Perceptual Quality and Bit Rate Models for Omnidirectional Video

Francisco Lopes
Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
francisco.garcao.lopes@tecnico.ulisboa.pt

***Abstract*—The immersion sensation provided by 360° videos gives users a completely different user experience of performing 2D videos, forcing a validation on the use of traditional assessment metrics in the omnidirectional context. In fact, several 2D inherited objective metrics have been used, some adapted to the reality of omnidirectional video. This work, based on subjective tests performed to measure the subjective visual impact of spatial/temporal resolutions and compression variations in 360° videos, evaluates the performance of some objective metrics commonly used in the literature, proposing an evolution of a 2D metric adapted to 360° video. In addition, a set of metrics based on uncompressed video for subjective quality prediction are tested, following the considerations of several works available in the literature for 2D video. These metrics are video content dependent and can be obtained with linear combinations of original videos characteristics. Aiming at selecting video representations (combinations of temporal/spatial resolution and compression) to be stored on a multimedia server and following a predictive context, a procedure similar to that of uncompressed quality prediction is done to predict bit rate, considering the impact of spatial/temporal resolution and compression variations. The selection algorithm is also combined with the expected quality for each representation. After an evaluation of the predictive metrics based on characteristics of the original videos, it was possible to observe, as proved for 2D video, that it can be used in the context of omnidirectional video, presenting metrics with high correlations with the experienced subjective perceptions and real bit rates.**

***Index Terms*—360° Video, Bit Rate, Omnidirectional, Predictive Models, Quality Assessment, Quality of Experience.**

I. Introduction

Nowadays, omnidirectional video, or simply 360° video is gaining more and more importance. The sensation of immersiveness given by these videos, pictures or games creates a very different user experience than the one provided by traditional 2D video, putting the user at the center of the action. Generating omnidirectional content results in videos or contents of high bit rate. Of course, in contexts of finite bandwidth and even sometimes reduced, like in the case of mobile network where the internet's bandwidth oscillates a lot, a tradeoff between quality and bit rate that can be provided to service subscribers appears. Thus, quality experience assessment metrics adapted for 360° video have been developed, many of them as adaptations of 2D conventional metrics such as Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity Index (SSIM). In this way, a combination of quality prediction and bit rate metrics can help in deciding which video representations should be made available by content providers. Maximizing quality while minimizing the bit rate.

A lot of work have been being done for quality of experience (QoE) and bit rate prediction for 2D video, with some models fully based on uncompressed video features, allowing to predict quality and bit rate without make any compression or other transformation to original videos. This paper provides a comparative study of traditional quality assessment metrics as well as metrics designed specifically for 360° video to conclude about the correlation between them and the human visual system (HVS) in the 360° video context. From this study arise two new metrics based on SSIM and adapted to 360° video. Also this work shows that with uncompressed video features it is possible to generate models for subjective quality and bit rate prediction, presenting an algorithm that selects the best video representations to be stored in a streaming server.

This paper is organized as follows: Section II makes an overview on some 360° video basic principles, as well as some 360° video streaming solutions. Section III presents the main image impairments and distortion causes and then approaches some subjective and objective quality assessment metrics designed for image and video assessment. Section IV describes the procedures used to access 360° video subjectively with a group of 360° videos which were subject to transformations in spatial/temporal resolution and Q_p, as well as the results of four performed subjective test sessions. Section V assesses the traditional video quality metrics using the subjective results obtained previously. The assessment of the different metrics is presented and discussed. Then, following some state-of-the-art quality prediction metrics developed for 2D video, is proposed a quality prediction metric based solely on uncompressed video and designed for 360° video. Section VI describes the testing of a proposed model from the literature for bit rate prediction considering uncompressed 360° video. Then, it is presents an algorithm for best representation video selection, following some state-of-the-art solutions, considering the quality and bit rate prediction model developed before. Section 0I concludes the work, highlighting the most important conclusions and making future work suggestions.

II. Omnidirectional Video: Basics and Key Processing Solution

360° video is a spherical video format that stores visual information from all viewing directions. This allows a user to navigate in all viewing directions. It is common to wear Head Mounted Devices (HMD) which are special glasses that put the user in the center of the viewing sphere, and can change the direction of view simply by moving the head. This section introduces the principles of 360° video. The transmission chain of 360° video presented in Fig. 1, can be divided in a sequence of modules.

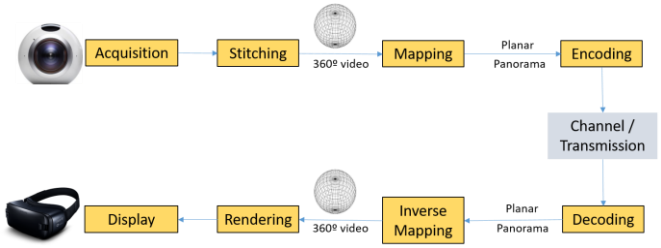

Fig. 1. Generic architecture of an omnidirectional video transmission chain [12] [13] [14]

The main functions of each module are:

- **Acquisition:** The acquisition of 360° 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 area, so that each camera acquires a 2D image corresponding to a portion of the spherical view around it.
- **Stitching:** After the acquisition, the 2D images recorded by each camera are fused to create an 360° frame, containing the 360º information of the scene with a process called *stitching* [1].
- **Mapping:** To be transmitted, the spherical 360° frame is mapped into a planar representation by a process called mapping. The most used planar projection is the equirectangular projection (ERP).
- **Encoding:** Since a planar representation was obtained in the previous step a 2D video codec can be used. In most cases, before encoding there is an additional step called tilling that divides the 360° frame into several tiles which are independently encoded.
- **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.
- **Decoding:** The decoding step of this processing chain performs the inverse operation of the encoder and at the end the reconstructed 360° video is obtained.
- **Inverse Mapping:** When 360° video is rendered, a spherical representation is often used. Therefore, it is needed to map the planar 360° video into a sphere, by applying the corresponding inverse mapping transformation of the sender.
- **Rendering:** In 360° video, the images that are presented to the user are a part of the entire viewing sphere. A selected part of the sphere is projected on a 2D plane which is 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 360° video are of two types: the first corresponds to a navigable image on, a standard 2D display, where the viewing direction can be controlled by mouse or by moving the display and the second type corresponds to a HMD, which tracks user's head movements to compute the corresponding viewport.

III. State-of-the-art on Quality Assessment

After being acquired, an 360° video signal undergoes several transformations till it is displayed in the viewer's screen, and some of these transformations introduce artifacts on the video. To assess the artifacts impact of the video quality, several objective metrics have been developed, which must be adapted to the HVS, to achieve a good match between the predicted quality and the subjective quality. This section presents some methodologies for subjective and objective quality assessment of 360° video. In this work, the objective quality assessment metrics used are compared with the result of subjective tests using the Pearson Linear Correlation Coefficient (PLCC), the Spearman Rank Correlation Coefficient (SRCC) and the Root Mean Square Error (RMSE).

A. *Objective Quality Assessment Methodologies*

The differences between conventional 2D video and 360° video are vast and so dedicated metrics for objective quality assessment applied for spherical surfaces were created, many as evolutions of conventional 2D metrics like PSNR. Besides, the study in [2] reveals that not all the sphere locations are equally observed by the users. Fig. 2 presents a heat map with the locations of the sphere accessed by users, parametrized by the latitude and longitude on the sphere, showing that the users motion observations are mostly concentrated in the main viewport.

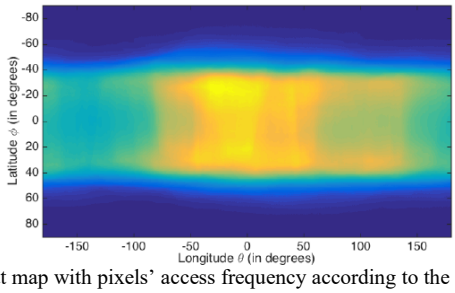

Fig. 2. Heat map with pixels' access frequency according to the head motion trajectories [2].

Some of the mostly used objective metrics for 360° video are described next [2]:

- **Viewport PSNR:** In VPSNR, only the image pixels belonging to a viewport are subject of a traditional PSNR evaluation, that is applied between corresponding viewports of the reference and impaired videos.
- **Spherical PSNR:** In SPSNR, for each image sample from a total of 655362 on the spherical video, the correspondent image samples on the reference and impaired planar

projections are firstly obtained; then, the PSNR is applied between them.

- **Weighted to Spherical PSNR:** WS-PSNR weighs the pixel error, computed between pixels on the reference and impaired 2D images, by the corresponding pixel area on the spherical surface. In fact, when mapping from the 2D representation space to the spherical observation space, pixels are stretched or condensed at various level according to its location on projection plane. To measure the stretchiness, the luminance for each pixel in the position (i, j) of the frame is weighted by a factor that depends on a scaling factor. For example, for the ERP the scaling factor is given by:

$$w(i,j) = \cos((i - \frac{height}{2} + 0.5) \times \frac{\pi}{height}) \quad (1)$$

- **Latitude SPSNR:** 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 and in the back, almost never accessed.

B. *Quality Prediction Based on Video Features*

Traditional objective metrics obligate to create the impaired sequences before knowing its quality level. To address this issue, some metrics have been developed, aimed at predicting 2D video quality based only on some intrinsic characteristics of the original video, as its spatial and temporal activities. In [3] the authors found possible to predict the subjective impact of reducing the frame rate, using an objective metric called Temporal Correction Factor (*TCF*) based only on uncompressed video features and obtained from subjective tests. Further, in [4], the same authors proved to be possible to combine independent quality assessment metrics (like *TCF* but for spatial resolution and Q_p) in one, capable of assessing the combined effect of temporal/spatial and Q_p variations.

IV. Subjective Quality Assessment of Omnidirectional Video

This sections describes the procedures followed on the subjective assessment of 360° videos, and presents and analyses the subjective assessment results.

A. *Subjective Assessment Framework*

Fig. 3 presents, schematically, the subjective assessment framework architecture used. 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.

B. *Subjective Tests Procedures*

The use of HMD might induce tiredness and queasiness caused by long time exposure. Therefore, the subjective tests duration should be rather short. Two test sessions were made. The first session included two stages: the quality assessment of the spatial resolution and the quality assessment of the temporal resolution impact on quality.

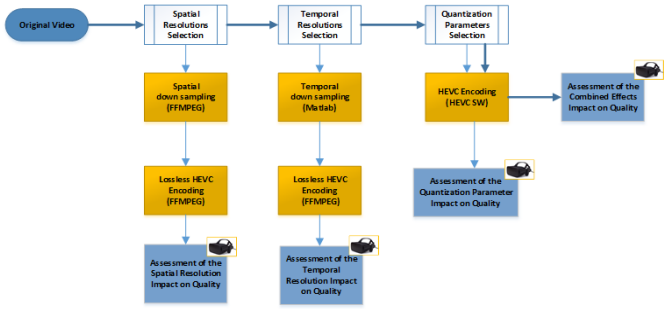

Fig. 3. Subjective Assessment Framework Architecture

The second session included also two stages: the quality assessment of the compression impact and the quality assessment of the combined effects impact on quality. Thus, single stimulus assessment methodologies were selected, namely, the Absolute Category Rating with Hidden Reference (ACR-HR) in the first three subjective test groups and the Absolute Category Rating (ACR) in the fourth test [5]. In ACR-HR the reference stimuli are also assessed, but the subject does not know which test video it is, while in ACR the reference stimuli are not assessed. In the three first stages, subjects were presented the reference video sequences. These correspond to the video sequences that present the minimal distortion of a certain type. In the fourth stage, however, the reference sequences were not presented.

The HMD used was Oculus Rift [6], running under the Oculus Software, with the GoPro VR Player 2.3 application as VR media player. The graphic card used, GEFORCE GTX 1060 3 GB, can display videos with up to 7680×4320@60 of resolution. For this reason, the maximum spatial resolution considered in the tests was 7680×3840. Videos with a higher original resolution were down sampled to 7680×3840 and this resolution was considered their maximum spatial resolution. A swivel chair, that allows subjects to freely move and explore the entire 360° view while sitting, was used.

At the beginning of each test, and without using the HMD, the subjects were introduced to the objectives of the test session. Then, with the HMD put on, 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. During the test session, each test video was displayed for 10 seconds. After that, a still image with the evaluation scale (1 to 5) was also displayed during 10 seconds, where the subject should inform the score of the previous video sequence. The score was then register by the test host. 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. Between the two test stages, subjects had to take off the HMD for one or two minutes to rest the eyesight and avoid dizziness.

The number of participants in the two sessions was, respectively, 20 and 17. Each session lasted around 25 minutes. Subjects were from both genders, mostly male subjects, with ages between 22 and 55 years old, and included experts and non-experts in the image processing research field.

C. Statistical Analysis and Validation

The results obtained in each test were validated by applying the procedure recommended in [7]. From the results obtained in the subjective assessment sessions, the initial MOS was calculated. Then, the kurtosis coefficient was computed to verify if the initial MOS follows a normal distribution. According to this result, the values of two counters P_i and Q_i defined in [7] were computed for outliers removal. Finally, the final MOS was calculated without the outliers. In addition to MOS it was also calculated the Differential MOS (DMOS)

D. Characterization of the Video Test Set

The video test set was chosen from a group of 10 YUV videos from the JVET dataset. These videos are in the 4:2:0 format and have a length of 10 seconds each. Table I summarizes their main characteristics. A sub-set was selected based on the videos spatial and temporal perceptual information, namely *SI* and *TI*, presented in [7]. Also, to adapt these metrics to 360° video, *TI* and SI were combined with (1) to give more importance to edges or pixel intensity differences in the center of the ERP.

After computing the *SI* and *TI* values for each video, these values were normalized relatively to the maximum SI and TI values in the video set, resulting in *NSI* and *NTI*, respectively. Table I presents the resulting *SI*, *TI*, *NSI* and *NTI* for each video in the set. The sequences to be used during the subjective tests were selected according to the following rational: the videos labelled as *e, f, g, h, i, j* were first chosen because they populate very well the *NTI* range of values and have the highest spatial resolution. Next, the video labelled as *d* was also included, since it is the one with the highest spatial information, populating in this way the *NSI* range of values.

TABLE I
VIDEO TEST DATABASE CHARACTERISTICS

Sequence	Label	Width [px]	Height [px]	#Frames	Frame Rate [fps]
AerialCity	a	3840	1920	301	30
DrivingInCity	b	3840	1920	301	30
DrivingInCountry	c	3840	1920	301	30
PoleVault	d	3840	1920	300	30
Harbor	e	8192	4096	300	30
KiteFlite	f	8192	4096	300	30
SkateboardInLot	g	8192	4096	300	30
ChairliftRide	h	8192	4096	300	30
SkateboardTrick	i	8192	4096	520	60
Train	j	8192	4096	600	60

TABLE II
VIDEOS AND CORRESPONDING SI, TI, NSI AND NTI VALUES

Sequence	SI	TI	NSI	NTI
AerialCity	39.89	3.71	0.71	0.27
DrivingInCity	45.24	9.56	0.81	0.69
DrivingInCountry	46.07	11.72	0.82	0.84
PoleVault	55.99	3.64	1.00	0.26
Harbor	27.74	1.95	0.50	0.14
KiteFlite	36.06	4.19	0.64	0.30
SkateboardInLot	25.42	13.88	0.45	1.00
ChairliftRide	22.51	7.16	0.40	0.52
SkateboardTrick	19.51	3.80	0.35	0.27
Train	20.90	9.27	0.37	0.67

E. Characterization of the Subjective Assessment Test Sessions

For the spatial resolution subjective test stage were used the spatial resolutions presented on Table III. The down sampling was applied using a Lanczos Filter included in the FFMPEG software. The temporal resolution was kept at a fixed value of 30 fps. For the sequences with 60 fps this was achieved by skipping 1 in every two frames in the YUV components.

For the temporal resolution subjective test stage were used the frame rates presented on Table IV. The frame rate down sampling was done by skipping frames in the YUV components. In all cases, the spatial resolution was kept at 3840×1920, so that all test videos could be used. In both subjective test stages, the test videos were encoded losslessly using the HEVC standard using the FFMPEG encoder [8].

For the Q_p subjective test stage were used the Q_p of 15, 30, 35, 40 and 45. After processing the data from the first session, *KiteFlite* was removed from the video test set due to the singular (and unusual) behavior that this sequence showed in the previous two stages. All the sequences were kept at their maximum spatial resolution, namely 3840×1920 for *d* 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 [9], with the GoP structure having one I frame for fifteen B frames and Q_p constant for all GoPs. The set of spatial/temporal resolutions and Q_p frame were found to be representative of the 5 quality perceptual levels.

The videos selected for assessing the combined distortions effect were *Train*, *SkateboardTrick* and *SkateboardInLot*, 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. The test conditions are presented in Table V. In terms of spatial resolution, the four values of Table III were kept. Since the temporal resolution results showed too low MOS values for the 7.5 and 10 fps, these temporal resolutions were not considered fir this test. In terms of Q_p, was used the set from 15 to 35, since above that value the perceptual quality was found to be too low.

TABLE III
CHOSEN SPATIAL RESOLUTIONS

Sequence	Spatial Resolution
d	960×480, 1920×960, 3840×1920
e,f,g,h,i,j	960×480, 1920×960, 3840×1920, 7680×3840

TABLE IV
CHOSEN TEMPORAL RESOLUTIONS

Sequence	Temporal Resolutions
d, e, f, g, h	7.5, 10, 15, 30
i, j	7.5,10,15, 30,60

TABLE V
SELECTED COMBINATIONS OF SPATIAL/TEMPORAL RESOLUTION AND Q_p

Q_p		Spatial Resolution			
		960 × 480	1920 × 960	3840 × 1920	7680 × 3840
Frame Rate	15	–	30, 35	30, 35	30
	30	30	30, 35	30, 35	30, 35
	60	–	–	15, 30	15

F. Subjective Assessment Results

The results of the four subjective test stages are presented on Fig. 4 and Table VI. As expected, the MOS grows with the spatial/temporal resolution and decreases with the Q_p, with the MOS presenting a global stagnation more or less accentuated

between the 4 and 5 quality levels. Some remarks can be done:

- In the spatial resolution subjective test results it is possible to observe a stagnation of the MOS results starting on 3840×1920, which may result from a possible limitation of the Oculus Rift technology to present spatial resolutions higher than 4K or due to a limitation of the HVS to distinguish details when the resolution is very high.
- In the Q_p subjective test results, *ChairliftRide* does not present a stagnation between MOS of 4 or 5, which might be explained by the fact that that was the video used in the training, which could have induced subjects to be more observant to the sequence and thus.

V. Objective Quality Assessment of Omnidirectional Video

This section presents and evaluates objective quality assessment metrics for omnidirectional video, to be applied when the video is subject to the distortions considered before.

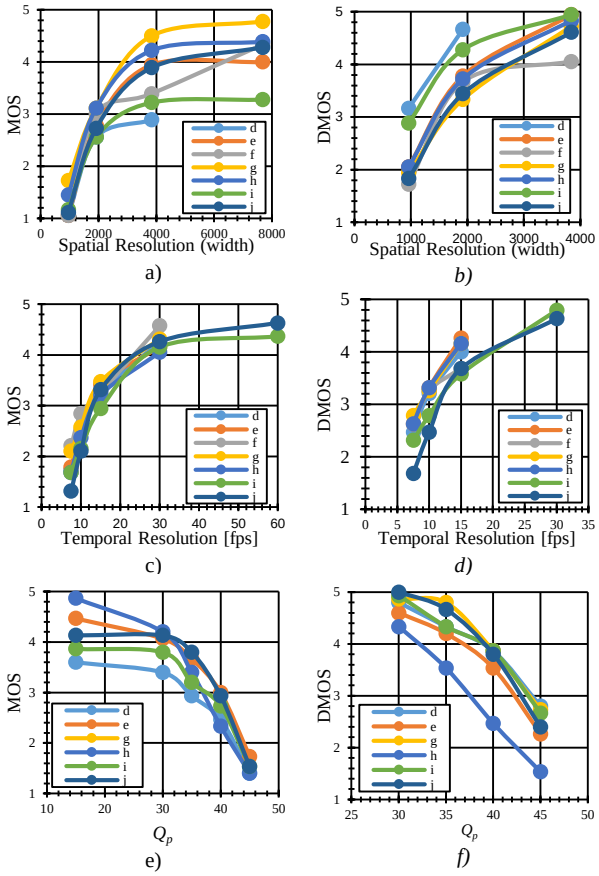

Fig. 4. Left column: Final MOS; Right Column: Final DMOS. a) and b) refer to the spatial resolution subjective quality assessment stage, c) and d) refer to the temporal resolution subjective quality assessment stage and e) and f) refer to the Q_p subjective quality assessment stage.

A. Evaluation of 2D Conventional and Omnidirectional Video Quality Metrics

To evaluate the impact of the used spatial resolution and Q_p, on the perceived video quality, the metrics SSIM, Multi-Scale SSIM (MS-SSIM), PSNR, SPSNR, W-PSNR and VPSNR were used.

TABLE VI
Final MOS of the combined effects impact test.

Sequence	Spatial Resolution	Temporal Resolution	Q_p	MOS
Train	1920 × 960	15	35	1.12
SkateboardTrick	960 × 480	30	30	1.24
Train	960 × 480	30	30	1.35
Train	1920 × 960	15	30	1.76
SkateboardTrick	1920 × 960	15	35	1.88
SkateboardInLot	960 × 480	30	30	1.94
Train	1920 × 960	30	30	2.71
Train	3840 × 1920	15	30	2.76
SkateboardInLot	1920 × 960	30	35	2.88
SkateboardInLot	3840 × 1920	15	35	2.94
SkateboardInLot	1920 × 960	15	30	3.00
Train	7680 × 3840	15	30	3.24
Train	3840 × 1920	30	35	3.41
SkateboardInLot	3840 × 1920	15	30	3.41
Train	7680 × 3840	30	35	3.71
SkateboardInLot	1920 × 960	30	30	3.88
Train	3840 × 1920	60	30	4.24
Train	7680 × 3840	30	30	4.41
SkateboardInLot	3840 × 1920	30	30	4.41
Train	7680 × 3840	60	15	4.47
SkateboardTrick	3840 × 1920	60	15	4.76

Additionally, two new metrics are also considered, resulting from adaptations of SSIM and MS-SSIM to omnidirectional content:

1) W-SSIM

In this case, for an impaired frame F and the unimpaired frame F^R at time n, SSIM is computed block, b by block of 11×11 pixel. It is then applied the scaling correction factor (1) to the SSIM value of each block, generating $SSIM_b$, where i is the vertical axis coordinate of the block middle pixel, and then all the weighted SSIM relative to blocks are summed and divide by the sum of all frame weights to generate the W-SSIM value for frame n (3). Finally, W-SSIM is the mean of the frame W-SSIM over all frames (4). In (2) l, c and s stand for the luminance, contrast and structure respectively and α, β and γ define the relative importance of each component. This way it is given more importance to the SSIM values near the center of the frame instead of near the poles or back.

$$SSIM_{b,n}(F_n, F_n^R) = [l(F_n, F_n^R)]_b^\alpha \times [c(F_n, F_n^R)]_b^\beta \times [s(F_n, F_n^R)]_b^\gamma \quad (2)$$

$$W\_SSIM_n = \frac{\sum_{b=1}^{B} SSIM_{b,n}(F_n, F_n^R) \times w_b(i,j)}{\sum_{b=1}^{B} w_b} \quad (3)$$

$$W\_SSIM = \frac{W\_SSIM_n}{N} \quad (4)$$

2) WMS-SSIM

To MS_SSIM is applied a similar procedure. However, in this case, the multiplication of the scaling factors is applied twice, in the luminance and in the combination of structure with contrast component (5) generating WMS_SSIM for a certain block b at time n (6). The rest of the procedure is equal to the one explained for W_SSIM. In (5), A is the number of time that block b is spatially down-sampled.

$$M = \prod_{a}^{A} [c_a(F_n, F_n^R)]_b^{\beta_a} \times [s_a(F_n, F_n^R)]_b^{\gamma_a} \times w_b(i,j) \quad (5)$$

$$WMS\_SSIM_{b,n}(F_n, F_n^R) = [l(F_n, F_n^R)]_b^\alpha \times w_b(j) \times M \quad (6)$$

3) Objective Metrics Assessment

Table VII presents the result of the metrics referred before, between the DMOS values obtained from the spatial

resolution subjective test stage modeled with the logistic (7) and the objective metric computed for the same sequences, considering the 7680x3840 as reference. Note that *PoleVault* was not considered in this comparison, since its maximum spatial resolution is 3840×1920.

$$DMOSp = \beta_1 + \frac{\beta_2 - \beta_1}{1 + 10^{\beta_4(\beta_3 - m)}} \quad (7)$$

Table VIII presents the result of the same metrics, between the DMOS values obtained from the Q_p subjective test stage modeled with the logistic (7) and the objective metric computed for the same sequences, considering a reference Q_p of 15. *PoleVault* was also removed from this study due to its lower maximum spatial resolution. The objective metrics were also evaluated when compression and spatial down sampling are simultaneously applied to the videos. The original videos were down sampled spatially, encoded with a certain Q_p and up sampled spatially to the original spatial resolution. Table IX presents the result of the same metrics, between the DMOS values obtained from the Q_p and combined effects subjective test stages with the logistic (7) and the objective metric computed for the same sequences. In this case only test sequences with a frame rate of 30 fps were used. Some remarks can be done:

- In the spatial resolution and Q_p cases, all metrics perform a good match with the subjective scores. Also, W-SSIM and WMS-SSIM perform slightly better than SSIM and MS-SSIM.
- In the combination of spatial resolution and Q_p case, again W-SSIM and WMS-SSIM perform slightly better than SSIM and MS-SSIM. However, only the multi-scaled metrics present acceptable values of correlation with the subjective scores, probably due to the fact that video is assessed in different resolutions.

TABLE VII
SPATIAL RESOLUTION OBJECTIVE METRICS ASSESSMENT

	SSIM	MS_SSIM	W_SSIM	WMS_SSIM	PSNR	SPSNR	WS_PSNR	VPSNR
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

TABLE VIII
Q_p OBJECTIVE METRICS ASSESSMENT

	SSIM	MS_SSIM	W_SSIM	WMS_SSIM	PSNR	SPSNR	WS_PSNR	VPSNR
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

TABLE IX
SPATIAL RESOLUTION PLUS Q_P OBJECTIVE METRICS ASSESSMENT

	SSIM	MS_SSIM	W_SSIM	WMS_SSIM	PSNR	SPSNR	WS_PSNR	VPSNR
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

B. Temporal Correction Factor Based on Uncompressed Video Features

The subjective tests have shown that the temporal sub-sampling may have a high impact on the perceived video quality. Accordingly, and inspired by the work presented in [3], in this section a temporal correction factor (TCF) is proposed, that reduces the quality predicted by the previous metrics when the frame rate decreases. As in [3] the temporal factor is based on content-dependent parameters, extracted from the original video throughout features like: SI, TI, NTI, NSI, $\log_{10}(NSI)$, $\log_{10}(NTI)$, $TI \times SI$, $NTI \times NSI$, $\log_{10}(NSI \times NTI)$, $\log_{10}(TI \times SI)$, Frame Difference (*FD*), Frame Standard Deviation (*STD*), Normalized Frame Difference (*NFD*), Motion Vector Magnitude (*MVM*), Displaced Frame Difference (*DFD*), Motion Activity Intensity (*MAI*), Motion Direction Activity (MDA), *MVM* normalized by *STD* (*NMV_STD*), *MVM* normalized by *MAI* (*NMV_MAI*), and *MVM* normalized by *MDA* (*NMV_MDA*), proposed in [3], [10]. This time was used MOS instead of DMOS, since in the DMOS, the temporal resolution used as reference must be the same and the video sequences used in the temporal resolution subjective test session had different maximum temporal resolutions, namely 30 and 60 fps.

The MOS versus frame rate curve resulting from the subjective tests (see Fig. 4 c)), shows a behavior that can be well modeled by an inverted exponential function, described by (5).

$$TCF = \frac{1 - e^{-b\left(\frac{f}{f_{max}}\right)}}{1 - e^{-b}}, \qquad f_{max} = 60 \text{ fps} \quad (8)$$

This function was firstly proposed in [3] as a model of the frame rate impact on 2D video quality using a model parameter content-dependent, b. *TCF* goes from 0 to 1, where 1 means no loss of quality and 0 total loss of quality. Fig. 5 presents the model fitting (using the nonlinear least squares method and the trust-region algorithm) of the NMOS (MOS normalized by 5), by the model (8). For all videos, the NMOS predicted by (8) is quite close to the actual NMOS. Can be concluded that the model parameter b, is sequence dependent. To predict it from the video characteristics, a bidirectional stepwise regression was applied assuming a linear model between b and the content-dependent video features described before resulting in (6), with $x_1 = 5.120$, $x_2 = -0.372$, $x_3 = 3.103$, $x_4 = 0.640$ and $x_5 = 2.616$.

$$b' = x_1 + x_2 NFD + x_3 \log_{10}(NTI) \quad (9)$$

The parameterization (9) presented a marginally significant p-value of 7%, slightly higher than 5%, which means that there is a high probability that the null hypothesis (H_o: "*The regression b' does not explain b.*") can be ruled out. The parameterization (9) were obtained with all test videos. To validate the model, the leave-one-out cross-validation (LOOCV) technique was then applied, with five rounds. In each round, the video sequences are split in two groups: the training group, with five sequences and used to obtain the model coefficients; and the test group, with one sequence and used to test the model. This technique was used to avoid training and testing with the same video sequences because the number of videos is very low. From now on b' refers to the values obtained with this technique. The model *TCF*, when compared with the subjective scores (see Fig. 6), presented a PLCC of 0.97, SRCC of 0.97 and RMSE of 0.05, showing that the parameterization b' can be inserted into TCF generating a reliable objective metric for the frame rate impact on quality.

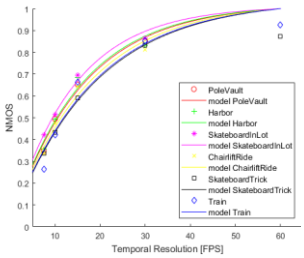

Fig. 5. Fitting of the NMOS from the temporal resolution assessment subjective test with $TCF(f,b)$.

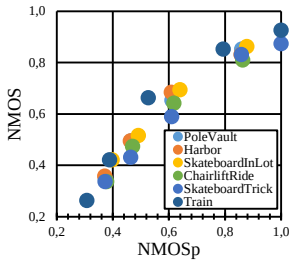

Fig. 6. NMOS vs NMOSp with $NMOSp = TCF(f, b = b')$.

C. Quality Prediction Model Fully Based on Uncompressed

Like previously done for the temporal resolution, models to assess the impact on quality of the spatial resolution and Q_p fully based on uncompressed video were also tested following [4] but in the omnidirectional video context.

1) Spatial and Quantization Correction Factor

For the spatial resolution case, the subjective scores of the spatial resolution subjective test stage were subject to model fitting with the proposed model in [4] called Spatial Corrections Factor (SCF) and described by (10). For the Q_p case, the subjective scores of the Q_p subjective test stage were subject to model fitting with the proposed model in [4] and described by (11) called Quantization Correction Factor (SQF). Note that (11) is in function of the quantization stepsize, q, which relates to Q_p with (12). The minimum Q_p used was 15, which corresponds to a q of 3.564.

$$SCF = \frac{1-e^{-c\left(\frac{s}{s_{max}}\right)^{0.6}}}{1-e^{-c}}, \qquad s_{max} = 7680 \times 3840 \tag{10}$$

$$QCF = \frac{e^{-a\left(\frac{q}{q_{min}}\right)}}{e^{-a}}, \qquad q_{min} = 3.564 \tag{11}$$

$$q(Q_p) = 2^{\frac{Q_p-4}{6}} \tag{12}$$

Using the same logic as for the prediction b', the parameters c and a were verified to be possible to predict for each video using the combination of features expressed in (13) for the spatial resolution case with $x_1 = 1.572$, $x_2 = 0.263$, $x_3 = 0.445$, $x_4 = -13.978$ and $x_5 = -30.547$ (see Fig. 7) and expressed in (14) for the Q_p case with $x_1 = 0.116$, $x_2 = -0.089$,$x_3 = 0.076$ and $x_4 = 0.018$ (see Fig. 9). In (13) was assumed a quadratic model between c and the video features.

$$c' = x_1 + x_2 MVM + x_3 NVV\_MDA + x_4 \log_{10}(NSI) + x_5 (\log_{10}(NSI))^2 \tag{13}$$

$$a' = x_1 + x_2 NTI + x_3 \log_{10}(NTI) + x_4 \log_{10}(NSI) \tag{14}$$

Both parameterization (13) and (14) presented a highly significant p-value of 1%, thus lower than 5%, which means that the null hypothesis (H_o: "*The regression does not explain the model parameter.*") can be rejected. Then, to avoid training and testing with the same sequences, the LOOCV technique was applied to obtain new values of c' and a' using the same features, that from now on refer to the values obtained with this technique. The model *SCF* when compared with the subjective scores (see Fig. 8) presented a PLCC of 0.94, SRCC of 0.95 and RMSE of 0.11, showing that the parameterization c' can be inserted into SCF generating a reliable objective metric for the spatial resolution impact on quality. The model QCF when compared with the subjective scores (see Fig. 10) presented a PLCC of 0.94, SRCC of 0.93 and RMSE of 0.09, showing that the parameterization a' can be inserted into QCF generating a reliable objective metric for the Q_p impact on quality. Other models were tested but the ones presented here were the ones that presented the highest correlations and lower RMSE when compared to the subjective scores.

2) Full Model Performance

As case study, the models *TCF*, *SQF* and *QCF* where combined to verify the independency (15), low-bounded to 1 and up-bounded to 5, of the three distortion causes.

$$MOSp = 5 \times SCF \times TCF \times QCF \tag{15}$$

In each individual prediction model, were used the parameters obtained with LOOCV. The predicted MOS (15), was applied to the sequences obtained for the quantization subjective test session plus the ones obtained for the combined effect subjective test session and then compared to the respective MOS (see Fig. 11), resulting in a PLCC of 0.92, a SRCC of 0.94 and a RMSE of 0.72. The points follow a linear trend, however mostly above the line MOS=MOSp, what was expected due to the relatively high RMSE of 0.72. This fact might that indicate that the three distortions causes are not completely independent from each other's. To reduce the RMSE, was found that a factor of 1.2 could improve the RMSE outcome. Using this factor, the new MOSp comes as (16), again bounded between 1 and 5.

$$MOSp' = 6 \times SCF \times TCF \times QCF \tag{16}$$

With the multiplicative factor and due to the boundaries, the PLCC changed to 0.94. The RMSE however was reduced to 0.48. Fig. 12 shows the relation between the MOS and MOSp', where the points show a better track of the MOS=MOSp line, thus explaining the sharp decrease in RMSE.

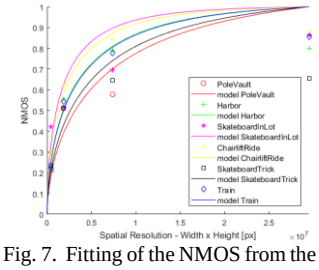

Fig. 7. Fitting of the NMOS from the spatial resolution assessment subjective test with $SCF(s,c)$.

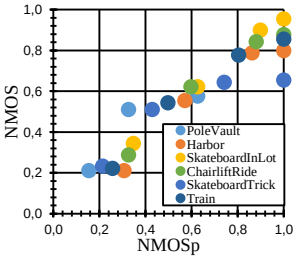

Fig. 8. NMOS vs NMOSp with $NMOSp = SCF(s, c = c')$.

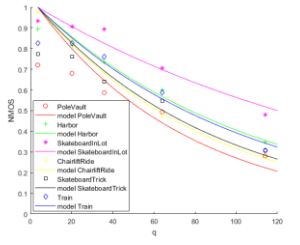

Fig. 9. Fitting of the NMOS from the Q_p assessment subjective test with $QCF(q,a)$.

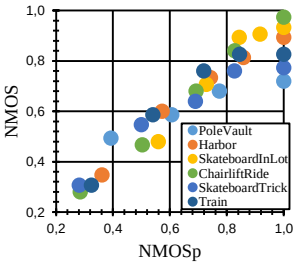

Fig. 10. NMOS vs NMOSp with $NMOSp = QCF(q, a = a')$.

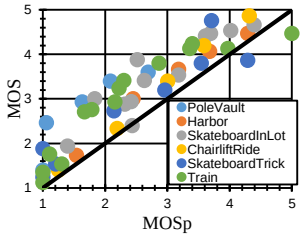

Fig. 11. Comparison between MOS and MOSp.

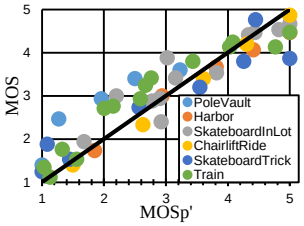

Fig. 12. Comparison between MOS and MOSp'.

VI. Bit Rate Prediction of Omnidirectional Video

To achieve a high Quality of Service (QoS) in a streaming environment it is necessary a methodology to perform rate-adaptation considered the available network bandwidth. This adaptation can be obtained by encoding the same video with different characteristics like the spatial/temporal resolution and the Q_p, which creates a tradeoff between the network bandwidth and media quality. In a bandwidth constrained streaming environment, it may not be possible to transmit video with the highest spatial/temporal resolutions and quality, and one possible and very common strategy is to allow the decoder to select the best video representation (i.e. the video characteristics and the quality associated to the quantization process, considering the network conditions. This sections starts by testing the same bit rate prediction model as in [4], but this time considering high resolution video and then present an algorithm designed for select best representations.

A. Bit Rate Prediction Model Based on Uncompressed Video Features

The bit rate prediction model (17) proposed in [4] is here evaluated to test its applicability to this type of content. This model follows a global bit rate prediction that considers the independent impact on the bit rate of the three types of distortion effects.

$$R_p(q,f,s) = R_{max}\left(\frac{q}{q_{min}}\right)^{-a_r}\left(\frac{f}{f_{max}}\right)^{b_r}\left(\frac{s}{s_{max}}\right)^{c_r} \quad (17)$$

In (17), R_{max} is the maximum bit rate for a video with no distortions, q is the quantization stepsize, q_{min} is the minimum quantization stepsize used to encode sequences, namely 3.564, f is the frame rate, f_{max} is the video maximum frame rate, s is the frame resolution ($width \times height$) in pixels, s_{max} is the video maximum frame resolution and a_r, b_r and c_r are model parameters. Each model parameter can be obtained with (17) by isolating each distortion and R_{max} can be obtained by encoding the videos with maximum spatial/temporal resolution and minimum quantization stepsize.

$$R_p(q) = R_{max}\left(\frac{q}{q_{min}}\right)^{-a_r} \quad (18)$$

$$R_p(f) = R_{max}\left(\frac{f}{f_{max}}\right)^{b_r} \quad (19)$$

$$R_p(s) = R_{max}\left(\frac{s}{s_{max}}\right)^{c_r} \quad (20)$$

To predict the bit rate were used all the ten video sequences available, and presented on Table I. For the tests were only generated sequences with one second out of the total duration of each video, due to the time it would take to encode all needed sequences at full duration. After each transformation (changes in spatial/temporal resolutions and compression) the bit rate was computed with (21), where S is the size of the generated video sequence, N_f is the number of generated frames and f is the temporal resolution.

$$R(s,f,q) = f\frac{S}{N_f} \quad (21)$$

1) Estimation of the Maximum Bit Rate

The maximum bit rate, R_{max}, was found, by encoding the original video sequences with minimum quantization stepsize, and original temporal and spatial resolutions, depending on the video and then computing (21).

2) Estimation of Bit Rate Model Parameters

For each independent distortion bit rate model prediction, the content-dependent parameters a_r, b_r and c_r are estimated with a Full Search (FS) algorithm that searches for the values that minimize the relative error between a proposed model and real bit rates. After that, estimations of a_r, b_r and c_r are found with the same generalized linear model and LOOCV, using video features, as previously done in section V. To estimate the bit rate depending on the Q_p, each video was encoded with Q_p of 15, 30, 35, 40 and 45, each with the original temporal and spatial resolution and then the real bit rates $R(q)$ were computed. To estimate the bit rate depending on the frame rate, each video was temporally down sampled to the set next set of frame rates: 7.5, 10, 15, 30 fps, each with the original spatial resolution and minimum q and then the real bit rates $R(f)$ were computed. To estimate the bit rate depending on the spatial resolution, each video was spatially down sampled to the next set of spatial resolutions: 960×480, 1920×960, and 3840×1920, each with the original temporal resolution and minimum q and then the real bit rates $R(s)$ were computed. Then the algorithm FS was applied considering the models (18), (19) and (20) respectively, to find the content-dependent parameters a_r, b_r and c_r that better fit the bit rates obtained with $R(q)$, $R(f)$ and $R(s)$ respectively. As for the quality models, a bidirectional stepwise regression, was applied assuming a quadratic model between a_r and the video features. For b_r and c_r was assumed a linear model. The resulting model for the prediction of parameter a_r was (22) with $x_1 = -17.409$, $x_2 = 0.053$, $x_3 = 17.648$, $x_4 = -0.543$, $x_5 = -28.290$ and $x_6 = -13.667$, for b_r was (23) with $x_1 = -1.912$, $x_2 = -11.524$, $x_3 = 0.066$, $x_4 = 3.023$, $x_5 = -1.521$ and $x_6 = -3.218$, and for c_r was (24) with $x_1 = 1.427$, $x_2 = 0.112$, $x_3 = -4.100$, $x_4 = 1.879$ and $x_5 = -0.004$. The three models a_r', b_r' and c_r' presented significant p-values of 4%, 3% and 2% respectively, thus lower than 5%, indicating that there is sufficient evidence to reject the null hypothesis (H_o: "*The regression found does not explain the model parameter.*").

$$a_r' = x_1 + x_2 STD + x_3 NTI + x_4(NSI \times NTI) + x_5 \log_{10}(NTI) + x_6(\log_{10}(NTI))^2 \quad (22)$$

$$b_r' = x_1 + x_2 NFD + x_3 DFD + x_4 NSI + x_5(NSI \times NTI) + x_6 \log_{10}(NSI) \quad (23)$$

$$c_r' = x_1 + x_2 DFD + x_3(NSI \times NTI) + x_4 \log_{10}(NSI) + x_5(NVM\_MAI \times DFD) \quad (24)$$

To validate the model, the LOOCV technique was applied, with nine rounds. From now on, the predictions a_r', b_r' and c_r' stand for the parameterizations obtained with LOOCV. Between the real bit rates obtained with $R(q)$ and

$R_p(q, a = a'_r)$ was obtain a mean relative error of 37.81%, between the bit rates obtained with *R(f)* and $R_p(f, b = b'_r)$ a mean relative error of 6.78% and between the bit rates obtained with *R(s)* and $R_p(s, c = c'_r)$ a mean relative error of 16.52%.

3) Performance Evaluation of the Complete Model

To test the performance of the bit rate prediction model (17) with the three effects combined, it was applied to every sequence ever generated throughout this work with Q_p of 15, 30, 35, 40, and 45, out of a total of 194 sequences. It was computed the real bit rate and the predicted bit rate with the model (17) for each aforementioned sequence and then was computed the mean relative error between the two bit rates. The combined model presented a mean of the relative errors of 30.84% and of only 16.58% of the videos presented a relative error above 50%. In this case, the almost blind (only based on the original uncompressed video sequence) bit rate prediction with a mean relative error of 30.84% can be considered good, though higher than 10%.

B. Best Representations Selection

With predictive models of bit rate (17) and subjective quality (16), it is possible to define an algorithm to select best representations (s, f, Q_p) under certain conditions, for example, considering a limited bandwidth environment. Similarly to [11], an algorithm for selecting best representations was developed, which is described next.

1) Representation Selection Algorithm

The algorithm designed to select best representations uses the predictions models presented before to get the best representation under a certain criterion. The algorithm receives the following inputs: a set of possible spatial and temporal resolutions and qualities (via a set of Q_p), a maximum bit rate (obtained with $s_{max}, f_{max}, q_{min}$), original temporal and spatial resolutions values and minimum Q_p as well as the model parameters estimated for each prediction model and video, such as b', c', a', $a_r{}'$, $b_r{}'$ and $c_r{}'$, a number of intervals N to each a representation must return and a selection criterion: selection by perceptual quality or selection by bit rate. In the first case it receives as input a minimum and maximum MOS plus a $\text{MOS}_{\text{fluctuation}}$ and in the second case it receives as input a minimum and maximum bit rates. Finally, it outputs a list of representations.

The algorithm starts by finding a list of all possible representations with corresponding MOS and bit rate, one representation for each combination of possible spatial and temporal resolution and Q_p, within the specified sets. Then, if the criterion is maximizing MOS for a specific bit rate range, it finds a set of N bit rate values between the maximum and minimum bit rate that are equally spaced. If the criterion is minimizing bit rate for a specific MOS it finds a set of N MOS values between the maximum and minimum MOS that are equally spaced. Finally, if the criterion is maximizing MOS for a specific bit rate range, for each bit rate in $n_i \in N$, the algorithm returns the representation that has the highest MOS and a bit rate up to the specified in the interval. On the other hand, if the criterion is minimizing bit rate for a specific MOS, for each MOS in $n_i \in N$, MOS_n, the algorithm returns the representation that has the minimum bit rate but a MOS inside the specified $\text{MOS} \pm \text{MOS}_{\text{fluctuation}}$ in the interval.

2) Selection by Perceptual Quality

In this case, the representation that for some MOS value presents the lowest bit rate is selected. The algorithm was tested using the sequence *SkateboardInLot*. Fig. 13 presents the result of the algorithm (blue circles) for a MOS range between 1 and 5, with $N = 7$ and a fluctuation of 10% of the specified MOS_n. The red circles represent the real bit rates obtained after encoding the selected representations. Note that the error between predicted and real bit rate increases when the Q_p decreases.

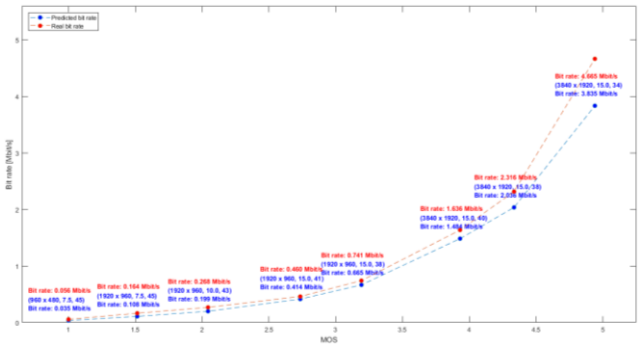

Fig. 13. Comparison between predicted and real bit rate when selecting representations with minimum bit rate for a specific MOS range.

3) Selection by Bit Rate

In this case, the representation that for a specific bit rate (from a set of bit rates) has the higher MOS is selected. The algorithm was evaluated using the sequence *Train*. Fig. 14 presents the result of the algorithm for a bit rate range between 1.7 Mbit/s and 13 Mbit/s, with a number of representations to be returned $N = 10$. Each point of the figure is labeled with the representation that has the highest MOS and a bit rate up to the bit rate specified in the interval. Fig. 14 presents, in blue, the predicted quality for each bit rate and in red the MOS obtained in the subjective tests for the same representation. The number of points are limited by the subjective scores available which is less than the amount of representations, which explain why only three points appear in the figure. The quality prediction tends to be rather close to the subjective score.

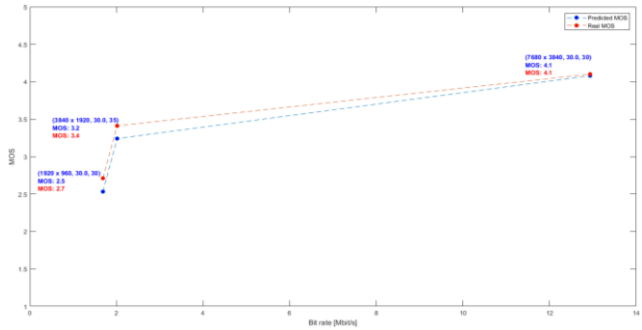

Fig. 14. Comparison between predicted and real MOS when selecting representations with maximum MOS up to a specific bit rate.

VII. Conclusion And Future Work

It was proven that the conventional objective metrics and the

ones designed for 360° video can assess with high accuracy the effects caused by spatial down sampling and Q_p variations independently. On this topic, it also proved the advantage of using multi scale SSIM metrics to access a mix of spatial resolution and Q_p variations. Also, it was verified that was possible to make quality prediction models based only on uncompressed video using features intrinsic to them. The developed quality prediction model developed considers as independent the three effects prorogued by changes in temporal/spatial resolution and quantization parameter. As observed, might not exist a complete independency between the three effects as considered here and in the literature. In the subject of bit rate based on an independent three factor model considering the impacts of spatial downscale, frame skipping and quantization increasing, was shown that it is possible to predict bit rate with a relative error 33% considering a not optimized procedures. After that, with quality and bit rate prediction models, it was possible to develop an algorithm for best representation selection using the prediction models.

Despite the promising results presented here, some limitations were found that might had a considerable impact on the conclusions. Next are presented some recommendations that could be done in order to improve the quality of the results presented here:

- **Increase subjective sessions duration** - Due to the subjective sessions duration limitations, only a few sequences were generated from the original videos to be presented, implying that the predictions models were based on few video sequences, reason why a similar procedure should be done with more testing sequences.
- **Avoid overfitting** - Naturally, making prediction models using 6 videos and 20 video features is prone to induce in overfitting, here plugged with LOOCV. Therefore, using more videos to create prediction models might increase the consistency of the model parameter prediction parameterizations.
- **Full sequence duration for bit rate prediction -** In the bit rate prediction, the bit rates used for modeling were obtained only from 1 second of each full video sequence. Probably, a solution based on the entire video sequence might be advantageous as well as divide videos in smaller segments, like groups of one or two GoPs and make a prediction for each group of GoPs separately.
- **Higher minimum Q_p -** The bit rate prediction model based on quantization variations has proved to have a considerably high mean relative error (around 37%), reason why it could be developed new fittings with a higher minimum Q_p, (even though reducing the maximum perceptual quality, subjective tests have proven that there is no significantly increase of QoE between Q_p of 15 and 30), that might lead to a more accurate bit rate prediction model based on the Q_p.
- **Distance between video representations -** It might be useful to include an additional consideration in the proposed representation selector algorithm which was a distance between representations, that could be a combination of the distance between each representation characteristic (spatial/temporal resolution and Q_p) each weighted by a factor according to the client´s preferences.

References

[1] Wikipedia, the free encyclopedia, “image stitching,” 18 January 2017. [Online]. Available: https://en.wikipedia.org/wiki/Image_stitching. [Accessed 3 March 2017].

[2] M. Yu and B. Girod, "A Framework to Evaluate Omnidirectional Video Coding Schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, Sekijo-machi, Hakata-ku Fukuoka, Japan, 2015.

[3] Y.-F. OU, Z. Ma, T. Liu and Y. Wang, "Perceptual Quality Assessment of Video Considering Both Frame Rate and Quantization Artifacts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 286-298, March 2011.

[4] H. Hu, Z. Ma and Y. Wang, "Optimization of spatial, temporal and amplitude resolution for rate-constrained video coding and scalable video adaptation," in *2012 19th IEEE International Conference on Image Processing*, Orlando, FL, USA, 2012.

[5] ITU-T, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," 2016.

[6] Oculus, "Oculus Rift," Oculus VR, LLC, 2017. [Online]. Available: https://www.oculus.com/rift/. [Accessed 2017 May 18].

[7] ITU-R, "BT-500-13 Methodology for the subjective assessment of the quality of television pictures," Geneva, 2012.

[8] FFPEG, "A complete, cross-platform solution to record, convert and stream audio and video.," [Online]. Available: https://www.ffmpeg.org/. [Accessed 24 April 2017].

[9] ITU-T Q.6/SG 16, ISO/IEC JTC 1/SC 29/WG 11, "High Efficiency Video Coding (HEVC)," Fraunhofer Heinrich Hertz Institute, [Online]. Available: https://hevc.hhi.fraunhofer.de/. [Accessed 27 April 2017].

[10] C. Lottermann, A. Machado, D. Schroeder, Y. Peng and E. Steinback, "Bit Rate Estimation For H.264/AVC Video Encoding Based On Temporal And Spatial Activities," in *IEEE International Conference on Image Processing*, Paris, 2014.

[11] L. Toni, G. Simon, R. Aparicio-Pardo, A. Blanc and P. Frossard, "Optimal Set of Video Representations in Adaptive Streaming," in *MMSys '14 Proceedings of the 5th ACM Multimedia Systems Conference*, Singapore, Singapore, 2015.

[12] Samsung, "Gear VR," [Online]. Available: http://www.samsung.com/global/galaxy/gear-vr/. [Accessed 7 March 2017].

[13] 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.

[14] Samsung, "Samsung Gear 360 2017," Samsung Electronics CO., LTD., [Online]. Available: http://www.samsung.com/global/galaxy/gear-360/. [Accessed 6 April 2017].

[15] X. Corbillon, G. Simon, A. Devlic and J. Chakareski, “Viewport-Adaptive Navigable 360-Degree Video Delivery,” in *2017 IEEE International Conference on Communications* , Paris, France, 2017.

[16] P. R. Alface, J.-F. Macq and N. Verzijp, "Interactive Omnidirectional Video Delivery: A Bandwidth-Effective Approach," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 135-147, March 2012.

[17] O. A. Niamut , E. Thomas, L. D'Acunto, C. Concolato , F. Denoual and S. Y. Lim, "MPEG DASH SRD - Spatial Relationship Description," 2016.

[18] B. Vishwanath, Y. He and Y. Ye, "AHG8: Area Weighted Spherical PSNR for 360 video quality evaluation," InterDigital Communications, Inc., San Diego, USA, 2016.