Quality and User Experience (2020) 5:4
https://doi.org/10.1007/s41233-020-00032-3

RESEARCH ARTICLE

Visual attention-aware quality estimation framework for omnidirectional video using spherical Voronoi diagram

Simone Croci[1] · Cagri Ozcinar[1] · Emin Zerman[1] · Sebastian Knorr[2] · Julián Cabrera[3] · Aljosa Smolic[1]

Received: 1 December 2019

Abstract

Omnidirectional video (ODV) enables viewers to look at every direction from a fixed point and provides a much more immersive experience than traditional 2D video. Assessing the video quality is important for delivering ODV to the end-user with the best possible quality. For this goal, two aspects of ODV should be considered. The first is the spherical nature of ODV and the related projection distortions when the ODV is stored in a planar format. The second is the interactive look-around consumption nature of ODV. Related to this aspect, visual attention, that identifies the regions that attract the viewer's attention, is important for ODV quality assessment. Considering these aspects, in this paper, we study in particular objective full-reference quality assessment for ODV. To this end, we propose a quality assessment framework based on the spherical Voronoi diagram and visual attention. In this framework, a given ODV is subdivided into multiple planar patches with low projection distortions using the spherical Voronoi diagram. Afterwards, each planar patch is analyzed separately by a quality metric for traditional 2D video, obtaining a quality score for each patch. Then, the patch scores are combined based on visual attention into a final quality score. To validate the proposed framework, we create a dataset of ODVs with scaling and compression distortions, and conduct subjective experiments in order to gather the subjective quality scores and the visual attention data for our ODV dataset. The evaluation of the proposed framework based on our dataset shows that both the use of the spherical Voronoi diagram and visual attention are crucial for achieving state-of-the-art performance.

Keywords Quality assessment · Omnidirectional video · 360° video · VR video · Spherical Voronoi diagram · Visual attention · Scaling distortion · Compression distortion

✉ Simone Croci
crocis@scss.tcd.ie

Cagri Ozcinar
ozcinarc@scss.tcd.ie

Emin Zerman
emin.zerman@scss.tcd.ie

Sebastian Knorr
knorr@tu-berlin.de

Julián Cabrera
julian.cabrera@gti.ssr.upm.es

Aljosa Smolic
smolica@scss.tcd.ie

[1] V-SENSE Project, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

[2] Communication Systems Group, Technical University of Berlin, Berlin, Germany

[3] Grupo de Tratamiento de Imágenes, Information Processing and Telecommunications Center and ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

Introduction

Omnidirectional video (ODV), also known as 360° or VR video, can be conceived as a spherical video where the viewers are placed at its center, allowing them to look at every direction. ODV is ideally viewed with a head-mounted display (HMD) that shows only the content in the direction where the viewer is looking at. In contrast to traditional 2D video, this emerging media type provides higher immersive and interactive viewing experience. Thanks to its immersive nature, ODV can be used in different applications such as entertainment [1, 2], communication [3], health-care [4], and education [5].

Compared to traditional 2D video, ODV introduces new technical challenges especially for storage and transmission [3]. For example, due to the large field of view of ODV [6], higher video resolution is necessary, and consequently, also higher memory requirements are demanded. For the development and evaluation of new solutions to these technical

Published online: 27 April 2020

challenges, like new compression and streaming approaches [3], subjective and in particular objective quality assessment methods are necessary to ensure a high quality of experience (QoE) [7]. There are already quality metrics for ODV like [8–13], but these metrics have limited correlation with the subjective quality scores. Thus, in order to improve the quality estimation performance, in this paper we propose a new quality estimation framework.

Quality assessment for ODV requires to consider its unique aspects. First, ODV is spherical in nature, but it is stored and transmitted in planar formats to be compatible with the existing video delivery pipelines. Different projection techniques can be used to map the spherical content to the 2D plane [14], such as equirectangular projection (ERP) and cubemap projection (CMP). These projections inevitably introduce distortions which must be taken into account to accurately estimate the video quality [15]. Second, HMDs allow the viewer to freely look around a scene [16], but they show only a part of the video, called viewport. In [11], it was found that less than 65% of ODV area is viewed by the viewers and consequently only such a partial region determines the perceived quality. Therefore, it is important to consider the viewing behavior while exploring ODV with an HMD [17–19], and to identify in particular the ODV regions that attract the visual attention [1, 16, 20–22] and are consequently viewed with high probability. Various previous research works emphasize the importance of visual attention in quality assessment [15, 23], and existing studies show that visual attention improves the performance of quality assessment [11–13, 24, 25].

In this paper, we propose an objective full-reference quality assessment framework that takes into account the spherical nature of ODV and its viewing characteristics. The framework first subdivides the ODV into planar patches using the spherical Voronoi diagram [26, 27]. These planar patches are characterized by low projection distortions, and we call them planar Voronoi patches. Afterwards, the framework applies a quality metric for traditional 2D video to each planar Voronoi patch, obtaining a quality score for each patch. To further consider the viewing characteristics of ODV, the proposed framework integrates visual attention by multiplying each patch score with a weight that accounts for the probability of the patch being viewed. Finally, the framework averages the weighted patch scores obtaining the final ODV quality score. The results show that both the ODV subdivision into planar Voronoi patches and the integration of visual attention improve the performance of ODV quality assessment.

This paper extends in several ways the contributions of our previous conference paper [28], where the original Voronoi-based framework without visual attention was initially proposed. These additional contributions are as follows:

- We integrate visual attention into the original Voronoi-based quality assessment framework;
- We extend the ODV quality dataset introduced in our previous study with 45 new distorted videos. In total, we create an ODV quality dataset of 120 distorted ODVs with scaling and compression distortions from eight undistorted reference ODVs, and we conduct a second subjective experiment in order to gather the subjective quality scores and the viewport trajectories for the new ODVs;
- We perform an extensive analysis of the parameters of the proposed quality assessment framework, such as the number and angular resolution of the planar Voronoi patches, the visual attention estimation method, and the temporal pooling of the frame scores;
- We perform a comparative analysis with existing quality metrics.

Our new dataset and the code of the proposed framework are made publicly available with this paper.[1] We expect that the developed framework and the established dataset will be beneficial for future research in ODV quality assessment, compression, and streaming.

The rest of this paper is organized as follows. Section "Related work" discusses the related work on both subjective and objective ODV quality assessment. Then, Section "Proposed quality assessment framework" describes the proposed quality assessment framework. The details of our extended ODV dataset and the related subjective experiments are explained in Section "Dataset and subjective experiments". Based on the proposed dataset, Section "Analysis and evaluation" presents the study of the framework parameter space and the extensive comparative analysis with several existing quality metrics. Finally, Section "Conclusion" concludes the paper.

Related work

Although there are many studies about subjective and objective ODV quality assessment, in the following, we outline only those that are most related to our work together with their limitations. For a comprehensive overview of recent research in the field, we recommend the overview paper of Li et al. [29].

[1] https://v-sense.scss.tcd.ie/research/voronoi-based-objective-metrics/.

Subjective quality assessment

Creating datasets and gathering subjective quality scores are fundamental requirements to understand the perceived quality of distorted omnidirectional images [23] and videos [11, 30–32]. For this purpose, Li et al.[11] conducted a subjective experiment to establish an ODV quality dataset. Their dataset contains subjective scores for 600 compressed ODVs across 221 participants. Eye and head movement data were also gathered during the subjective experiment. Another recent work [30] established a dataset that contains subjective quality scores of 30 participants across 50 different ODVs compressed with the HEVC/H.265 video coding standard [33]. In this work, the optimal resolution of ODVs displayed by the HMD was used in order to reduce the sampling distortions when extracting the viewport from the ODV. Furthermore, Singla et al.[31] and Schatz et al.[32] conducted subjective experiments to assess the perceived quality of ODV streaming.

At the time of writing this paper, most of the existing studies related to quality assessment, e.g., [15, 30, 34–36], consider only compression distortions of ODVs with low spatial resolution due to the computational complexity of ODV rendering. However, hardware for the rendering of 8K ODV is now on the market, providing higher quality of experience. Thus, in this paper, we extend our ODV dataset established in [28], which is based on the typical visual distortions in adaptive streaming systems, namely, compression and scaling distortions, applied to uncompressed ODVs with 8K resolution. We also organize a second subjective experiment to collect the subjective scores together with the viewport trajectories for the new ODVs.

Objective quality assessment

Many quality metrics developed for ODV are the extended versions of the traditional PSNR metric. Sun et al.[8], for instance, developed the weighted spherical PSNR metric (WS-PSNR) with weights that consider the projection distortions of the pixels in the planar format. The Craster parabolic projection PSNR metric (CPP-PSNR) [9] computes the PSNR in the Craster parabolic projection characterized by low projection distortions. Furthermore, the Spherical PSNR metric (S-PSNR) [10] estimates the PSNR for uniformly sampled points on the sphere. This quality metric has two different variants, namely, S-PSNR-NN and S-PSNR-I. When sampling pixels, they use the nearest neighbor or bicubic interpolation, respectively.

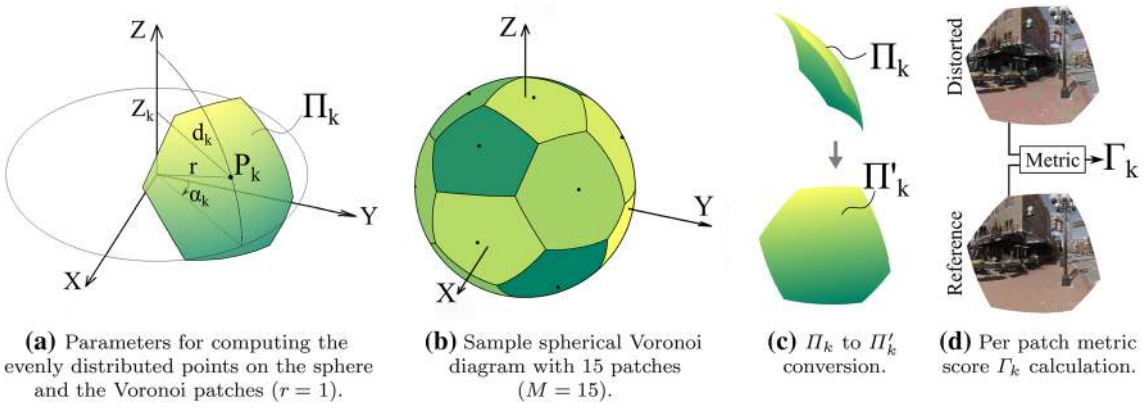
Subjective quality studies reported various findings about the PSNR-based quality metrics for ODV. On one hand, Zhang et al.[30] and Sun et al.[15] recently reported that the existing PSNR-based quality metrics for ODV have superior performance than the traditional PSNR. On the other hand, Tran et al.[35] claimed that the traditional PSNR is the most appropriate metric for quality evaluation in ODV communication. Furthermore, Upenik et al.[37] showed that the existing PSNR-based quality metrics for ODV do not have high correlation with subjective scores. A similar conclusion was reached in another study [34].

In addition to the PSNR-based metrics, the structure similarity index metric (SSIM) was also extended to ODV by Chen et al.[38] based on weights that take into account the projection distortions. Moreover, two recent studies [28, 36] investigated the performance of the video multimethod assessment fusion metric (VMAF) [39] applied to ODV, which is a metric for traditional 2D video developed to evaluate the distortions introduced by the adaptive streaming systems (i.e., compression and scaling distortions), and characterized by high correlation with subjective scores [40–42]. The work in [36] created a dataset of ODVs in ERP compressed using constant quantization parameters, and showed that VMAF can be used as a metric also for ODVs without modifications. Differently, in our previous work [28], we showed based on an ODV dataset with compression and scaling distortions, that the performance of VMAF can be improved using planar Voronoi patches.

In our previous work [28], we did not only study VMAF, but we developed a new objective quality assessment framework for ODV based on planar Voronoi patches. With our framework existing quality metrics for traditional 2D video (e.g., VMAF) can be applied to ODV based on planar Voronoi patches achieving high correlation with subjective scores. However, in our framework we did not consider visual attention.

Visual attention in objective quality assessment

As already shown in [15, 23], visual attention is crucial when evaluating the quality of ODV. Similarly, Li et al.[11] showed that the incorporation of head and eye movement data in objective quality assessment, more specifically in PSNR, increases the quality prediction performance. Upenik et al.[12] also proposed to incorporate visual attention in PSNR for ODV quality assessment. Furthermore, Ozcinar et al.[13] developed a quality metric based on PSNR that considers visual attention and projection distortions, with the aim of ODV streaming optimization. However, these works [11–13] that use visual attention are based on PSNR, which does not correlate well with subjective scores. Differently, in this paper, we develop a new quality assessment framework, which works with visual attention and robust quality metrics for traditional 2D video.



(a) Parameters for computing the evenly distributed points on the sphere and the Voronoi patches ($r = 1$).

(b) Sample spherical Voronoi diagram with 15 patches ($M = 15$).

(c) Π_k to Π'_k conversion.

(d) Per patch metric score Γ_k calculation.

Fig. 1 Figures of the proposed Voronoi-based quality assessment framework, showing patch extraction and patch metric score calculation

Proposed quality assessment framework

This section introduces the proposed framework for objective full-reference quality assessment based first on planar Voronoi patches obtained with the spherical Voronoi diagram [26], and second on visual attention. Initially, we introduce the planar Voronoi patches, and then we describe the original Voronoi-based framework presented in [28] and the new proposed Voronoi-based framework integrated with visual attention.

Planar Voronoi patches

For the extraction of M planar Voronoi patches from a given ODV, the spherical Voronoi diagram [26] of M evenly distributed points on the sphere [27] is computed as illustrated in Fig. 1a, b. The M evenly distributed points $\mathbf{P}_k = (X_k, Y_k, Z_k)$ on the sphere, where $k \in [0, M-1]$, are obtained according to the following equations:

$$\alpha_k = k\pi\left(3 - \sqrt{5}\right), \tag{1}$$

$$Z_k = \left(1 - \frac{1}{M}\right)\left(1 - \frac{2k}{M-1}\right), \tag{2}$$

$$d_k = \sqrt{1 - Z_k^2}, \tag{3}$$

$$X_k = d_k\, cos(\alpha_k) \quad \text{and} \tag{4}$$

$$Y_k = d_k\, sin(\alpha_k), \tag{5}$$

where α_k is the azimuthal angle and d_k is the distance of the point from the z-axis.

The spherical Voronoi diagram defines for each input point $\mathbf{P}_k$ the spherical patch Π_k on the surface of the sphere Ω_S that contains all the points that are closer to $\mathbf{P}_k$ than to any of the other input points $\mathbf{P}_l$:

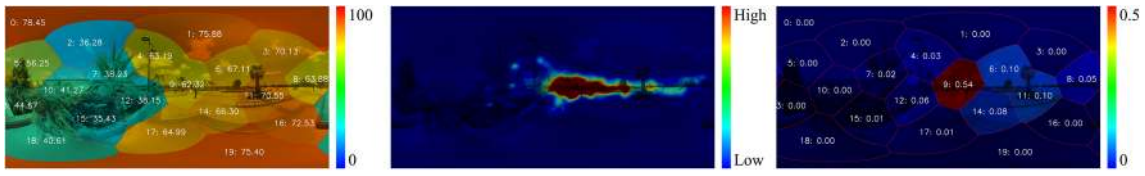
$$\Pi_k = \{\mathbf{P} \in \Omega_S \mid d_S(\mathbf{P}, \mathbf{P}_k) \leq d_S(\mathbf{P}, \mathbf{P}_l)\ \forall l \neq k\}, \tag{6}$$

where $d_S(\mathbf{P}, \mathbf{P}_k)$ is the spherical distance between the point $\mathbf{P}$ and the point $\mathbf{P}_k$, i.e., the length of the shortest path on the surface of the sphere connecting these two points. Notice that by using evenly distributed points $\mathbf{P}_k$ on the sphere, we guarantee that the spherical Voronoi patches Π_k have approximately equal size.

After the computation of the spherical Voronoi diagram, for each spherical Voronoi patch Π_k a planar Voronoi patch Π'_k is extracted from the ODV, as illustrated in Fig. 1c. This operation is obtained by first positioning the plane of the planar patch Π'_k on the centroid of the spherical patch Π_k, tangent to the sphere. The points on the sphere and the planar patch Π'_k are related by central projection, and the pixels of Π'_k are computed by sampling the ODV in ERP using bilinear interpolation. The angular resolution of each planar Voronoi patch Π'_k is defined by the pixels per visual angle, a parameter that is kept constant for each patch.

Original Voronoi-based quality framework

The quality framework presented in this section extends full-reference metrics for traditional 2D video to ODV. The extended metrics for ODV are called VI-METRIC, where VI stands for Voronoi, and METRIC $\in$ { PSNR , SSIM, MS-SSIM, VMAF, …} is a full-reference metric for traditional 2D video. Since we are dealing with full-reference quality assessment, the inputs of the framework are a distorted (e.g., compressed) ODV and the corresponding undistorted reference ODV. Initially, the quality framework



(a) VMAF scores of 20 Voronoi patches. **(b)** Visual attention map generated with the Kent distribution method [26]. **(c)** Visual attention patch weights $\nu_{i,k}$ of 20 Voronoi patches from the visual attention map in (b).

Fig. 2 Visualization of the VMAF patch scores, visual attention map, and the visual attention patch weights $\nu_{i,k}$. Please refer to the color bars beside the figures for the used color code

extracts M planar Voronoi patches Π'_k from the distorted ODV and other M from the reference ODV. Then, a full-reference metric for traditional 2D video is applied to the planar Voronoi patches Π'_k of the distorted and reference ODV, obtaining M patch scores Γ_k as illustrated in Fig. 1d. In our study, we apply the following full-reference metrics: PSNR, SSIM [43], MS-SSIM [44], and VMAF [39]. Since these metrics take rectangular video frames as input, we modified the first three of them, so that they can deal with any patch shape. For VMAF we took the bounding box of the patch as input, as it is not straightforward to modify VMAF for different patch shapes. In the end, the final ODV quality score is obtained by computing the arithmetic mean of the patch scores Γ_k as follows:

$$\text{VI-METRIC} = \frac{\sum_{k=0}^{M-1} \Gamma_k}{M}. \tag{7}$$

Proposed Voronoi-based framework integrated with visual attention

When viewing an ODV with an HMD, only a part of the ODV, the so-called viewport, is visible, and usually different viewers tend to look only at particular regions that attract their visual attention. Since different parts of an ODV can be of different quality, it is important during quality assessment to give more weight to the regions that attract the visual attention, i.e., the regions that are most likely to be viewed. A way to represent these regions is the visual attention map, which defines for each pixel of the ODV the probability of being viewed.

To take visual perception into consideration for ODV quality assessment, we now propose to integrate visual attention into the original Voronoi-based framework, and refer to its metrics as VI-VA-METRIC, where VA stands for visual attention. Different methods can be used for the computation of visual attention maps. We investigate the effects of different visual attention estimation methods in Section "Investigation of applying different visual attention estimation methods". Figure 2 shows a sample visual attention map generated using Kent method [22].

For the computation of the VI-VA-METRICs, first a quality score for each video frame of the distorted ODV is computed based on visual attention, and then the frame scores are pooled into a final quality score. For the computation of the frame scores, initially M planar Voronoi patches Π'_k are extracted from each frame i of the distorted and reference ODV. Then, a full-reference metric for traditional 2D video is applied to the planar Voronoi patches Π'_k of each frame i, obtaining M patch scores $\Gamma_{i,k}$ for each frame. At this point, the visual attention map Ψ_i of each frame i of the distorted ODV is estimated. Then, M planar Voronoi patches Π'_k are extracted from each visual attention map Ψ_i, and the sums $\nu_{i,k}$ of the visual attention pixel values inside each patch Π'_k of each map Ψ_i are computed. The sum $\nu_{i,k}$ is related to the probability of patch Π'_k of frame i being viewed. Next, the frame scores T_i are obtained through a weighted average of the patch scores $\Gamma_{i,k}$ using the visual attention sums $\nu_{i,k}$ as weights according to the following equation:

$$T_i = \frac{\sum_{k=0}^{M-1} \nu_{i,k} \Gamma_{i,k}}{\sum_{k=0}^{M-1} \nu_{i,k}}. \tag{8}$$

In the last step, the frame scores T_i are combined using a pooling approach P obtaining the final video score:

$$\text{VI-VA-METRIC} = P(T_0, T_1, \ldots, T_{N-1}), \tag{9}$$

where N is the number of frames. Different pooling approaches P can be applied, like the arithmetic and harmonic mean, the median, the minimum, etc. In this study, we analyze the following metrics obtained with the framework: VI-VA-PSNR, VI-VA-SSIM, VI-VA-MS-SSIM, and VI-VA-VMAF.

Figure 2 shows the patch scores obtained by applying VMAF to 20 Voronoi patches, the visual attention map computed by the Kent distribution method [22] from the viewport trajectories obtained in our subjective experiments, and the visual attention patch weights $\nu_{i,k}$ corresponding to 20 Voronoi patches. As can be seen in the figure, different

Fig. 3 Sample frames of the nine reference ODVs used in the subjective experiments. The top five ODVs were rated in the first subjective experiment [28], and the bottom left three ODVs were rated in the second subjective experiment. *Train* was used for the experiment training

regions of the ODV can have noticeably different qualities, and also clearly different visual attention values. For this reason, we integrate visual attention in our proposed quality assessment framework in a way to give more importance to patches that attract visual attention.

Dataset and subjective experiments

In this section, we introduce our dataset, and we describe the technical details of the two subjective experiments that we conducted in order to collect the subjective quality scores and the viewport trajectories for our dataset. This section terminates with the analysis of the collected subjective data.

Omnidirectional video quality dataset

Considering a streaming application scenario, we built our dataset using ODVs with different spatial resolutions and different compression levels. For our dataset and subjective experiments, we first selected a total of nine *uncompressed* reference ODVs in YUV420p format of 10 *sec.* length, $8K \times 4K$ ERP resolution, and with different characteristics. These ODVs were selected from the videos of the joint video exploration team of ITU-T VCEG and ISO/IEC MPEG [45–47]. The selected videos are *Basketball*, *Dancing*, *Gaslamp*, *Harbor*, *JamSession*, *KiteFlite*, *SkateboardTrick*, *Train*, and *Trolley*. Sample frames of these videos are shown in Fig. 3. *Basketball*, *Dancing*, *Harbor*, *JamSession*, *KiteFlite* were rated in the first subjective experiment already reported in [28], and *Gaslamp*, *SkateboardTrick*, *Trolley* were rated in the second experiment reported here. The *Train* sequence was used only as training material in both experiments.

After the selection of the nine reference ODVs, they were downsampled to three different resolutions in ERP format: 8128×4064, 3600×1800, and 2032×1016. For the downsampling, we used the bicubic scaling algorithm of the FFmpeg software (*ver.* 4.0.3-1 18.04). Next, the ODVs were compressed with the HEVC/H.265 video coding standard [33]. For this, we used the *libx265* codec (*ver.* 2.9) [48] in FFmpeg [49] with the video buffering verifier method to set the target bitrates. As this database is created to understand possible cases which might be encountered in an adaptive streaming scenario, to ensure constant bitrate, each ODV was compressed using two-pass encoding with 150 percent constrained variable bitrate configuration, following the recommendations of streaming providers [50]. We also defined the buffer size during encoding to limit the output bitrate to twice the maximum bitrate for handling large bitrate spikes. To avoid any possible impact of the unknown resampling algorithm used by the video player, we upsampled the decoded ODVs to 8128×4064 resolution using the bicubic scaling algorithm of FFmpeg. For the downsampling and compression of the reference ODVs, we used the following FFmpeg commands:

```
ffmpeg -y -f rawvideo -pix_fmt
    iVideoFormat -s iVideoRes -r
    iVideoFramerate -i iVideoFn -c:v libx265
     -preset medium -frames:v
    iVideoFrames -vf scale=oVideoRes -x265-
    params profile=main:keyint=48:min-
    keyint=48:scenecut=0:ref=5:bframes
    =3:b-adapt=2:bitrate=oVideoBitRate:vbv
    -maxrate=oVideoMaxRate:vbv-bufsize=
    oVideoBufSize:pass=1 -f mp4 /dev/null

ffmpeg -y -f rawvideo -pix_fmt
    iVideoFormat -s iVideoRes -r
    iVideoFramerate -i iVideoFn -c:v libx265
     -preset medium -frames:v
    iVideoFrames -vf scale=oVideoRes -x265-
    params profile=main:keyint=48:min-
    keyint=48:scenecut=0:ref=5:bframes
    =3:b-adapt=2:bitrate=oVideoBitRate:vbv
    -maxrate=oVideoMaxRate:vbv-bufsize=
    oVideoBufSize:pass=2 oVideoFn
```

Table 1 Bitrates (in Kbps) for the selected ODVs

ODV	BR1	BR2	BR3	BR4	BR5
Basketball Dancing	500	1000	2000	5000	13000
Harbor JamSession Gaslamp SkateboardTrick Trolley	500	1000	2000	7000	13000
KiteFlite	500	1000	5000	7000	13000

Table 2 Statistics of the stimuli and the participants in the subjective quality assessment experiments

Subjective Experiment	# of Stimuli	# of Participants	Min – Mean – Max Age	Ratio of women (%)
First [28]	75 + 5 Ref	24	22 – 29.7 – 38	16
Second	45 + 3 Ref	23	25 – 31.6— 42	26

where

- *iVideoFn*: filename of input video,
- *iVideoRes*: resolution of input video,
- *iVideoFormat*: format of input video (in our case yuv420p),
- *iVideoFramerate*: framerate of input video,
- *iVideoFrames*: number of frames of input video,
- *oVideoFn*: filename of output video,
- *oVideoRes*: resolution of output video,
- *oVideoBitRate*: target bitrate of output video in Kbps,
- *oVideoMaxRate*: maximum bitrate in Kbps (in our case $1.5 \times oVideoBitRate$),
- *oVideoBufSize*: buffer size in Kbps (in our case $2 \times oVideoMaxRate$).

To ensure that the distorted ODVs within the database are uniformly distributed across different quality levels, five different target bitrates were selected independently for each reference ODV in a pilot test with three experts using HTC Vive HMD. For this pilot test, before encoding, the reference ODVs were resized to the resolution 3600×1800, which was found to be the optimal ODV resolution for HTC Vive HMD by Zhang et al.[30], after their calculation considering the HMD's display resolution and its field of view. The ODVs were then encoded with different bitrates $\in \{500, 1000, 2000, 5000, 7000, 10000, 13000, 15000\}$Kbps, and among them five different bitrates were selected in the pilot test to correspond to five different quality levels, namely, *"bad"*, *"poor"*, *"fair"*, *"good"*, and *"excellent"*, which are reported in Table 1.

Subjective experiments

This section describes the technical details of the two subjective experiments that we organized. Their main characteristics are shown in Table 2.

Experiment setup

The subjective experiments were conducted in a dedicated experiment room equipped with an HTC Vive HMD, which was used to present the stimuli to the viewers. Participants were seated in a swivel chair and allowed to turn freely. To ensure that the participants could vote without removing the HMD, we used the Virtual Desktop application. Virtual Desktop is an ODV player and an application that enables the users to watch and interact with the desktop using the HMD and VR controllers. Using this application and the open-source MATLAB GUI presented in [51, 52], participants were able to vote each stimulus. Additionally, with a special application, the viewport trajectories were also recorded during the presentation of each stimulus for the computation of the visual attention maps.

Methodology

The modified-absolute category rating (M-ACR) [53] methodology was chosen for our subjective experiments in order to lengthen the exposure to the stimuli, since in this methodology each stimulus is presented twice with a short mid-gray screen (in our case a three second long one) between the two presentations. The reference sequences were also included in the subjective experiments as hidden references. That is, the participants were not told of reference sequences, and they voted the hidden references as any other stimulus.

The subjective quality scores for all the videos were collected in two experiments with different ODVs and participants. The first experiment, which was presented in [28], comprised of two sessions of 30 minutes, one hour in total. The second experiment had only one session of 30 minutes. At the beginning of both experiments, there was a training phase when the *Train* video sequence with five different quality levels was displayed. After the training phase, the experiment ODVs were randomly displayed, and the quality scores were assigned by the participants based on a continuous grading scale in the range [0,100], with 100 corresponding to the best score, as recommended in ITU-R BT.500-13 [54].

Participants

24 participants, 20 males and four females, took part in the first experiment. These participants were aged between 22 and 38 years with an average of 29.7 years. 23 participants, 17 males and six females, took part in the second experiment. These participants were aged between 25 and 42 years with an average of 31.6 years. The gathered quality scores were screened for outliers using the outlier detection method recommended in ITU-R BT.500-13 [54]. Three outliers in the first experiment and two outliers in the second experiment were found and removed. All participants were screened for visual acuity and found to have normal or corrected-to-normal vision.

Subjective quality analysis

To represent the subjective quality of each stimulus, differential mean opinion scores (DMOS) [55] are calculated by applying the standard approach described in [56]. First, the difference scores are computed as: $d_{ij} = s^r_{ij} - s_{ij}$, where s_{ij} and s^r_{ij} are the raw subjective score assigned by participant i to the distorted ODV j and the raw subjective score assigned to the corresponding hidden reference ODV, respectively. These difference scores d_{ij} are converted to z-scores as follows: $z_{ij} = (d_{ij} - \mu_i)/\sigma_i$, where μ_i and σ_i are the mean and standard deviation of the raw scores assigned by the participant i. Then, the z-scores are linearly rescaled in the interval [0,100] as follows: $z'_{ij} = 100(z_{ij} + 3)/6$. The rescaling is based on the assumption that the z-scores z_{ij} are normally distributed with mean equal to zero and standard deviation equal to one, which means, that 99% of the z-scores z_{ij} are in the interval [-3,3], and consequently 99% of the rescaled z-scores z'_{ij} are in the interval [0,100]. The final DMOS value of ODV j is then obtained by averaging the rescaled z-scores z'_{ij} of the K participants excluding the outliers as follows:

$$\text{DMOS}_j = \frac{1}{K}\sum_{i=1}^{K} z'_{ij}. \tag{10}$$

Small DMOS values indicate that the distorted stimulus is closer to the reference, and hence small DMOS is better. Figure 4 shows the DMOS values of the ODVs included in the experiments. As expected, we can notice that there is an inverse relationship between DMOS and bitrate. From the plots we can also see that the ODVs with highest spatial resolution have the worst quality (highest DMOS) for low bitrate and the best quality for high bitrate. This shows that the 8128×4064 ODVs are coarsely compressed at the low bitrates due to the high number of pixels present. As the bitrate increases, the perceived quality for these videos gets better. Conversely, the perceived quality of the 2032×1016 ODVs becomes the worst at high bitrates, due to the scaling distortions [3].

Visual attention analysis

Table 3 shows the comparison between the visual attention maps of the reference ODVs and the corresponding ODVs with resolution 8128×4064 and encoded at the five bitrates reported in Table 1. For the comparison, first uniformly distributed points on the sphere are sampled from the visual attention maps, and then the Pearson's linear correlation coefficient (PLCC) and the Kullback–Leibler divergence (KLD) are applied to the sampled points [18]. Large PLCC values and small KLD values correspond to high similarity. As can be noticed from Table 3, the visual attention maps of the reference and corresponding distorted ODVs can be different, especially for the smallest bitrate BR1. This can also be noticed in Fig. 5, where the visual attention maps of the *JamSession* reference ODV and the corresponding encoded ODVs at the smallest and largest bitrates with resolution 8128×4064 are shown. In Table 3, there is also the average of the PLCC and KLD

Table 3 Pearson's linear correlation coefficient (PLCC) and Kullback-Leibler divergence (KLD) computed between the visual attention maps of the reference ODVs and the corresponding ODVs with resolution 8128×4064 and encoded at the five bitrates reported in Table 1

ODV	BR1		BR2		BR3		BR4		BR5	
	PLCC	KLD	PLCC	KLD	PLCC	KLD	PLCC	KLD	PLCC	KLD
Basketball	0.8914	0.5939	0.9134	0.6394	0.8838	0.7101	0.9019	0.8640	0.9195	0.6801
Dancing	0.6410	1.3625	0.6911	1.0891	0.7226	1.2005	0.7841	0.7137	0.7205	1.0115
Harbor	0.7316	0.7843	0.7134	0.6718	0.8341	0.4486	0.8348	0.5310	0.8536	0.4550
JamSession	0.5781	1.4356	0.8312	0.7140	0.7313	0.8753	0.8640	0.5990	0.8457	0.4435
KiteFlite	0.7273	0.8362	0.8136	1.0684	0.8486	0.5353	0.8352	0.6136	0.8557	0.5614
Gaslamp	0.7769	0.8339	0.7981	0.6213	0.8457	0.4773	0.8739	0.5137	0.8421	0.6517
SkateboardTrick	0.8705	0.7316	0.8713	1.0611	0.9413	0.4834	0.8901	0.5517	0.8976	0.3951
Trolley	0.8586	0.8713	0.7891	0.9610	0.8232	0.8207	0.8945	0.5879	0.9162	0.5906
Average	0.7594	0.9312	0.8026	0.8533	0.8288	0.6939	0.8598	0.6218	0.8564	0.5986

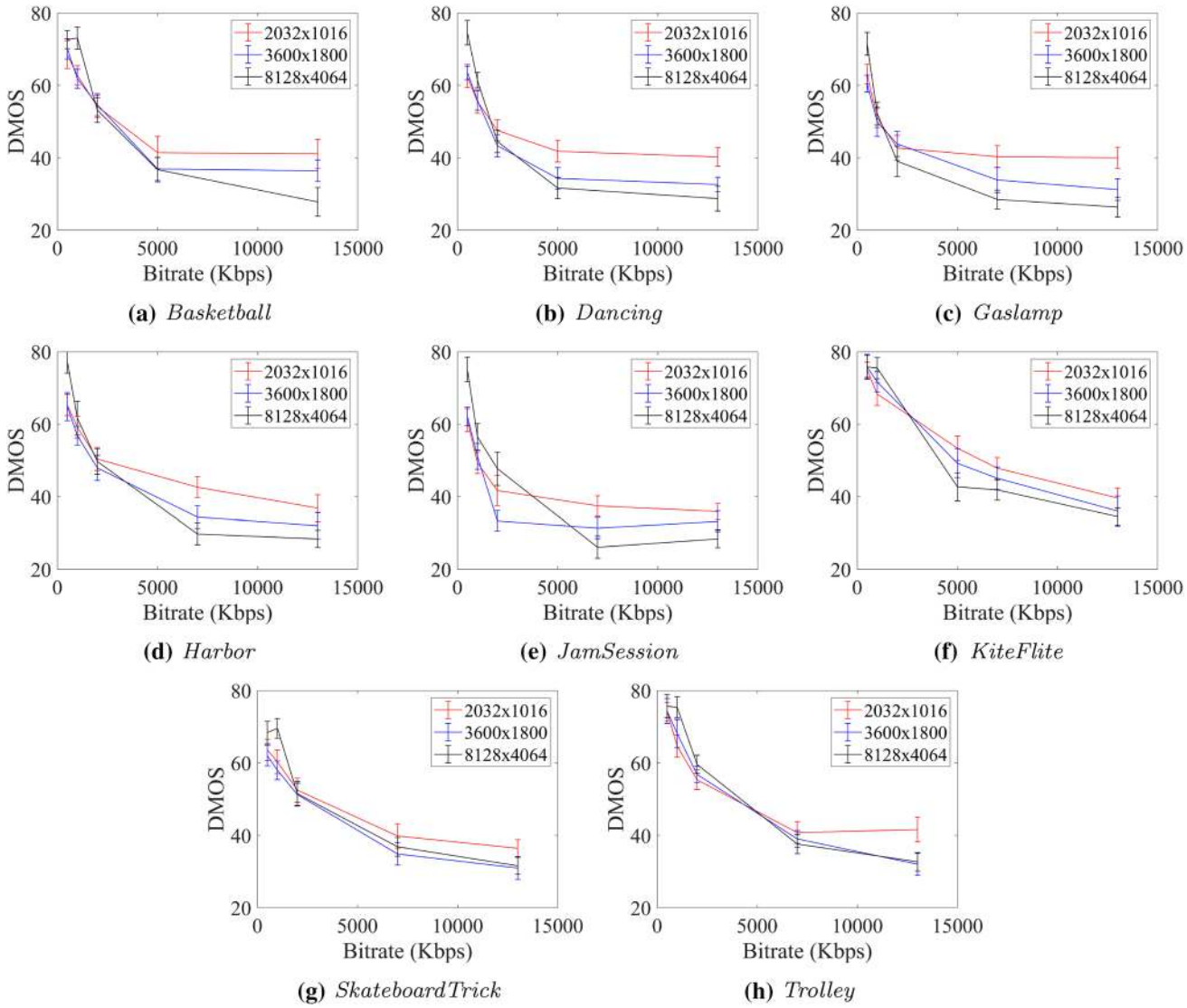


Fig. 4 Bitrate vs. DMOS plots of each ODV used in the subjective experiments. The vertical bars show 95% confidence intervals

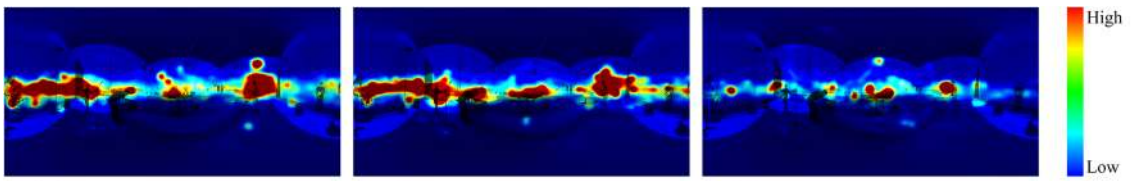


Fig. 5 Comparison of the visual attention maps of the *JamSession* reference ODV in (**a**) and two corresponding encoded ODVs in (**b**) and (**c**). See the color bar in (**d**) for the used color code

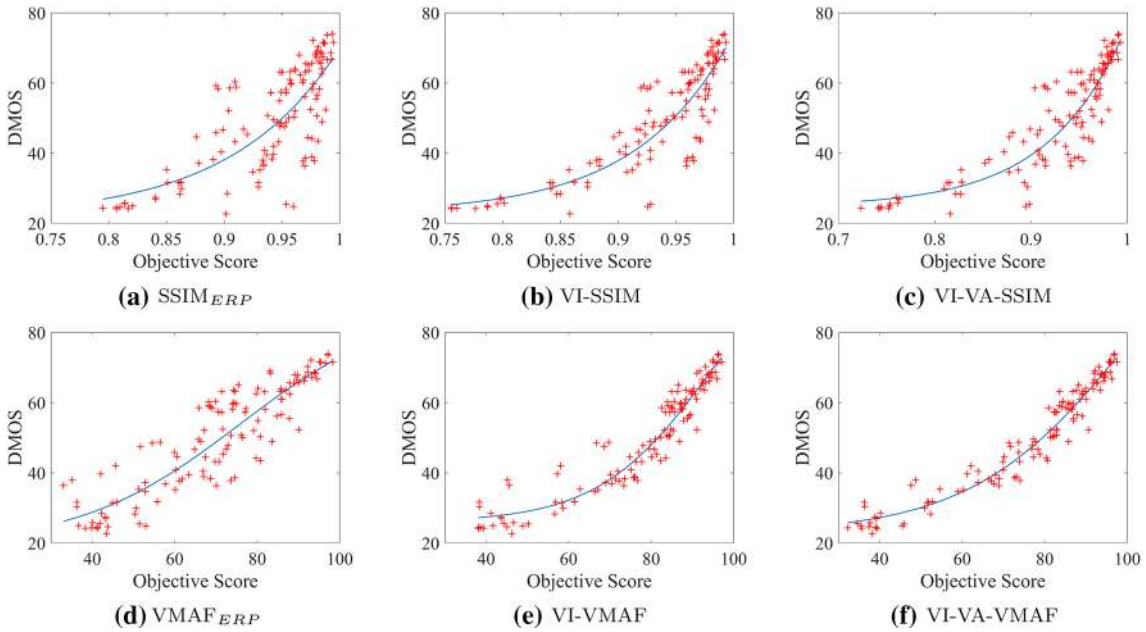


Fig. 6 Metric versus subjective score plots with the fitted logistic functions. Red points indicate the data points, and blue lines indicate the logistic functions

values for each bitrate. It can be seen that by increasing the bitrate the average PLCC increases while the average KLD decreases. Based on these observations and to ensure the most accurate results, in our framework we use, for each undistorted and distorted ODV, the corresponding visual attention map and not only the visual attention maps of the undistorted ODVs.

Analysis and evaluation

In this section, we first determine the optimal parameter values of the proposed framework, and then we compare the metrics of the proposed framework with existing quality metrics. With this aim, we use our ODV dataset with the gathered subjective quality scores presented in Section "Dataset and subjective experiments", and we analyze the correlation between the metric scores and the subjective scores. For the correlation analysis, we first convert the metric scores into the subjective scores by fitting a logistic function. We use the logistic function proposed in [57], and defined as follows:

$$s' = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{s-\beta_3}{\|\beta_4\|}}} + \beta_2, \tag{11}$$

where s' is the predicted subjective score of the metric score s, and $\beta_{1,\ldots,4}$ are the parameters that are estimated during the fitting. Here, the subjective score predicted by the logistic function is the reversed DMOS (i.e., subtracted from 100).

To evaluate how well the logistic function predicts the subjective scores, i.e., how well the metric estimates the subjective quality, the following measures are applied to the real and predicted subjective scores: Pearson's linear correlation coefficient (PLCC), Spearman's rank ordered correlation coefficient (SROCC), root mean squared prediction error (RMSE), and mean absolute prediction error (MAE). PLCC and SROCC measure the prediction accuracy and the monotonicity, respectively. The larger they are, the more accurate and monotonic the prediction is. For RMSE and MAE, the smaller they are, the better the prediction is.

To visualize the relationship between the metric and subjective scores, sample plots are shown in Fig. 6 for the metrics SSIM and VMAF applied to the ERP format (SSIM $_{ERP}$ and VMAF $_{ERP}$), and in the original and proposed Voronoi-based quality assessment framework. In these plots, the increase of the correlation between the metric scores and DMOS is noticeable for the VI-METRICs and the VI-VA-METRICs compared to the metrics calculated in ERP format.

Selection of optimal parameter values for the proposed framework

In this section, we fine-tune the proposed framework by determining the optimal parameter values.

Table 4 PLCC and SROCC of the Voronoi-based metrics with different angular resolutions and numbers of patches. The best performance values for each resolution (i.e., each row) are in bold, while the best performance values among all the metrics are in italics

Metrics	Resolutions	10 patches		15 patches		20 patches	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
VI-PSNR	10 pix/deg	0.8700	0.8584	**0.8775**	**0.8634**	0.8676	0.8551
	15 pix/deg	0.8700	0.8584	**0.8775**	**0.8636**	0.8675	0.8553
	20 pix/deg	0.8700	0.8584	**0.8775**	**0.8634**	0.8676	0.8553
VI-SSIM	10 pix/deg	0.8757	0.8667	0.8821	**0.8763**	**0.8823**	**0.8763**
	15 pix/deg	0.8423	0.8301	0.8509	0.8411	**0.8516**	**0.8414**
	20 pix/deg	0.8132	0.7995	0.8227	0.8072	**0.8237**	**0.8079**
VI-MS-SSIM	10 pix/deg	0.9468	0.9432	**0.9488**	0.9446	0.9486	**0.9450**
	15 pix/deg	0.9385	0.9361	**0.9411**	0.9381	0.9409	**0.9398**
	20 pix/deg	0.9314	0.9260	**0.9343**	**0.9303**	0.9339	0.9291
VI-VMAF	10 pix/deg	0.9634	0.9553	0.9615	0.9529	**0.9646**	**0.9581**
	15 pix/deg	0.9532	0.9444	0.9544	0.9470	**0.9581**	**0.9497**
	20 pix/deg	0.9387	0.9288	0.9435	0.9363	**0.9476**	**0.9401**
VI-VA-PSNR	10 pix/deg	**0.8977**	**0.8812**	0.8760	0.8563	0.8876	0.8712
	15 pix/deg	**0.8977**	**0.8817**	0.8760	0.8564	0.8876	0.8708
	20 pix/deg	**0.8977**	**0.8817**	0.8760	0.8564	0.8876	0.8707
VI-VA-SSIM	10 pix/deg	0.8947	0.8848	0.8921	0.8832	**0.9106**	**0.9007**
	15 pix/deg	0.8633	0.8510	0.8537	0.8426	**0.8777**	**0.8663**
	20 pix/deg	0.8353	0.8214	0.8188	0.8136	**0.8463**	**0.8323**
VI-VA-MS-SSIM	10 pix/deg	0.9563	0.9505	0.9628	0.9581	**0.9676**	**0.9635**
	15 pix/deg	0.9501	0.9438	0.9552	0.9506	**0.9627**	**0.9573**
	20 pix/deg	0.9445	0.9371	0.9482	0.9424	**0.9572**	**0.9517**
VI-VA-VMAF	10 pix/deg	0.9661	0.9589	0.9738	0.9667	*0.9773*	*0.9717*
	15 pix/deg	0.9580	0.9491	0.9678	0.9599	**0.9723**	**0.9658**
	20 pix/deg	0.9444	0.9349	0.9553	0.9482	**0.9623**	**0.9564**

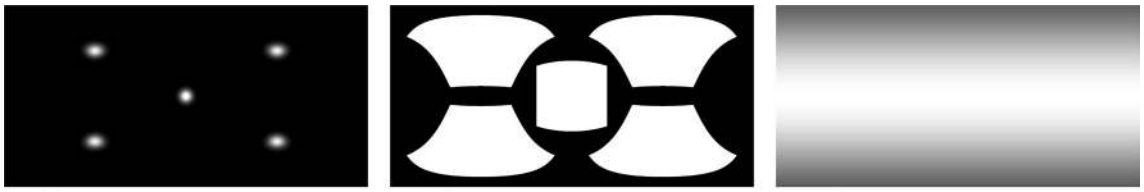
Estimation of the optimal angular resolution and number of planar Voronoi patches

We first analyze the two main parameters of the original and proposed frameworks that have an impact on the accuracy of the quality estimation, namely, the angular resolution and the number of the planar Voronoi patches.

For the Voronoi-based metrics obtained with the original and proposed framework, i.e., VI-METRICs and VI-VA-METRICs, Table 4 shows PLCC and SROCC based on different parameter values. Three angular resolutions are investigated, namely $\{10, 15, 20\}$ pix/deg, which are close to the resolution of the HTC Vive HMD used in our subjective experiments. Moreover, we also consider three different numbers of planar Voronoi patches, that is, $M = \{10, 15, 20\}$. For the estimation of the visual attention maps of the VI-VA-METRICs, we use the Kent method [22].

As can be seen in the table, the reduction of the patch resolution improves the performance of the Voronoi-based metrics in most of the cases. For the other cases, the performance remains almost constant. On the other hand, increasing the number of patches seems to positively influence the performance of the Voronoi-based metrics almost always, except for VI-PSNR and VI-VA-PSNR. This can be explained by the reduction of the projection distortions when the number of patches increases and consequently the patch size decreases. For the VI-VA-METRICs that use visual attention, the improvement of the performance can also be explained by the fact that with more patches the visual attention weights $v_{i,k}$ are localized to smaller regions and consequently more accurate.

As a result of this analysis, we select 10 pix/deg and 20 patches ($M = 20$) as the optimal parameter values for our proposed framework. We use these two parameters for the rest of this paper. Please note that although we select these optimal parameter values, independently of the studied parameter values, the Voronoi-based metrics are characterized by a better performance than the performance of the corresponding original metrics for traditional 2D video applied to the ERP and CMP formats, as shown later in Table 7.



(a) Kent distribution method. (b) Uniform viewport method. (c) Equator bias method.

Fig. 7 Visual attention maps computed with different methods using as input five different viewport positions

Table 5 Performance evaluation of the Voronoi-based metrics integrated with visual attention estimated with three methods. The best performance values are in bold

Metrics	Vis. Att.	PLCC	SROCC	RMSE	MAE
VI-PSNR	–	0.8676	0.8551	7.5743	5.8377
VI-VA-PSNR	Equator-bias	0.8781	0.8628	7.2995	5.5508
VI-VA-PSNR	Uniform	0.8774	0.8585	7.4141	5.7168
VI-VA-PSNR	Kent	0.8876	0.8712	7.1818	5.5072
VI-SSIM	–	0.8823	0.8763	7.1172	5.2867
VI-VA-SSIM	Equator-bias	0.8879	0.8850	6.9454	5.1687
VI-VA-SSIM	Uniform	0.8981	0.8929	6.8103	5.0647
VI-VA-SSIM	Kent	0.9106	0.9007	6.4345	4.8097
VI-MS-SSIM	–	0.9486	0.9450	4.8743	3.8475
VI-VA-MS-SSIM	Equator-bias	0.9486	0.9450	4.8790	3.8343
VI-VA-MS-SSIM	Uniform	0.9634	0.9583	4.1350	3.3506
VI-VA-MS-SSIM	Kent	0.9676	0.9635	3.8982	3.1526
VI-VMAF	–	0.9646	0.9581	4.2096	3.1548
VI-VA-VMAF	Equator-bias	0.9650	0.9576	4.1959	3.1393
VI-VA-VMAF	Uniform	0.9749	0.9671	3.5602	2.7569
VI-VA-VMAF	Kent	**0.9773**	**0.9717**	**3.3753**	**2.5948**

Table 6 Comparison of different temporal pooling methods for the combination of the frame scores applied in VI-VA-VMAF

Pooling	PLCC	SROCC	RMSE	MAE
Mean	0.9773	0.9717	3.3753	2.5948
Harmonic Mean	0.9775	0.9718	3.3681	2.5911
Min	0.9753	0.9705	3.4920	2.6887
Median	0.9761	0.9715	3.4093	2.6275
5th Percentile	0.9759	0.9708	3.4489	2.6437
10th Percentile	0.9776	0.9711	3.3636	2.5776
20th Percentile	0.9764	0.9714	3.3866	2.6041

Investigation of applying different visual attention estimation methods

The proposed quality framework can make use of different visual attention estimation methods, as the visual attention weights $v_{i,k}$ can be computed from any visual attention map generated by different algorithms. Here, we investigate the effect of three different visual attention methods on VI-VA-METRIC performance, namely, Kent distribution method [22], uniform viewport method, and equator-bias method. The first of the three estimation methods is based on Kent distribution, which is a Gaussian distribution defined on the surface of a unit sphere, as explained in [22]. With this method, we compute the visual attention maps using the viewport trajectories gathered in our subjective experiments and the default parameters proposed in [22]. For the second method, we also use viewport trajectories. In this method, each point of the viewport trajectories is replaced with a uniform viewport that is projected to ERP. The final visual attention map is obtained as the summation of the projected viewports. The last method does not require the viewport trajectories. Instead, it computes the visual attention map as a vertical bias from the equator defined by the Gaussian curve centered on the equator. Figure 7 shows the visual attention maps obtained with these three methods based on five discrete viewport positions.

Table 5 shows the performance of the Voronoi-based metrics integrated with visual attention. As can be noticed, both the Kent distribution method [22] and the uniform viewport method are able to improve the performance of the Voronoi-based metrics. On the other hand, the equator-bias method is capable to improve VI-PSNR and VI-SSIM, while the performance values of VI-MS-SSIM and VI-VMAF remain almost constant. In conclusion, these results show that adding a characterization of the actual parts of the ODV that are likely watched improves the performance of the Voronoi-based metrics. As can be seen from the table, the metrics of the proposed framework achieve the best performance when applying the Kent distribution method. Since this method is the most plausible and similar to the human eye-tracking results [21], it is expected to perform better than the other methods. Therefore, we use the visual attention maps estimated by the Kent distribution method in the rest of this paper.

Table 7 Performance evaluation of the selected existing metrics and our Voronoi-based metrics together with two projection formats, namely ERP and CMP. The best performance values are in bold

Metrics	PLCC	SROCC	RMSE	MAE
PSNR $_{ERP}$	0.8408	0.8237	8.2326	6.3169
PSNR $_{CMP}$	0.8480	0.8323	8.0419	6.2085
S-PSNR-I	0.8580	0.8438	7.8207	5.9715
S-PSNR-NN	0.8584	0.8433	7.8066	5.9648
WS-PSNR	0.8582	0.8430	7.8107	5.9772
CPP-PSNR	0.8579	0.8439	7.8200	5.9779
SSIM $_{ERP}$	0.7659	0.7551	9.7734	7.7396
SSIM $_{CMP}$	0.7701	0.7546	9.6583	7.6036
MS-SSIM $_{ERP}$	0.9224	0.9160	5.8232	4.4205
MS-SSIM $_{CMP}$	0.9132	0.9081	6.1422	4.7378
VMAF $_{ERP}$	0.8978	0.8864	6.7433	5.3631
VMAF $_{CMP}$	0.9063	0.8945	6.5630	5.2229
VI-PSNR	0.8676	0.8551	7.5743	5.8377
VI-SSIM	0.8823	0.8763	7.1172	5.2867
VI-MS-SSIM	0.9486	0.9450	4.8743	3.8475
VI-VMAF	0.9646	0.9581	4.2096	3.1548
VI-VA-PSNR	0.8876	0.8712	7.1818	5.5072
VI-VA-SSIM	0.9106	0.9007	6.4345	4.8097
VI-VA-MS-SSIM	0.9676	0.9635	3.8982	3.1526
VI-VA-VMAF	**0.9773**	**0.9717**	**3.3753**	**2.5948**

Investigation of different temporal pooling methods of the frame scores

Since the selection of the temporal pooling method P for the combination of the frame scores T_i (see Eq. 9) might affect the overall performance, in this paper, we also investigate its effect. For this purpose, motivated by the pooling methods which are used in VMAF code [58], we evaluate the following ones: mean, harmonic mean, min, median, 5th percentile, 10th percentile, and 20th percentile. Table 6 shows the performance of VI-VA-VMAF with these pooling methods. As can be noticed, the performance is not influenced too much by the choice of the pooling method. Therefore, in the rest of the paper, we consider only the mean pooling method.

Comparison with existing metrics

This section evaluates the performance of the Voronoi-based metrics and existing well-known metrics used in ODV quality assessment studies. Four of the existing metrics that we evaluate were developed for traditional 2D image/video quality assessment: PSNR, SSIM [43], MS-SSIM [44], and VMAF [39]. These metrics were applied to ODVs in two different formats, namely, ERP and CMP, and to distinguish them we use a subscript, e.g. PSNR $_{ERP}$ and PSNR $_{CMP}$. Moreover, we analyze extra four metrics which were specifically designed for ODV: S-PSNR-I [10], S-PSNR-NN [10], WS-PSNR [8], and CPP-PSNR [9]. The implementation used in our evaluation for PSNR, SSIM, and MS-SSIM is the one provided by the Video Quality Measurement Tool [59]; for VMAF we used the code provided by its developers [58]; while for S-PSNR-I, S-PSNR-NN, WS-PSNR, and CPP-PSNR, we used the 360Lib standard software [60].

Table 7 shows the performance evaluation of the selected existing metrics and our Voronoi-based metrics. By looking at the results, we can notice a slightly higher correlation between the subjective and metric scores when the metrics PSNR, SSIM, and VMAF are applied to the CMP format instead of the ERP format. The reason of this could be the lower projection distortions of CMP compared to ERP. We also observe that the performance of the PSNR-based metrics developed for ODV is better than the performance of the traditional PSNR. Furthermore, among all the evaluated metrics in Table 7, SSIM is characterized by the worst performance, even worse than PSNR. The reason might be that the inevitable projection distortions negatively affect the performance of SSIM, as some regions are stretched to much bigger areas (especially the top and bottom parts of ERP). Therefore, SSIM scores could be dominated by these regions, and this could cause SSIM to have lower correlation with subjective scores than PSNR, even though, for traditional 2D video, SSIM is much closer to human perception than PSNR. On the other hand, among the selected existing metrics that are not Voronoi-based, MS-SSIM and VMAF have the best performance. This is not unexpected, since these metrics, which have state-of-the-art performance for traditional 2D video [42], consider scaling and compression distortions that characterize our dataset. Between these two metrics, MS-SSIM is slightly better than VMAF for both projection formats. The reason can be explained with the fact that VMAF was neither modeled for 8K nor ODV.

The results also show that when the metrics are applied to planar Voronoi patches instead of the ERP and CMP formats, they achieve a better performance. This is expected because of the lower projection distortions of the planar Voronoi patches compared to ERP and CMP, and because of the similar angular resolutions of the patches and the HMD viewport. Moreover, as already noticed before, the Voronoi-based metrics integrated with visual attention (i.e., VI-VA-METRICs) achieve better performance than the corresponding ones without visual attention (i.e., VI-METRICs). The best performing metric among all compared is VI-VA-VMAF followed by VI-VA-MS-SSIM.

In addition to the numerical results, a statistical significance analysis of the difference between PLCC, SROCC, and RMSE of the quality metrics was conducted according ITU-T Recommendation P.1401 [61]. Figure 8 illustrates the statistical significance analysis of the evaluated metrics in Table 7. The

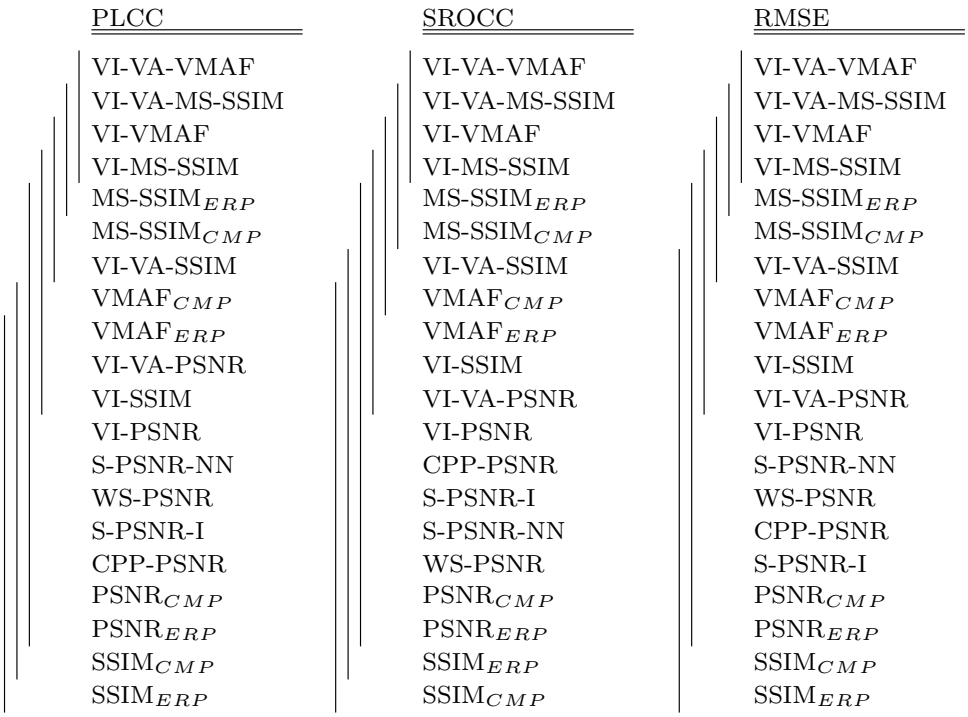


Fig. 8 Statistical significance analysis of the difference between PLCC, SROCC, and RMSE of the quality metrics, obtained according to ITU-T Recommendation P.1401 [61]. There is statistically significant equivalence between two quality metrics, only if there is a vertical bar aligned with them; e.g., there is a statistically significant difference between VI-VA-VMAF and MS-SSIM$_{ERP}$ in terms of PCC, SROCC, and RMSE

Table 8 PLCC and SROCC of the evaluated metrics computed separately for the resolutions 2K, 4K, and 8K. The best performance values for each resolution are in bold

	2K		*4K*		*8K*	
Metrics	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR $_{ERP}$	0.7388	0.6139	0.8360	0.8343	0.9202	0.9183
PSNR $_{CMP}$	0.7517	0.6203	0.8431	0.8450	0.9221	0.9163
S-PSNR-I	0.7634	0.6469	0.8568	0.8615	0.9304	0.9228
S-PSNR-NN	0.7649	0.6433	0.8570	0.8574	0.9300	0.9227
WS-PSNR	0.7650	0.6366	0.8570	0.8574	0.9299	0.9230
CPP-PSNR	0.7638	0.6432	0.8567	0.8615	0.9302	0.9230
SSIM $_{ERP}$	0.6996	0.5570	0.7703	0.7951	0.8600	0.8482
SSIM $_{CMP}$	0.7011	0.5591	0.7714	0.7878	0.8565	0.8484
MS-SSIM $_{ERP}$	0.8841	0.7992	0.9150	0.9351	0.9652	0.9478
MS-SSIM $_{CMP}$	0.8673	0.7824	0.9071	0.9276	0.9583	0.9446
VMAF $_{ERP}$	0.9202	0.8735	0.9203	0.9071	0.9515	0.9240
VMAF $_{CMP}$	0.9226	0.8790	0.9309	0.9156	0.9567	0.9285
VI-PSNR	0.7640	0.6321	0.8660	0.8769	0.9358	0.9247
VI-SSIM	0.8346	0.7109	0.8794	0.9060	0.9367	0.9249
VI-MS-SSIM	0.8642	0.8807	0.8140	0.9437	0.9767	0.9557
VI-VMAF	0.9627	0.9287	0.9577	0.9458	0.9789	0.9500
VI-VA-PSNR	0.7960	0.6644	0.9050	0.9006	0.9451	0.9321
VI-VA-SSIM	0.8434	0.7326	0.9200	0.9321	0.9593	0.9392
VI-VA-MS-SSIM	0.9529	0.9105	0.8332	**0.9674**	0.9829	**0.9634**
VI-VA-VMAF	**0.9762**	**0.9493**	**0.9737**	0.9625	**0.9862**	0.9593

vertical bars show that there is no statistically significant difference between the metrics aligned with the same bar. As can be noticed in Fig. 8, the first four best quality metrics are statistically equivalent. The significance analysis results also show that the addition of visual attention might not always yield a statistically significant difference. Nevertheless, the numerical results show that integrating visual attention improved the metric performance in all the cases, as we can also see in Table 4.

To further evaluate the Voronoi-based metrics in a different condition and analyze the effect of different spatial

resolutions of the ODVs, we calculate the correlation coefficients separately for each spatial resolution of our dataset (i.e., 2K, 4K, and 8K). The results of this analysis are shown in Table 8. It is interesting to notice that for most of the selected existing and Voronoi-based metrics the correlations PLCC and SROCC improve when the resolution is increased. This can be attributed to scaling distortions (blur) present at 2K and 4K resolutions. Assuming that most of the metrics were developed mainly for compression distortions and/or noise, the presence of scaling distortions could decrease the correlation between DMOS and metric scores in the cases of 2K and 4K. Nevertheless, we notice again that the integration of visual attention increases the performance of the Voronoi-based metrics.

Regardless of the case, the integration of visual attention (i.e., VI-VA-METRICs) improves the Voronoi-based metrics (i.e., VI-METRICs) in every situation. These improvements can be seen not only in Table 7 but also in Tables 5 and 8 . This consistent improvement shows that the proposed integration of visual attention is an important factor to consider in the objective ODV quality assessment, and it needs to be taken into account to increase the metric performance.

Limitations of the proposed framework and future improvements

As discussed in the previous subsection, the proposed framework integrated with visual attention achieves state-of-the-art performance. Nevertheless, it has also limitations that we plan to tackle in future work.

First, the current framework only considers visual attention maps generated using the viewport trajectories collected from the participants of a subjective experiment. In practice, this type of data is not available, as it is not possible to find the viewport trajectories for new content without conducting a subjective experiment first. Instead, automatic saliency estimation algorithms [62] might be used for most of the practical cases. Nevertheless, the integration of the said automatic saliency estimation methods and the performance analysis in this case remain as future work.

Second, in our study and in particular in our dataset, we have considered only the typical artifacts introduced by the encoding pipeline of the adaptive streaming systems, i.e., compression and scaling distortions. However, the end-to-end ODV distribution pipeline can introduce other visual artifacts [63, 64], such as artifacts introduced during capturing (e.g., noise and motion blur), stitching artifacts (e.g., visible seams and missing information), blending artifacts (e.g., ghosting and exposure difference), and warping artifacts. The perceptual impact of the other visual artifacts can be investigated and integrated into our proposed framework.

Third, with the current unoptimized code, the computation of VI-VA-VMAF requires considerable computational resources. For an 8K ODV with 300 frames, the computation of VI-VA-VMAF with 20 patches and with 10 pix/deg patch resolution takes about three minutes using a PC with a 4GHz Intel Core i7-6700K processor. Moreover, VI-VA-VMAF requires as input also a visual attention map for each frame. On a machine with two Intel Xeon Gold 6134 processors, the parallel computation of 400×800 visual attention maps using the code of the Kent method provided in [22] takes about nine seconds per map.

Conclusion

This paper presented a framework for objective ODV quality assessment that takes into account the spherical nature of ODV and the ODV viewing characteristics. The proposed framework is based on the subdivision of ODV into planar Voronoi patches with low projection distortions obtained with the spherical Voronoi diagram. Furthermore, it also exploits visual attention to identify the regions that are consumed by the viewer with high probability, which have a big influence on the perception of the video quality. For the evaluation of the framework, our previously established ODV dataset was extended in this study, creating a dataset with a total of 120 distorted videos from 8 undistorted reference videos. Subjective scores and viewport trajectories for the new ODVs were also collected in a subjective experiment.

In the evaluation of the framework, first the framework parameter space was analyzed. This analysis showed how planar Voronoi patches and visual attention are important to achieve high correlation between subjective and metric scores. Moreover, the framework was also compared with exiting metrics, and this showed that our framework can achieve state-of-the-art performance.

As future work, we plan to further explore the visual attention methods for ODV that do not require viewport trajectories. We also intend to extend our framework to distortions different from the ones considered here, i.e., compression and scaling distortions.

Acknowledgements This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776. This work has also been partially supported by the Ministerio de Economía, Industria y Competitividad (AEI/FEDER) of the Spanish Government under project TEC2016-75981 (IVME)

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Knorr S, Ozcinar C, Fearghail CO, Smolic A (2018) Director's cut—a combined dataset for visual attention analysis in cinematic VR content. In: The 15th ACM SIGGRAPH European conference on visual media production. https://doi.org/10.1145/3278471.3278472
2. Rana A, Ozcinar C, Smolic A (2019) Towards generating ambisonics using audio-visual cue for virtual reality. In: 44th international conference on acoustics, speech, and signal processing (ICASSP)
3. Ozcinar C, De Abreu A, Knorr S, Smolic A (2017) Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems. In: The 19th IEEE international symposium on multimedia (ISM 2017). Taichung, Taiwan
4. Warburton DE, Bredin SS, Horita LT, Zbogar D, Scott JM, Esch BT, Rhodes RE (2007) The health benefits of interactive video game exercise. Appl Physiol Nutrit Metabol 32(4):655–663
5. Freina L, Ott M (2015) A literature review on immersive virtual reality in education: state of the art and perspectives. In: The international scientific conference elearning and software for education, vol 1. " Carol I" National Defence University, p 133
6. Ozcinar C, De Abreu A, Smolic A (2017) Viewport-aware adaptive 360° video streaming using tiles for virtual reality. In: 2017 international conference on image processing (ICIP). Beijing, China
7. Möller S, Raake A (2014) Quality of experience: advanced concepts, applications and methods. Springer, Berlin
8. Sun Y, Lu A, Yu L (2017) Weighted-to-spherically-uniform quality evaluation for omnidirectional video. IEEE Signal Process Lett 24(9):1408–1412. https://doi.org/10.1109/LSP.2017.2720693
9. Zakharchenko V, Choi KP, Park JH (2016) Quality metric for spherical panoramic video. Proc SPIE 9970:9970. https://doi.org/10.1117/12.2235885
10. Yu M, Lakshman H, Girod B (2015) A framework to evaluate omnidirectional video coding schemes. In: 2015 IEEE international symposium on mixed and augmented reality, pp 31–36. https://doi.org/10.1109/ISMAR.2015.12
11. Li C, Xu M, Du X, Wang Z (2018) Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. CoRR **abs/1807.10990**
12. Upenik E, Ebrahimi T (2019) Saliency driven perceptual quality metric for omnidirectional visual content. In: 2019 IEEE international conference on image processing (ICIP), pp 4335–4339. https://doi.org/10.1109/ICIP.2019.8803637
13. Ozcinar C, Cabrera J, Smolic A (2019) Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality. IEEE J Emerg Sel Topics Circuits Syst 9(1):217–230. https://doi.org/10.1109/JETCAS.2019.2895096
14. Ye Y, Alshina E, Boyce J (2017) Algorithm descriptions of projection format conversion and video quality metrics in 360lib. Technical Report JVET-F1003, ISO/IEC JTC1/SC29/WG11/N16888, Hobart, AU
15. Sun W, Gu K, Ma S, Zhu W, Liu N, Zhai G (2018) A Large-Scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In: 2018 IEEE 20th international workshop on multimedia signal processing (MMSP), pp 1–6. https://doi.org/10.1109/MMSP.2018.8547102
16. Ozcinar C, Smolic A (2018) Visual attention in omnidirectional video for virtual reality applications. In: 10th international conference on quality of multimedia experience (QoMEX 2018). Sardinia, Italy
17. Singla A, Fremerey S, Raake A, List P, Feiten B (2017) AhG8: Measurement of user exploration behavior for omnidirectinal (360°) videos with a head mounted display. Technical report Macau, China
18. Gutiérrez J, David E, Rai Y, Le Callet P (2018) Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images. Signal Process Image Commun 69:35–42
19. David EJ, Gutiérrez J, Coutrot A, Da Silva MP, Callet PL (2018) A dataset of head and eye movements for 360° videos. In: Proceedings of the 9th ACM multimedia systems conference. ACM, pp 432–437
20. De Abreu A, Ozcinar C, Smolic A (2017) Look around you: saliency maps for omnidirectional images in VR applications. In: 2017 ninth international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
21. Rai Y, Le Callet P, Guillotel P (2017) Which saliency weighting for omni directional image quality assessment? In: 2017 ninth international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
22. John B, Raiturkar P, Le Meur O, Jain E (2018) A Benchmark of Four Methods for Generating 360° Saliency Maps from Eye Tracking Data. In: Proceedings of the first IEEE international conference on artificial intelligence and virtual reality. Taichung, Taiwan
23. Duan H, Zhai G, Min X, Zhu Y, Fang Y, Yang X (2018) Perceptual quality assessment of omnidirectional images. In: 2018 IEEE international symposium on circuits and systems (ISCAS), pp 1–5. https://doi.org/10.1109/ISCAS.2018.8351786
24. Luz G, Ascenso J, Brites C, Pereira F (2017) Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment. In: 2017 IEEE 19th international workshop on multimedia signal processing (MMSP), pp 1–6. https://doi.org/10.1109/MMSP.2017.8122228
25. Kim HG, Lim H, Ro YM (2019) Deep virtual reality image quality assessment with human perception guider for omnidirectional image. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2019.2898732
26. Aurenhammer F (1991) Voronoi diagrams—a survey of a fundamental data structure. ACM Comput Surv 23(3):345–405. https://doi.org/10.1145/116873.116880
27. Croci S, Knorr S, Goldmann L, Smolic A (2017) A framework for quality control in cinematic VR based on voronoi patches and saliency. In: International conference on 3D immersion. Brussels, Belgium
28. Croci S, Ozcinar C, Zerman E, Cabrera J, Smolic A (2019) Voronoi-based objective quality metrics for omnidirectional video. In: 11th international conference on quality of multimedia experience (QoMEX 2019)
29. Li C, Xu M, Zhang S, Callet PL (2019) State-of-the-art in 360° video/image processing: perception, assessment and compression. CoRR **abs/1905.00161**. http://arxiv.org/abs/1905.00161
30. Zhang Y, Wang Y, Liu F, Liu Z, Li Y, Yang D, Chen Z (2018) Subjective panoramic video quality assessment database for coding applications. IEEE Trans Broadcast 64(2):461–473. https://doi.org/10.1109/TBC.2018.2811627
31. Singla A Goring S, Raake A, Meixner B, Koenen R, Buchholz T (2019) Subjective quality evaluation of tile-based streaming for omnidirectional videos. In: 10th ACM multimedia systems conference (MMSys 2019)
32. Schatz R, Sackl A, Timmerer C, Gardlo B (2017) Towards subjective quality of experience assessment for omnidirectional video streaming. In: Proceedings of 9th international conference on quality multimedia expo. (QoMEX), pp 1–6
33. Ohm JR, Sullivan G (2011) Vision, applications and requirements for high efficiency video coding (HEVC). Technical Report MPEG2011/N11891, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland

34. Upenik E, Rerabek M, Ebrahimi T (2017) On the performance of objective metrics for omnidirectional visual content. In: 2017 ninth international conference on quality of multimedia experience (QoMEX)
35. Tran HTT, Ngoc NP, Bui CM, Pham MH, Thang TC (2017) An evaluation of quality metrics for 360 videos. In: 2017 ninth international conference on ubiquitous and future networks (ICUFN), pp 7–11. https://doi.org/10.1109/ICUFN.2017.7993736
36. Orduna M, Díaz C, Muñoz L, Pérez P, Benito I, García N (2019) Video multimethod assessment fusion (VMAF) on 360vr contents. CoRR **abs/1901.06279**
37. Upenik E, Reřábek M, Ebrahimi T (2016) A testbed for subjective evaluation of omnidirectional visual content. In: Proceedings of the picture coding symposium (PCS)
38. Chen S, Zhang Y, Li Y, Chen Z, Wang Z (2018) Spherical structural similarity index for objective omnidirectional video quality assessment. In: 2018 IEEE international conference on multimedia and expo (ICME), pp 1–6. https://doi.org/10.1109/ICME.2018.8486584
39. Li Z, Aaron A, Katsavounidis I, Moorthy A, Manohara M (2019) Toward a practical perceptual video quality metric. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652
40. Barman N, Schmidt S, Zadtootaghaj S, Martini MG, Möller S (2018) An evaluation of video quality assessment metrics for passive gaming video streaming. In: Proceedings of the 23rd Packet Video Workshop. ACM, pp 7–12. https://doi.org/10.1145/3210424.3210434
41. Rassool R (2017) VMAF reproducibility: Validating a perceptual practical video quality metric. In: 2017 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB), pp 1–2. https://doi.org/10.1109/BMSB.2017.7986143
42. Bampis CG, Li Z, Bovik AC (2018) Spatiotemporal feature integration and model fusion for full reference video quality assessment. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2018.2868262
43. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861
44. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems computers, 2003, vol 2, pp 1398–1402 . https://doi.org/10.1109/ACSSC.2003.1292216
45. Abbas A, Adsumilli B (2016) AhG8: New GoPro test sequences for virtual reality video coding. Technical Report JVET-D0026, JTC1/SC29/WG11, ISO/IEC, Chengdu, China
46. Asbun E, He H, Y, H, Ye Y (2016) AhG8: InterDigital test sequences for virtual reality video coding. Technical Report JVET-D0039, JTC1/SC29/WG11, ISO/IEC, Chengdu, China
47. Bang G, Lafruit G, Tanimoto M (2016) Description of 360 3D video application exploration experiments on divergent multiview video Technical Report MPEG2015/ M16129, JTC1/SC29/WG11, ISO/IEC, Chengdu, China
48. x265 HEVC Encoder / H.265 Video Codec. http://x265.org/ (2018)
49. FFmpeg. https://ffmpeg.org. Accessed 15 Jan 2019
50. HLS Authoring Specification for Apple Devices. https://developer.apple.com (2018)
51. Xu M, Li C, Chen Z, Wang Z, Guan Z (2018) Assessing visual quality of omnidirectional videos. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2018.2886277
52. https://github.com/Archer-Tatsu/Evaluation_VR-onebar-vive. Accessed 15 Jan 2019
53. Singla A, Fremerey S, Robitza W, Lebreton P, Raake A (2017) Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution. In: Proceedings of the on thematic workshops of ACM multimedia 2017, Thematic workshops '17. ACM, New York, NY, USA, pp 511–519. https://doi.org/10.1145/3126686.3126768
54. ITU-R: Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-13 (2012)
55. ITU-T: Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910 (2008)
56. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. IEEE Trans Image Process 19(6):1427–1441. https://doi.org/10.1109/TIP.2010.2042111
57. VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment. Technical report, ITU, COM 9-80-E, Geneva, Switzerland (2000)
58. Video multi-method assessment fusion (VMAF). https://github.com/Netflix/vmaf. Accessed 15 Jan 2019
59. Video quality measurement tool (VQMT). https://mmspg.epfl.ch/vqmt. Accessed 15 Jan 2019
60. 360lib. https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/trunk. Accessed 15 Jan 2019
61. ITU-T: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. ITU-T Recommendation P.1401 (2012)
62. Zhang Z, Xu Y, Yu J, Gao S (2018) Saliency detection in 360° videos: 15th European conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, pp 504–520. https://doi.org/10.1007/978-3-030-01234-2_30
63. Knorr S, Croci S, Smolic A (2017) A modular scheme for artifact detection in stereoscopic omni-directional images. In: Irish machine vision and image processing conference. Maynooth, Ireland
64. de Albuquerque Azevedo RG, Birkbeck N, Simone FD, Janatra I, Adsumilli B, Frossard P (2019) Visual distortions in 360-degree videos. CoRR **abs/1901.01848**