnidirectional contents, are also proposed and evaluated.

Keywords: omnidirectional video, 360º video, subjective quality assessment, objective quality assessment, virtual reality, quality-of-experience.

1. INTRODUCTION

Nowadays, omnidirectional video is gaining more and more importance. The sensation of immersion given by these videos, pictures or games, create a much better user experience than the one provided by traditional 2D video, putting the user at the center of the action and even making it part of that action. Virtual reality (VR) based games are being exploiting more and more by important companies, such as Sony with its PlayStation VR head mounted display (HMD). Also, content providers, like YouTube, noticed a growing demand for virtual reality content. Nowadays, a lot of omnidirectional videos can be accessed through YouTube, which can be rendered on common smartphones, equipped with a gyroscope, or even with the mobile phone inserted into a HMD. Other HMDs have been developed and some of them, already in the market, can render omnidirectional videos without the use of any smartphone, like the Oculus Rift, though more expensive. However, to fulfill the expectation of an high quality-of-experience (QoE), the video content quality delivered to the end users must be reliable measured. In fact, although recognizing that the assessment of virtual reality experiences is more than just assessing the video quality, this is still a critical component in providing a truly immersive experience.

To automatically evaluate the video quality, objective quality assessment metrics are required. Although several quality metrics have been developed in the last years for 2D video, in the omnidirectional case new types of distortions can be present, due to the fact that the processing (e.g., compression) and transmission are typically done in a planar domain; however, for rendering, the video is projected back to a sphere and then to the viewport, and the distortions introduced by the previous processing are influenced due to warping and interpolation [1]. Additional, the perceptual impact of those distortions may also be influenced by the rendering of visual signals in HMDs, and in an extent that may not be accurately predicted by existing 2D metrics.

Although some objective metrics have already been developed for the specific case of omnidirectional videos, namely spherical peak-signal to noise ratio (S-PSNR) [2], latitude PSNR (L-PSNR) [2] weighted to spherically uniform PSNR (WS-PSNR) [3], area weighted spherical PSNR (AW-PSNR) [4], and viewport PSNR (VPSNR) [2], they are slight variations of the conventional PSNR, and were not proven to accurately model the perceived quality, especially when

* joao.ascenso@lx.it.pt

head mounted devices (HMD) are used to render and display omnidirectional videos. An assessment of these metrics with respect to subjective ground-truth data was conducted in [5], but limited to compression artifacts in still images.

This paper starts by describing and analyzing a subjective assessment test campaign, that was conducted to evaluate the impact, on quality, of HEVC compression and/or spatial/temporal subsampling of omnidirectional videos, when the videos are displayed in a head mounted device (HDM). The subjective assessment results, that were made available in [6], are then used as ground-truth to evaluate conventional quality assessment metrics developed for 2D video, as well as some of the recently proposed metrics for omnidirectional video, namely S-PSNR, WS-PSNR, AW-PSNR and VP-PSNR; in the context of this study, the adaptation of two SSIM based metrics to omnidirectional contents are also proposed and evaluated.

This paper is organized as follows: after the introductory section, Section 2 overviews the key steps of a typical omnidirectional video processing chain. Section 3 presents a review on objective video quality assessment metrics, developed specifically for omnidirectional videos; new adaptations of SSIM [7] and MS-SSIM [8], to omnidirectional content, are also proposed. Section 4 describes the procedures followed on the subjective assessment tests and presents and analyses the subjective tests results. The main objective is to evaluate the impact, on the perceived video quality, of three main sources of distortion, namely spatial and temporal subsampling, HEVC compression, and their combined effects. Section 5 evaluates the objective quality assessment metrics, using the subjective results as ground-truth data. Section 6 concludes the paper, highlighting the most important conclusions and putting forward some suggestions of future work.

2. OMNIDIRECTIONAL VIDEO PROCESSING CHAIN

This section presents the typical omnidirectional (or 360°) video communication chain, describing the main concepts behind the creation, compression, rendering and visualization of omnidirectional video content. Figure 1 presents the key processes involved on the transmission of 360° video, from acquisition to display.

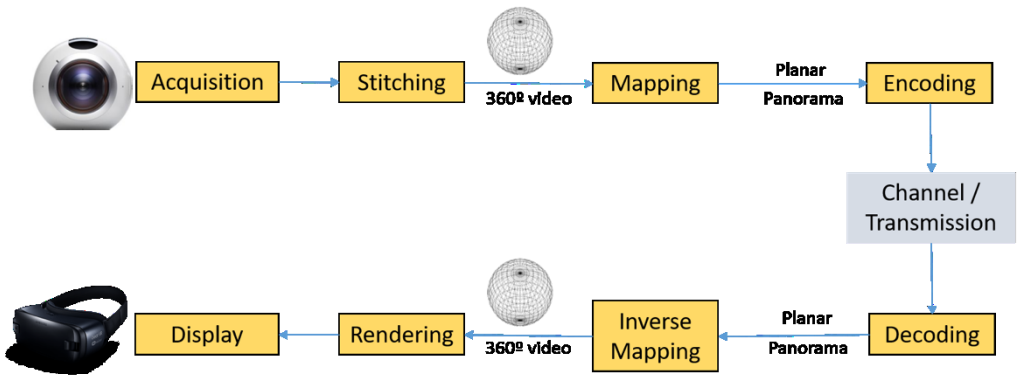


Figure 1. Generic architecture of an omnidirectional video transmission chain [9-11].

The main functions of each module are:

- **Acquisition:** The acquisition of omnidirectional video is typically done with multiple cameras, that are time synchronized, calibrated, and uniformly placed along a rig; each camera´s lens points to a different direction, so that each camera acquires a 2D image corresponding to a portion of the spherical view around it; together, and to obtain omnidirectional video, the multiple cameras cover the whole sphere.
- **Stitching:** After the acquisition, the 2D images recorded by each camera are combined to create an omnidirectional image, containing the 360º information of the scene. This process is called *stitching* and registers, aligns and blends images that typically overlap [12]. In this step, the omnidirectional image is a non-planar frame; usually, a spherical representation is used.
- **Mapping:** To be transmitted, the spherical omnidirectional frame is mapped into a planar representation. The process of transforming a non-planar frame into a planar frame is called mapping. The most used planar projection is the equirectangular projection but there are many other projections, with different properties.

- **Encoding:** Since a planar representation of each omnidirectional video frame was obtained in the previous step, a 2D video codec can be used, such as MPEG-4/AVC [13] or HEVC [14]. Before encoding, an additional step - called tilling - can be applied, that divides the omnidirectional image into several tiles which are independently encoded. This is useful to control which quality each tile will have, e.g. tiles not perceptually relevant can be encoded with lower quality, and to stream only parts of the content that will be watched by the users.
- **Channel/Transmission:** The bit stream generated by the encoding step is then stored or sent to the client over a fixed or wireless communication channel. Nowadays, omnidirectional video content is typically transmitted from a server to a client with an HTTP based protocol.
- **Decoding:** The decoding step of the processing chain performs the inverse operation of the encoder and, at the user end, the reconstructed omnidirectional video is obtained.
- **Inverse Mapping:** For rendering the omnidirectional video, a spherical representation is often used. Therefore, the transmitted planar video has to be mapped into a sphere, by applying the corresponding inverse mapping transformation of the sender.
- **Rendering:** In omnidirectional video, the images that are presented to the user are a part of the entire viewing sphere. Depending on the user viewing direction, a selected part of the sphere is projected on a 2D plane, resulting in the so-called viewport. There are several projections that can be used to perform rendering but the popular perspective projection is widely used nowadays.
- **Display:** The output of the rendering step is a 2D image that can be presented on a display. The displays for omnidirectional video are of two types: the first corresponds to a navigable image on a standard 2D display (e.g., a computer or a smartphone screen), where the viewing direction can be controlled by a mouse or by moving the display; the second type corresponds to a head mounted display (HMD), which is a display that a user wears on his head, and tracks user's head movements to compute the corresponding viewport.

Each one of the aforementioned steps may induce distortions on the omnidirectional video content; the characterization of these distortions, and the development of quality metrics able to capture the impact of perceivable impairments introduced by the different components of the communication system, are still open and challenging issues.

3. OBJECTIVE QUALITY ASSESSMENT METRICS FOR OMNIDIRECTIONAL VIDEO

As mentioned in the previous section, the omnidirectional video undergoes several transformations along the transmission chain, till being displayed in the user terminal screen, and some of these transformations introduce artifacts on the video content. To assess the artifacts impact of the video quality, objective metrics are needed, whose quality prediction should be well correlated with the quality perceived by the users. This paper is mainly concerned with the quality degradation introduced by the encoding, which results mainly from quantization, spatial subsampling and/or temporal subsampling.

In the last years, several objective metrics have been developed for 2D images and video. However, the differences between conventional 2D content and omnidirectional content are vast, and dedicated metrics for objective quality assessment of spherical visual content have been recently proposed [2-4]. Most of them are simple adaptations of the full-reference peak signal-to-noise ratio (PSNR) metric recognizing that, for 360^o video quality evaluation, the metric must consider sampling on the sphere, instead of a given planar format; they are overviewed on Section 3.1. In Section 3.2, adaptations of the SSIM [7] and MS-SSIM [8] metrics to omnidirectional content are also proposed.

3.1 Overview of objective quality metrics for omnidirectional images

In the following, some of the metrics developed for omnidirectional video are described:

- **Viewport PSNR (VPSNR) [2] -** In VPSNR, only the image samples belonging to a viewport are subject of a traditional PSNR evaluation, which is applied between corresponding viewports of the reference and impaired videos. This metric was proposed to evaluate the impact of various panoramic projections on the coding efficiency of a video encoder. However, since the actual head motion data is not known beforehand, it may not give an accurate measure of the perceived (and global) video quality.

- **Spherical PSNR (SPSNR) [2] -** In SPSNR, for each image sample on the spherical video, the correspondent image samples on the reference and impaired planar projections are firstly obtained, and the error between them is computed; then, the PSNR is obtained by averaging the errors over the entire set of points of the sphere. Figure 2 presents graphically the SPSNR procedure. One of the SPSNR disadvantages results from the fact that a sample on the sphere might not correspond to an integer position on the planar projections, requiring the use of some interpolation procedure, which may condition the metric result.

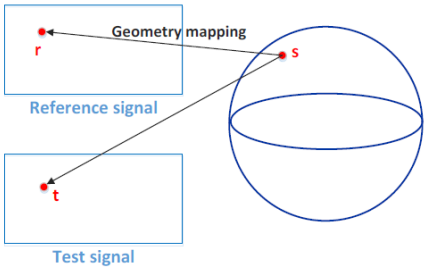


Figure 2. SPSNR procedure: starting in s, the quality degradation is measured between the reference sample *r* and the impaired sample *t* [2].

- **Latitude SPSNR (L-SPSNR) [2]:** L-SPSNR weights the sphere points per their corresponding latitude access frequency, thus giving more weight to pixels belonging to the front central areas of the image, and less importance to the areas near the poles, almost never accessed.
- **Weighted to Spherically Uniform PSNR (WS-PSNR) [3]** and **Area Weighted Spherical PSNR (AW-SPSNR) [4] -** AW-SPSNR and WS-PSNR are two independently proposed 360° video quality metrics based on the same concept. These metrics weighs the pixel error computed between pixels on the reference and impaired planar images, according to the corresponding pixel area (in WS-PSNR), or solid angle (in AW-PSNR), on the spherical surface; it is formally described by (1) and (2), where X_{ij} and Y_{ij} refers to the intensity of the reference and impaired images at pixel (i,j), respectively, M is the maximum possible intensity of the image, W and H are the width and height of equirectangular image, respectively, and w_{ij} is the weighting factor for pixel (i,j). This weight is given by (3), where W_{ij} is the scaling factor of area (in WS-PSNR) or solid angle (in AW-PSNR) from planar to spherical surface, at pixel (i,j). For the equirectangular planar projection, W_{ij} is given by $\cos(\theta)$ for both metrics, where θ is the latitude corresponding to pixel (i,j); this factor represents the inverse of the horizontal stretching resulting from the equirectangular planar projection. WS-PSNR and AW-SPSNR do not need interpolations, which is the main advantage relatively to SPSNR.

$$WMSE = \sum_{j=1,i=1}^{W,H} (X_{ij} - Y_{ij})^2 . w_{ij} \tag{1}$$

$$WS\_PSNR = 10 \log\left(\frac{M^2}{WMSE}\right) \tag{2}$$

$$w_{ij} = \frac{W_{ij}}{\sum_{j=1,i=1}^{W,H} W_{ij}} \tag{3}$$

It is worth to mentioned that any of the previous metrics can be combined with saliency maps, under the assumption that distortions in salient areas are more visible and, therefore, more annoying.

3.2 Proposed adaptations of SSIM and MS-SSIM to omnidirectional visual content

In this section, two new metrics are proposed, resulting from adaptations of the Structural Similarity Index (SSIM) [7] and of the Multi-Scale Structural Similarity Index (MS_SSIM) [8], both developed for 2D images, to omnidirectional content.

The SSIM metric measures the perceptual similarity between two images; it is based on three comparison between the samples of X and Y: luminance, l, contrast, c, and structure, s. These functions are computed according to (4) to (6), where u_X and u_Y refer to the mean, σ_X and σ_Y to the standard deviation and σ_{XY} to the cross correlation, of pixel values inside a window, b, of $N \times N$ pixels, and C_1, C_2, C_3 are small stabilizing constants. SSIM is then obtained by a weighted combination of those comparative measures, according to (7), where α, β and γ define the relative importance of each component. To obtain a single SSIM value per frame (or per video), the SSIM values per pixel (or per frame) are then averaged.

$$l_b(X,Y) = \frac{2u_X u_Y + C_1}{{u_X}^2 + {u_Y}^2 + C_1} \tag{4}$$

$$c_b(X,Y) = \frac{2\sigma_X \sigma_Y + C_2}{{\sigma_X}^2 + {\sigma_Y}^2 + C_2} \tag{5}$$

$$s_b(X,Y) = \frac{\sigma_{XY} + C_3}{\sigma_X \sigma_Y + C_3} \tag{6}$$

$$SSIM_b(X,Y) = [l_b(X,Y)]^{\alpha} \cdot [c_b(X,Y)]^{\beta} \cdot [s_b(X,Y)]^{\gamma} \tag{7}$$

The MS-SSIM metric is an extension of the SSIM, by incorporating image details in several spatial resolutions. First, the image is decomposed in L spatial resolutions, or scales, indexed by z in (8). Then, the luminance is compared only at scale L, while contrast and structure are compared in all spatial resolutions.

$$MS\_SSIM_b(X,Y) = [l_b^L(X,Y)]^{\alpha_L} \cdot \prod_{z=1}^{L} [c_b^z(X,Y)]^{\beta_z} \; [s_b^z(X,Y)]^{\gamma_z} \tag{8}$$

As for the SSIM, to obtain a single MS-SSIM value per frame (or per video), the MS-SSIM values per pixel (or per frame) are then averaged.

The adaptations of SSIM and MS_SSIM to omnidirectional videos are based on the same principles of WS_PSNR and AW-PSNR, where a correction factor is applied according to the relationship between the area, or the solid angle, covered by a pixel in the planar projection and on the sphere:

- **Weighted SSIM (W-SSIM)** - The W-SSIM metric can be summarized in the following steps:
    1. For each pixel of the equirectangular image under evaluation, compute its SSIM index according to (7), and considering a window, b, of 11×11 pixels around the pixel; the α, β, and γ, parameters values are those suggested in [7].
    2. Apply the area correction factor, w_b, given by (10), resulting in a weighted SSIM value per pixel, $SSIM_b^w$

$$SSIM_b^w(X,Y) = SSIM_b(X,Y).w_b \tag{9}$$

$$w_b = \frac{W_b}{\sum_n W_n}. \tag{10}$$

    In (10), $W_b = \cos(\theta_b)$, and θ_b is the latitude (in the sphere) corresponding to the center of window b.

    3. Compute a weighted SSIM value per frame, W_SSIM, by summing up all $SSIM_b^w(X,Y)$ values in the frame.
    4. Compute the W_SSIM mean considering all frames, to obtain a single value per video sequence.
- **Weighted MS-SSIM (WMS_SSIM)** - The procedure to extend the MS_SSIM index to omnidirectional content is similar to the one presented above for the SSIM. The weighted MS_SSIM index per pixel, $MS\_SSIM_b^w$, is now computed as:

$$MS\_SSIM_b^w(X,Y) = [l_b^L(X,Y)]^{\alpha_L}.w_b^L.\prod_{z=1}^{L} [c_b^z(X,Y)]^{\beta_z}.[s_b^z(X,Y)]^{\gamma_z} \tag{11}$$

with L=5, and the α, β, and γ, parameters values at each scale are those suggested in [8]. To obtain a weighted MS_SSIM value per frame, WMS_SSIM, and also per video sequences, steps 3 and 4 of the W_SSIM procedure are applied to (11).

4. SUBJECTIVE QUALITY ASSESSMENT OF OMNIDIRECTIONAL VIDEO

This section describes the procedures followed on the subjective assessment of omnidirectional videos, and presents and analyses the subjective assessment results. The main objective is to evaluate the impact, on the perceived video quality, of three main sources of distortion - spatial subsampling, temporal subsampling, HEVC compression - and of their combined effects. The subjective assessment results were made available in [6].

4.1 Subjective Assessment Framework

Figure 3 presents the subjective assessment framework architecture; each branch of the architecture corresponds to the assessment of one video distortion type - spatial down sampling, temporal down sampling, HEVC compression - or to their combined effects; the yellow boxes refer to the process of introducing a certain type of distortion in the video sequences and the blue boxes refer to the subjective assessment of the distortion impact on quality.

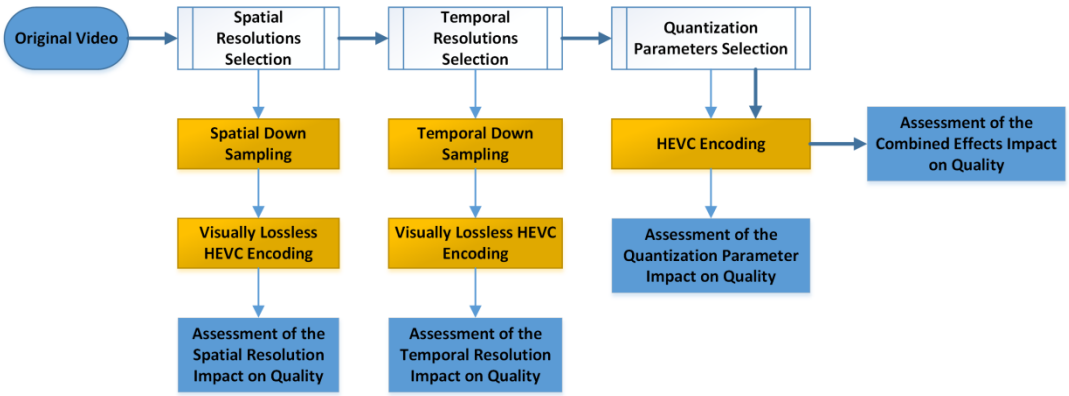


Figure 3. Subjective assessment framework.

4.2 Subjective Tests Procedures

The subjective evaluation of omnidirectional videos with HMD devices does not allow the use of double stimulus assessment methods, since it is not possible to display two stimuli in simultaneous in the screen; therefore, only single stimulus methods are allowed. Moreover, the use of HMD devices may cause tiredness and queasiness after long time viewings. These effects, which causes are still not clearly identified and understood, can be frequent, depending on the subject and on the video content characteristics. Therefore, the subjective tests should have short duration.

Two test sessions were conducted, each one composed of two stages: the first session included the assessment of both spatial and temporal subsampling impact on quality; in the second session, the impact of compression and of the combined distortions, were considered. The number of participants in the first and in the second sessions was, respectively, 20 and 17. Each session lasted around 25 minutes. Subjects were from both genders, with ages between 22 and 55 years old, and included experts and non-experts in the image processing research field. Between the two test stages, subjects had to take off the HMD for one or two minutes to rest the eyesight and avoid dizziness.

Two single stimulus assessment methodologies were selected, namely, the Absolute Category Rating with Hidden Reference (ACR-HR) for the individual distortions tests, and the Absolute Category Rating (ACR), for the combined distortions assessment; both methodologies are standard procedures defined in [15]. In both cases, the quality scale used by the subjects varies between 1 (Bad) and 5 (Excellent). With ACR-HR, it is possible to express the subjective evaluation in both mean opinion score (MOS) and differential mean opinion score (DMOS) scales, where MOS

measures the absolute perceived quality for a certain video sequence, and DMOS measures the quality relative to the reference video; with ACR, only MOS values can be obtained.

At the beginning of each test, and without using the HMD, the subjects were introduced to the objectives of the session; then, with the HMD already put on (see Figure 4-a)), a brief training session took place, so that the subjects could be familiarized with the evaluation interface and with the distortions types and their extreme cases. A swivel chair, that allows subjects to freely move and explore the entire omnidirectional view while sitting, was used.

a) b)

Figure 4. a) A subject during a subjective session; b) The scoring image presented after each test video sequence.

During the session, each test video was displayed for 10 seconds; after that, a still image (see Figure 4-b)), with the evaluation scale (1 to 5) was also displayed during 10 seconds, where the subject should tell, to a test host, the score given to the previous video sequence. Each session included a repeated video sequence, to evaluate the subject's consistency. In each test, the generated video sequences were shown in a random order, which was the same for all subjects. People with glasses kept them during the assessment sessions, inside the HMD.

The used HMD was Oculus Rift [16], running under the Oculus Software, with the GoPro VR Player 2.3 application as VR media player. The graphic card, GEFORCE GTX 1060 3 GB, can display videos with up to 7680×4320 pixels@60 Hz of resolution. For this reason, the maximum spatial resolution considered in the tests was 7680×3840 pixels.

The results obtained in each test were validated by applying the procedure recommended in [17]. Subjects consistency was also assessed by repeating one sequence in each test. Subjects with a score difference higher than 2 in the same sequences were also considered outliers and removed from the final MOS. This repeated test sequence was not included in the MOS computation. After outliers detection and removal, the resulting MOS and DMOS for the test video sequence k, MOS_k and $DMOS_k$, were then obtained by (12) and (13) respectively, where S is the number of consistent subjects, $x_{k,s}$ is the score given by subject s to the test video k, and r is the reference video corresponding to test video k.

$$MOS_k = \frac{\sum_{s=1}^{S} x_{k,s}}{S}, \tag{12}$$

$$DMOS_k = \frac{\sum_{s=1}^{S} (x_{k,s} - x_{r,s} + 5)}{S}. \tag{13}$$

4.3 Characterization of the omnidirectional video test set

The initial video test set was chosen from a group of ten YUV videos in the JVET dataset [18], which are available in the equirectangular format. They are in the 4:2:0 chroma subsampling format and have a length of 10 seconds each; a sub-set of six videos was selected, whose main characteristics are summarized in Table 1.

Table 1. Initial video test set characteristics.

Sequence	Width [px]	Height [px]	#Frames	Frame Rate [fps]	Observations
PoleVault	3840	1920	300	30	Many spatial details/texture
Harbor	8192	4096	300	30	Low motion
SkateboardInLot	8192	4096	300	30	Moving camera / High motion
ChairliftRide	8192	4096	300	30	Low motion
SkateboardTrick	8192	4096	520	60	Low motion
Train	8192	4096	600	60	High motion

4.4 Assessment of the spatial resolution impact on quality

Each video from the initial test set was iteratively down sampled, with a sampling factor of 2 in each iteration (in width and height), starting from its maximum resolution, till 960×480 pixels; a Lanczos filter was used. Since the available graphic card had a resolution of 7680×3840 pixels, videos with higher resolution were first down sampled to the card resolution. The resulting spatial resolutions for each test video are presented in Table 2. The sequence *Harbor* sequence, with a spatial resolution of 3840×1920, was used twice to verify the subjects' consistency.

Table 2. Spatial and temporal resolutions, and Q_p values, used on the subjective assessment of single distortions.

Sequence	Spatial Resolution (W× H)	Temporal Resolution [fps]	Q_p
PoleVault	960×480, 1920×960, 3840×1920	7.5, 10, 15, 30	15, 30, 35, 40, 45
Harbor	960×480, 1920×960, 3840×1920, 7680×3840		
SkateboardInLot			
ChairliftRide			
SkateboardTrick		7.5, 10, 15, 30, 60	
Train			

All sequences were encoded using the HEVC reference software [19], with a quantization parameter (Q_p) of 5, guaranteeing visually lossless encoded sequences. The videos temporal resolution were fixed at 30 fps; for the sequences with 60 fps, this was achieved by skipping one in every two frames. During the training session, five sequences were used, which result from down sampling the *ChairliftRide* full resolution sequence to 960×480, 1920×960, 2048×1024, 3840×1920 and 7680×3840 pixels. The assessment of spatial resolution impact on quality had a total of 33 sequences, including training, testing and repeated sequences. The final MOS values are presented in Figure 5.

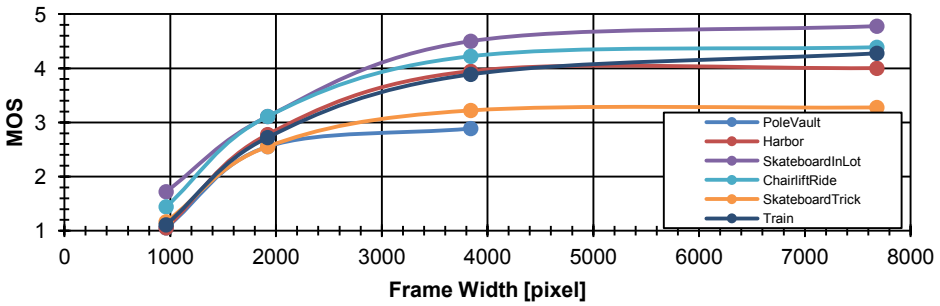


Figure 5. Resulting MOS values versus spatial resolution.

Both MOS and DMOS values increase with the videos spatial resolution, as expected. However, it is possible to observe a stagnation of the MOS values starting on 3840×1920, which may result from a possible limitation of the Oculus Rift technology to process spatial resolutions higher than 4K, and/or the limitation of the human visual system (HVS) to distinguish details when the resolution is very high.

4.5 Assessment of the frame rate impact on quality

For the subjective assessment of the frame rate impact on quality, the following temporal resolutions were considered: 7.5, 10, 15, 30 and 60 fps; Table 2 shows the used temporal resolutions per test video. The frame rate down sampling was done by skipping frames. The *Harbor* sequence at 15 fps, was used twice to verify the subjects' consistency. In all cases, the spatial resolution was kept at 3840×1920, so that all test videos could be used. In the training part, five sequences obtained from the full resolution *Train* sequence were used, each with one of the considered temporal resolutions. All sequences were visually lossless encoded using the HEVC standard (Q_p=5).

The whole test session had a total of 36 assessed sequences, including test, training and repeated sequences. The final MOS values are presented in Figure 6.

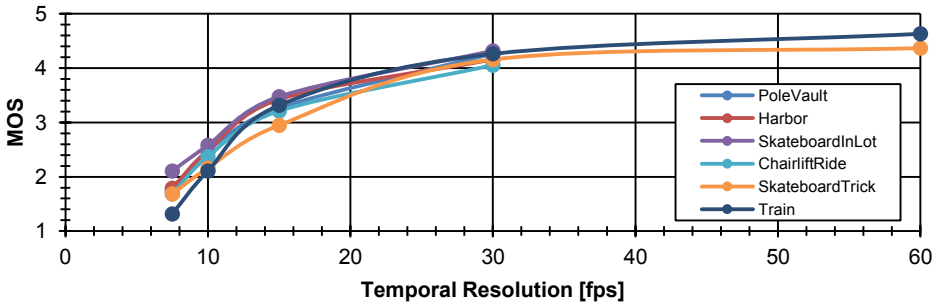


Figure 6. Resulting MOS values versus temporal resolution.

As expected, MOS (and also DMOS) increase with the temporal resolution, although not much from 30 fps to 60 fps. This stagnations might be due to the HVS limitation in distinguish temporal resolutions from a certain value.

4.6 Assessment of the quantization parameter impact on quality

For the assessment of the quantization parameter (Q_p) impact on quality, five Q_p values were selected, namely 15, 30, 35, 40 and 45, so that the resulting subjective scores may vary from very poor to very good. The *Harbor* sequence, with a Q_p of 35, was used twice to verify subjects' consistency. All the sequences were kept at their maximum spatial resolution, namely 3840×1920 for *PoleVault* and 7680×3840 for the rest of the videos, and with a temporal resolution of 30 fps. All sequences were encoded with the HEVC reference software, with the GOP structure having one I frame for fifteen B frames, constant Q_p and having the same value for all GoPs. For training, five sequences, resulting from the encoding of *ChairliftRide* with five different qualities, were used.

The assessment of the quantized parameter impact on quality had a total of 25 sequences, including training, testing and the repeated sequences. The final MOS values are presented in Figure 7.

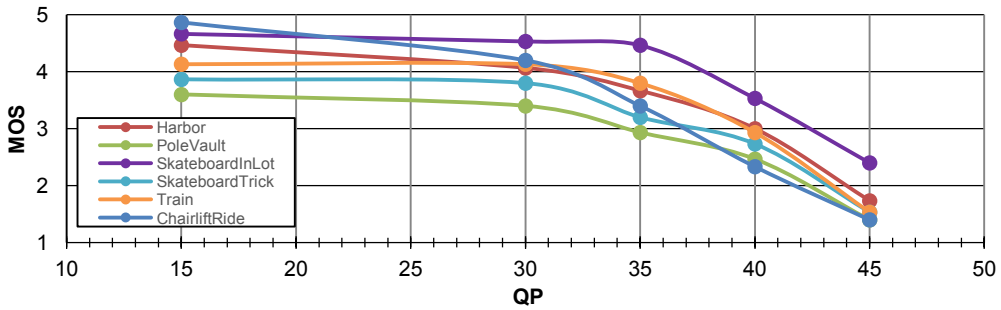


Figure 7. Resulting MOS values versus quantization parameter.

As expected, MOS (and DMOS) values decrease when Q_p increases, although for some sequences a stagnation of MOS values is observable for Q_p values below 30.

4.7 Assessment of combined distortions impact on quality

The videos selected for assessing the combined distortions effect were *Train*, *SkateboardTrick* and *SkateboardInLo*t, due to their good behavior on the temporal and spatial subjective quality assessment tests, resulting in MOS curves with smooth variations and with no, or minor, intersections between them. In terms of spatial resolution, four of the values used in section 4.4 were kept. Since the temporal resolution results showed too low MOS values for the 7.5 fps and 10 fps, these two temporal resolutions were not considered in this test. This choice came also from the fact that it was observed that low temporal resolutions are more likely to cause dizziness.

The selected spatial and temporal resolutions, and Q_p values, are present on Table 3. The sequence *Train,* with Q_p of 35, spatial resolution of 1920×960 and temporal resolution of 15 fps, was used twice to verify the subjects' consistency. The training included three sequences generated from *SkateboardInLot*, namely:

- *SkateboardInLot* with 960×480 pixels @ 15 fps and Q_p =30;
- *SkateboardInLot* with 1920×960 pixels @ 15 fps and Q_p =30;
- *SkateboardInLot* with 3840×1920 pixels @ 60 fps and Q_p =15.

Table 3. Selected combinations of spatial resolution, temporal resolution and Q_p.

		Spatial Resolution (W× H)			
		960 × 480	1920 × 960	3840 × 1920	7680 × 3840
		Q_p			
Frame rate [fps]	15	–	30, 35	30, 35	30
	30	30	30, 35	30, 35	30, 35
	60	–	–	15, 30	15

This assessment had a total of 25 sequences, including training, test and the repeated sequences. The final MOS values are presented on Figure 8. As shown, the quality range goes from very low quality to very high quality (MOS values from 1 to nearly 5).

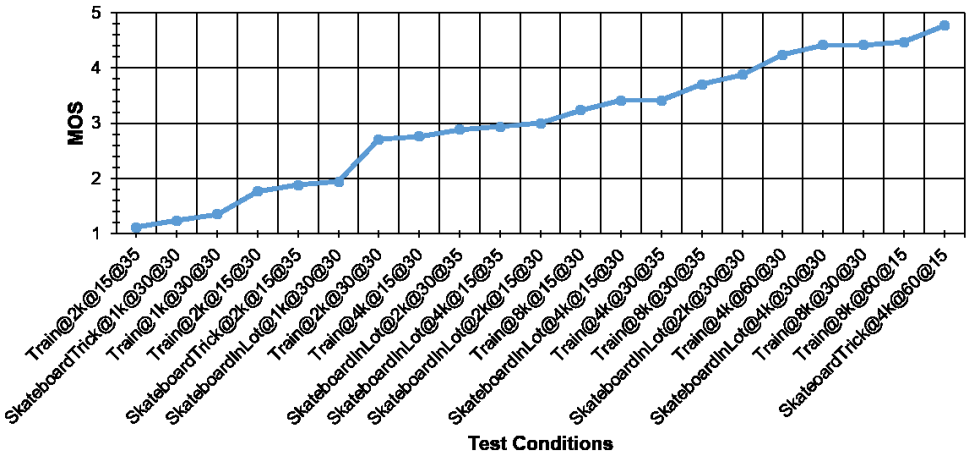


Figure 8. Resulting MOS values for the combined distortions.

5. OBJECTIVE METRICS EVALUATION

This section evaluates several objective quality assessment metrics for omnidirectional video, and when the video is subject to spatial subsampling and HEVC compression, applied as single distortions or in combination. The subjective quality assessment results, described in Section 4, are used as ground truth.

5.1 Considered objective quality metrics for omnidirectional video

To evaluate the impact of the used spatial resolution and quantization parameter (Q_p), on the perceived video quality, the following full-reference objective metrics were considered: SSIM, MS_SSIM, PSNR, SPSNR, WS_PSNR, VPSNR. These metrics can be divided in two groups: the conventional 2D metrics, like SSIM, MS_SSIM and PSNR, and the metrics developed specifically for omnidirectional images/video, like SPSNR, WS_PSNR and VPSNR. Additionally, the W_SSIM and WMS_SSIM metrics, proposed in Section 3 as an adaptation of SSIM and MS_SSIM to omnidirectional content, are also assessed. Although initially developed for quality assessment of images, all these metrics can also be extended to video quality assessment, by applying it to each video frame, followed by averaging the result over all frames. It is worth to mention that results from the temporal down sampling assessment test are not included, since none of the considered objective metrics are able to characterize the impact of the video temporal resolution.

5.2 Objective metrics evaluation procedure

To evaluate the quality of the aforementioned objective metrics, the DMOS values were used, since it measures the perceived quality difference between an impaired video and the reference one, and it is also in this way that the objective metrics work. Frequently, subjects experience difficulties in scoring video sequences with very high or very low quality,

leading to MOS or DMOS curves that present a flat behavior for very low and very high distortions. Therefore, the objective metric values are mapped in subjective scores using a logistic function, formally described by (14),

$$DMOSp = \beta_1 + \frac{\beta_2 - \beta_1}{1 + 10^{\beta_4(\beta_3 - m)}} \tag{14}$$

where m is the objective metric value, the β´s are the sigmoid function parameters obtained by curve fitting, and $DMOSp$ is the predicted DMOS value.

To evaluate the metrics, the following performance measures were used: Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SRCC) and Root Mean Square Error (RMSE).

5.3 Objective metrics behavior with spatial subsampling

To evaluate how well the objective metrics can predict the spatial resolution impact on quality, they were applied to the video sequences used in subsection 4.4. As mentioned in that subsection, the test sequences were generated by downscaling the original video sequences to spatial resolutions of 7680×3840, 3840×1920, 1920×960 or 960×480 pixels. To apply the metrics, the downscaled sequences were up sampled to their original spatial resolutions, using a Lanczos filter (with 6 taps) for interpolation.

Table 4 presents the result of the statistical measures, when applied between the DMOS values and the objective metric values (so, before applying the logistic) and also between DMOS and the predicted DMOS values (so, after applying the logistic). The *PoleVault* sequence was not considered in this comparison, since its maximum spatial resolution is 3840×1920 pixels.

Table 4. Objective assessment of the spatial subsampling impact on quality.

	SSIM	**MS_SSIM**	**W_SSIM**	**WMS_SSIM**	**PSNR**	**SPSNR**	**WS_PSNR**	**VPSNR**
Before logistic fitting (DMOS *vs.* objective metric value)								
PLCC	0.93	0.94	**0.96**	0.95	0.95	0.94	0.95	0.85
SRCC	0.89	0.90	0.89	**0.92**	0.88	0.91	0.90	0.80
After logistic fitting (DMOS *vs.* $DMOS_p$)								
PLCC	0.95	0.95	0.96	0.95	0.96	**0.97**	**0.97**	0.88
SRCC	0.89	0.90	0.89	**0.92**	0.88	0.91	0.90	0.80
RMSE	0.37	0.36	0.30	0.41	0.31	**0.29**	**0.29**	0.55

As can be seen from Table 4, both W_SSIM and WMS_SSIM present a better performance (although slightly) than the traditional SSIM and MS_SSIM. The WS_PSNR and SPSNR metrics show a better performance than PSNR, more visible in the SRCC measure. Overall, and excluding the VPSNR case, all metrics have similar behaviors. The lower PLCC and SRCC for VPSNR can be justified by the fact that, in this case, the PSNR is being applied to a video region that was not always the region observed, and evaluated, by the subjects. For a visual evaluation, Figure 9 (at the end of the paper) presents the DMOS versus objective scores for all metrics, and where the resulting logistic functions were also drawn.

5.4 Objective metrics behavior with HEVC compression

To evaluate how well the objective metrics can predict the compression impact on quality, the metrics were applied to the video sequences used in subsection 4.5. As mentioned in that subsection, the test sequences were generated by encoding the original ones with the HEVC reference software, with Q_p of 15, 30, 35, 40 and 45, a GoP structure with one I frame followed by fifteen B frames and Q_p constant inside each GoP, and having the same value in all GoPs. The reference sequences were considered to be the ones with Q_p of 15. *PoleVault* was also removed from this study due to its lower spatial resolution.

Table 5 presents the result of the statistical measures, between the DMOS values and the objective metric values (so, before applying the logistic fitting) and also between DMOS and the predicted DMOS values (so, after applying the logistic fitting).

Table 5. Objective assessment of the Q_p impact on quality.

	SSIM	**MS_SSIM**	**W_SSIM**	**WMS_SSIM**	**PSNR**	**SPSNR**	**WS_PSNR**	**VPSNR**
Before logistic fitting (DMOS *vs.* objective metric value)								
PLCC	0.95	0.94	**0.98**	0.96	0.95	0.93	0.94	0.83
SRCC	0.95	0.91	**0.98**	0.95	0.95	0.95	0.96	0.85
After logistic fitting (DMOS *vs.* $DMOS_p$)								
PLCC	0.96	0.95	**0.99**	0.98	0.97	0.95	0.95	0.85
SRCC	**0.96**	0.91	**0.96**	0.94	0.95	0.95	0.94	0.86
RMSE	0.28	0.31	**0.17**	0.22	0.27	0.31	0.31	0.53

In this case, the W_SSIM and WMS_SSIM metrics show a better performance when compared to the conventional SSIM and MS_SSIM metrics. The multi-scale options do not bring any improvement relatively to the single scale versions. The WS_PSNR and SPSNR shows a slightly lower performance than the PSNR. Overall, and excluding the VPSNR case, all metrics show a good performance. For a visual evaluation, Figure 10 (at the end of the paper) presents the DMOS versus objective scores for all metrics, and where the resulting logistic functions were also plotted.

5.5 Objective metrics behavior with joint HEVC compression and spatial subsampling

The objective metrics were also evaluated when compression and spatial subsampling were simultaneously applied to the videos. The original videos were subsampled spatially, encoded with a certain Q_p and upsampled spatially to the original spatial resolution. The test conditions considered in this comparison are presented in Table 6, whose subjective scores were obtained from experiments described in Sections 4.6 and 4.7. All sequences have a frame rate of 30 fps.

Table 6. Test conditions for joint HEVC compression and spatial subsampling.

	PoleVault	***Harbor***	***Skateboard InLot***	***ChairliftRide***	***Skateboard Trick***	***Train***
Spatial Resolution (W × H)	**Q_p**					
960 × 480	-	-	30	-	30	30
1920 × 960	-	-	30,35	-	-	30
3840 × 1920	30,35,40,45	-	30	-	-	35
7680 × 3840	-	30,35,40,45	30,35,40,45	30,35,40,45	30,35,40,45	30,35,40,45

Table 7 presents the result of the statistical measures, between the DMOS values and the objective metric values (before the logistic fitting) and also between DMOS and the predicted DMOS values (after the logistic fitting).

Table 7. Objective assessment of the joint Q_p and spatial subsampling impact on quality.

	SSIM	**MS_SSIM**	**W_SSIM**	**WMS_SSIM**	**PSNR**	**SPSNR**	**WS_PSNR**	**VPSNR**
Before logistic fitting (DMOS *vs.* objective metric value)								
PLCC	0.72	0.89	0.77	**0.90**	0.78	0.79	0.77	0.68
SRCC	0.83	0.88	0.86	**0.91**	0.81	0.82	0.82	0.69
After logistic fitting (DMOS *vs.* $DMOS_p$)								
PLCC	0.82	0.90	0.85	**0.92**	0.80	0.81	0.80	0.68
SRCC	0.83	0.88	0.84	**0.90**	0.81	0.82	0.80	0.69
RMSE	0.55	0.43	0.51	**0.38**	0.58	0.57	0.59	0.71

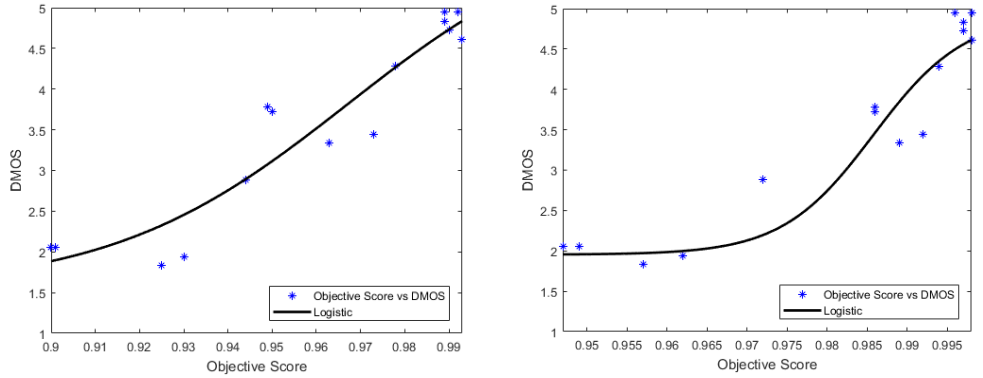
As in the previous cases, the W_SSIM and WMS_SSIM perform better than SSIM and MS_SSIM respectively. However, the only metrics having good performance are MS_SSIM and WMS_SSIM. This comes from the fact that in this case, there are two mixed effects: the change in quality due to compression and due to the variation of spatial resolution. In fact, the multi scale options measures the quality impairment in different spatial resolutions, which may explain its advantage against the others. For a visual evaluation, Figure 11 (at the end of the paper) presents the DMOS versus objective scores for all metrics, and where the resulting logistic functions were also drawn.

As can be concluded from the three analysis, with exception for VPSNR all the considered objective metrics perform quite well for each individual impairment. However, when the two considered impairments are combined, only the MS_SSIM and WMS_SSIM metrics achieve acceptable PLCC, SRCC and RMSE values.

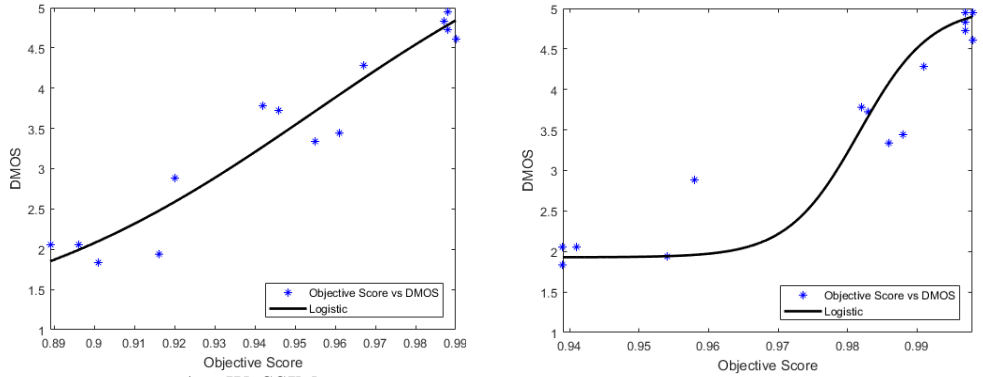
6. CONCLUSIONS

This paper evaluated the impact, on perceived quality, of HEVC compression, spatial subampling and temporal subsampling of omnidirectional videos, when the omnidirectional videos are displayed in a head mounted device (HDM). The subjective assessment results were then used as ground-truth data to evaluate conventional quality assessment metrics developed for 2D video, as well as some of the recently proposed metrics for omnidirectional video, namely S-PSNR, WS-PSNR, and VP-PSNR. The adaptation of two SSIM based metrics, to omnidirectional contents, were also proposed and assessed. The assessment of the metrics showed that the visual impact caused by spatial down sampling or HEVC compression, when applied as single distortions, can be well estimated by most of them (the exception was the VPSNR metric), with some advantage for the two metrics proposed in the paper, namely W-SSIM and WMS-SSIM. However, when both distortions are jointly applied to the videos, only the MS_SSIM and the WMS_SSIM metrics presented acceptable performance.

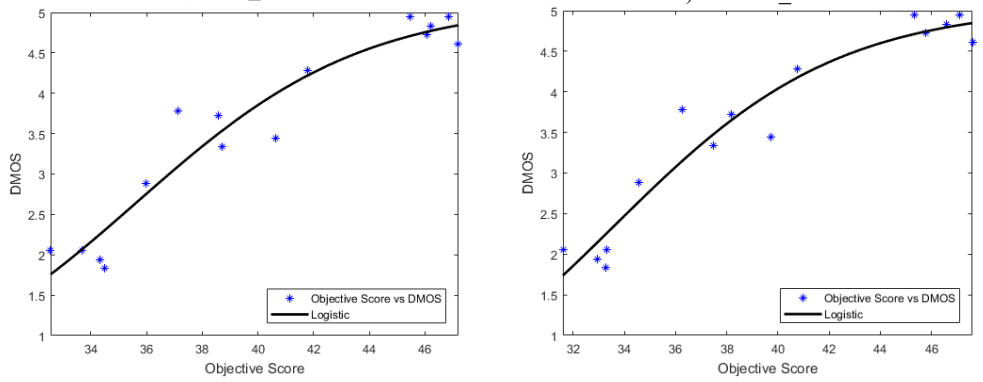
None of the considered quality metrics are able to characterize the impact of the video temporal resolution. However, the subjective assessment tests showed that the video frame rate may play an important role in the user QoE. How to objectively quantify the decrease in perceived quality due to temporal subsampling is an ongoing work, and will be reported in a future publication.



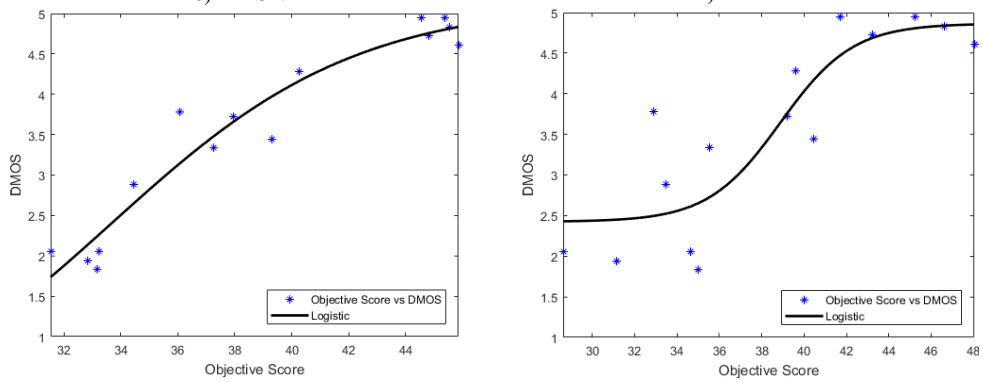
a) SSIM b) MS_SSIM



c) W_SSIM d) WMS_SSIM



e) PSNR f) SPSNR



g) WS_PSNR h) VPSNR

Figure 9. DMOS versus objective scores for the considered objective metrics when the videos are subjected to spatial subsampling.

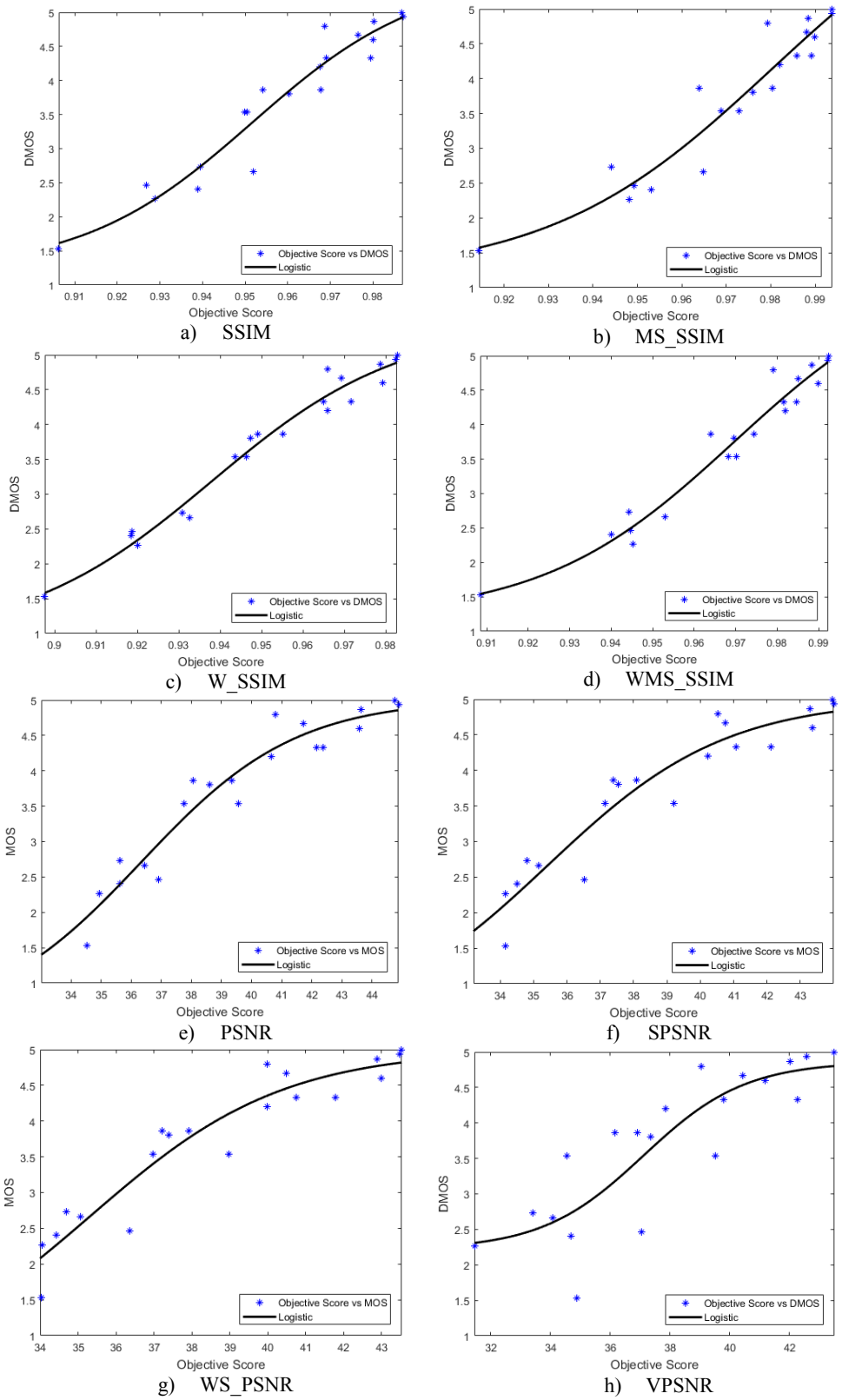
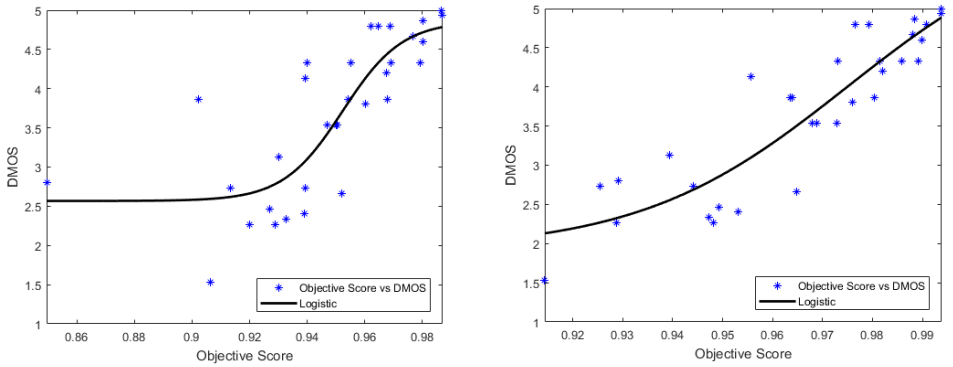
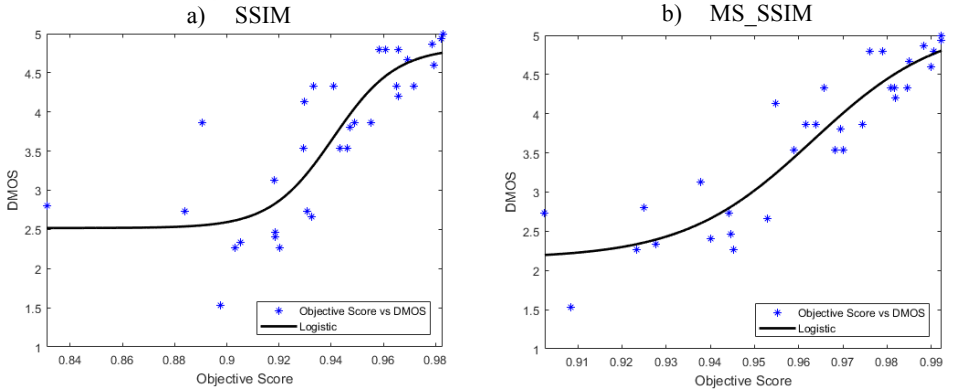


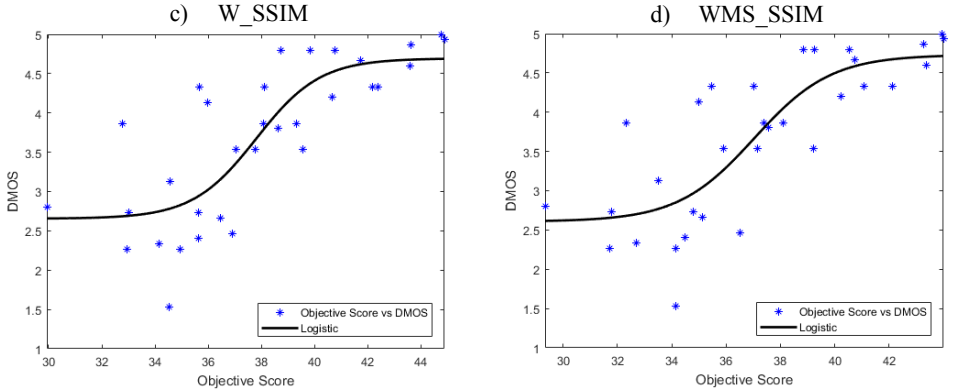
Figure 10. DMOS versus objective scores for the considered objective metrics when the videos are subjected to HEVC compression.



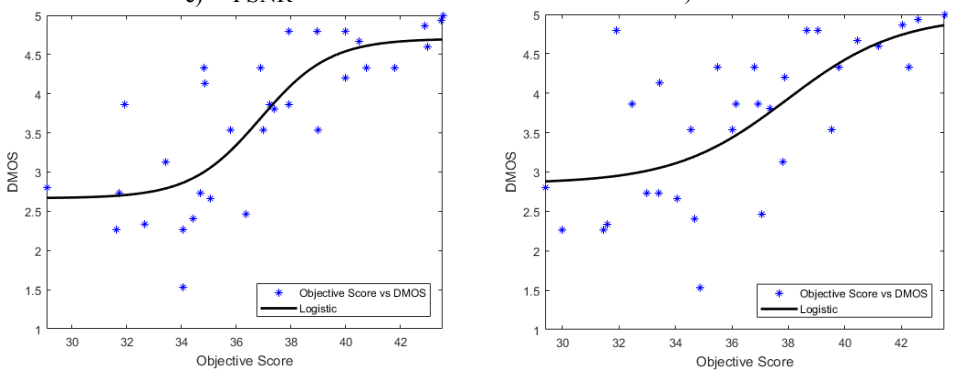
a) SSIM b) MS_SSIM



c) W_SSIM d) WMS_SSIM



e) PSNR f) SPSNR



g) WS_PSNR h) VPSNR

Figure 11. DMOS versus objective scores for the considered objective metrics when the videos are subjected to both spatial subsampling and HEVC compression.

ACKNOWLEDGMENTS

This work was funded by Instituto de Telecomunicações and Fundação para a Ciência e a Tecnologia (FCT) under project UID/EEA/50008/2013, and by project MESMOQoE (no. 023110 - 16/SI/2016) supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

REFERENCES

[1] Simone, F.D., Frossard, P., Brown C., Birkbeck, N. and Adsumilli, B., "Omnidirectional video communications: new challenges for the quality assessment community," VQEG eLetter, 3(1), (2017).
[2] Yu, M., Lakshman, H., and Girod, B., "A framework to evaluate omnidirectional video coding schemes," IEEE Int. Symp. on Mixed and Augmented Reality, Japan, (2015).
[3] Sun, Y., Lu, A. and Yu, L., "WS-PSNR for 360 video objective quality evaluation", Doc. JVET-D0040, (2016).
[4] Vishwanath, B., He, Y., and Ye Y., " Area Weighted Spherical PSNR for 360 video quality evaluation," JVET-D0072, (2016).
[5] Upenik, E., Rerabek, M. and Ebrahimi, T., "On the performance of objective metrics for omnidirectional visual content", Proc. of Int. Conf. on Quality of Multimedia Experience (2017).
[6] https://github.com/jascenso/ISTOmnidirectionalDataset
[7] Wang, Z., Bovik, A., Sheikh, H. and Simoncelli, E., "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. on Image Processing, 13(4), (2004).
[8] Wang Z., Simoncelli, E., and Bovik, A., "Multi-scale structural similarity for image quality assessment", Proc. of the 37th IEEE Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, USA (2003).
[9] Samsung, "Gear VR," [Online]. Available: http://www.samsung.com/global/galaxy/gear-vr/. [Accessed 7 March 2017].
[10] B. Choi, W. Ye-Kui and M. M. Hannuksela, "WD on ISO/IEC 23000-20 Omnidirectional Media Application Format," WD on ISO/IEC 23000-20 Omnidirectional Media Application Format, Geneva, Switzerland, (2016).
[11] Samsung, "Samsung Gear 360 2017," Samsung Electronics CO., LTD., [Online]. Available: http://www.samsung.com/global/galaxy/gear-360/. [Accessed 6 April 2017].
[12] Szeliski, R., "Image Alignment and Stitching: A Tutorial", Foundations and Trends in Computer Graphics and Vision, 2(1), (2006)
[13] Wiegand, T., Sullivan, G., Bjøntegaard, G. and Luthra, A., "Overview of the H.264/AVC video coding standard", IEEE Trans. on Circuits and Systems for Video Technology, 13(7), (2003)
[14] Sullivan, G., Ohm, J.-R., Han, W.-J., and Wiegand, T., " Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Trans. on Circuits and Systems for Video Technology, 22(12), (2012)
[15] ITU-T, "P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," (2016).
[16] Oculus, "Oculus Rift," Oculus VR, LLC, 2017. [Online]. Available: https://www.oculus.com/rift/. [Accessed 2017 May 18].
[17] TU-R, "BT-500-13: Methodology for the subjective assessment of the quality of television pictures," (2012).
[18] Boyce, J., Alshina, E., Abbas, A., and Ye, Y., "JVET common test conditions and evaluation procedures for 360° video," in Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-1030, (2016).
[19] "High Efficiency Video Coding (HEVC) reference software," Fraunhofer Heinrich Hertz Institute, 2015. [Online]. Available: https://hevc.hhi.fraunhofer.de/