International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.919

(10/2020)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Audiovisual quality in multimedia services

Subjective test methodologies for 360º video on head-mounted displays

Recommendation ITU-T P.919

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	P.10–P.19
Voice terminal characteristics	P.30–P.39
Reference systems	P.40–P.49
Objective measuring apparatus	P.50–P.59
Objective electro-acoustical measurements	P.60–P.69
Measurements related to speech loudness	P.70–P.79
Methods for objective and subjective assessment of speech quality	P.80–P.89
Voice terminal characteristics	P.300–P.399
Objective measuring apparatus	P.500–P.599
Measurements related to speech loudness	P.700–P.709
Methods for objective and subjective assessment of speech and video quality	P.800–P.899
Audiovisual quality in multimedia services	**P.900–P.999**
Transmission performance and QoS aspects of IP end-points	P.1000–P.1099
Communications involving vehicles	P.1100–P.1199
Models and tools for quality assessment of streamed media	P.1200–P.1299
Telemeeting assessment	P.1300–P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400–P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500–P.1599

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.919

Subjective test methodologies for 360º video on head-mounted displays

Summary

Recommendation ITU-T P.919 describes subjective assessment methods for evaluating quality of experience of short (between 10 s and 30 s) 360º videos. Recommendation ITU-T P.919 also outlines the characteristics of the source sequences to be used, such as duration, type of content and number of sequences. Details within Recommendation ITU-T P.919 are expected to change in subsequent editions, based on experiments into how best to conduct subjective tests with 360º content.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.919	2020-10-14	12	11.1002/1000/14429

Keywords

360º video, methodology, QoE, quality of experience, subjective test.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.919

Subjective test methodologies for 360º video on head-mounted displays

1 Scope

This Recommendation addresses the subjective evaluation of 360° video viewed with head-mounted displays (HMDs) that enable interaction of three degrees of freedom (3DoF) in head movement to explore content. As such, 360° video that may possibly include spatial audio presentation is different from traditional audiovisual media such as television or movies. 360º Videos are captured by cameras having a 360º field of view (FoV) and are thus able to capture the surrounding scene at each instant in time. Typically, when users view 360º videos on an HMD, they can turn around to view the immersive 360º space from different angles. Coupled with the large FoV presented by an HMD, 360º video can essentially provide a more immersive experience than that achievable with traditional video.

This Recommendation describes methods to evaluate aspects of quality of experience (QoE) for 360° video. In general, the test methods recommended in this Recommendation utilize a hierarchical design, where the entire evaluation process is divided into three abstraction layers. This Recommendation may be used to compare 360º viewing sessions where stimuli may differ in terms of, for example, the recording technique applied, the processing (such as projection schemes, coding or rendering-specific aspects), andthe HMD devices used. This Recommendation describes subjective evaluation of short (between 10 s and 30 s) 360º videos. Topics include assessment methods, subjective scales, environmental conditions, equipment and data analysis. These experiments can assess phenomena such as audiovisual quality and simulator sickness.

2 References

The following ITU-T Recommendations and other references contain provisions, which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.800.2] Recommendation ITU-T P.800.2 (2013), *Mean opinion score interpretation and reporting*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.913] Recommendation ITU-T P.913 (2016), *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*.

[ITU-R BT.500-14] Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images.*

[ITU-R BT.2420-1] Recommendation ITU-R BT.2420-1 (2020), *Collection of usage scenarios and current statuses of advanced immersive audio-visual systems.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 double stimulus [ITU-T P.913]: A quality rating method where the subject is presented with two stimuli; the subject then rates both stimuli in the context of the joint presentation (e.g., a rating that compares the quality of one stimulus to the quality of the other).

3.1.2 field of view (FoV) [b-ITU-R RS.1804]: The swath width and all areas covered when the instrument is scanned.

3.1.3 hypothetical reference circuit (HRC) [ITU-T P.913]: A fixed combination of a video encoder operating at a given bit rate, network condition and video decoder. The term HRC is preferred when vendor names should not be identified.

3.1.4 processed [ITU-T P.913]: The reference stimuli presented through a system under test.

3.1.5 processed video sequence (PVS) [ITU-T P.913]: The impaired version of a video sequence.

3.1.6 quality of experience (QoE) [b-ITU-T P.10]: The degree of delight or annoyance of the user of an application or service.

3.1.7 reference [ITU-T P.913]: The original version of each source stimulus. This is the highest quality version available of the audio sample, video clip or audiovisual sequence.

3.1.8 sequence [ITU-T P.913]: A continuous sample of audio, video or audiovisual content.

3.1.9 single stimulus [ITU-T P.913]: A quality rating method where the subject is presented with one stimulus and rates that stimulus in isolation (e.g., a viewer watches one video clip and then rates it).

3.1.10 source [ITU-T P.913]: The content material associated with one particular audio sample, video clip or audiovisual sequence (e.g., a video sequence depicting a ship floating in a harbour).

3.1.11 spatial information [ITU-T P.913]: The amount of detail in a video, e.g., from high contrast edges, fine detail and textures.

3.1.12 stimulus [ITU-T P.913]: Audio sequence, video sequence or audiovisual sequence.

3.1.13 subject [ITU-T P.913]: A person who evaluates stimuli by giving an opinion.

3.1.14 temporal information [ITU-T P.913]: The amount of temporal change in a video sequence.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 head mounted display (HMD): A display worn on the body that fits over a user's head, which has small display optics in front of the eyes and is usually equipped with additional sensors to track the viewer's head motions such as coordinate positions, pitch, roll, and yaw. In some instances, the position of the user's gaze is also captured.

NOTE – Based on [ITU-R BT.2420-1],

3.2.2 refresh rate: The frequency with which a display updates an image.

3.2.3 simulator sickness: A physiological condition arising when exposed to a virtual reality environment

NOTE – Definition paraphrased from [b-Kennedy].

3.2.4 three degrees of freedom (3DoF): Programme material in which the user can freely look around in any direction (yaw, pitch, and roll). A typical use case is a user sitting in a chair looking at 3D VR/360° content on a head-mounted display.

NOTE – Based on [ITU-R BT.2420-1],

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2D	two Dimensional
3D	three Dimensional
3DoF	3 Degrees of Freedom
ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DSIS	Double Stimulus Impairment Scale
FoV	Field of View
HMD	Head-Mounted Device
MOS	Mean Opinion Score
PVS	Processed Video Sequence
QoE	Quality of Experience
RGB	Red–Green–Blue
SI	Spatial Information
SSQ	Simulator Sickness Questionnaire
TI	Temporal Information
VR	Virtual Reality
VRSQ	Virtual Reality Sickness Questionnaire
VSR	Vertigo Score Rating
YUV	luminance–blue luminance–red luminance

5 Conventions

None.

6 Selection of 360º source content

In order to evaluate 360º video quality and other terms defined in this Recommendation in various circumstances, the content should cover a wide range of stimuli. In particular, 360º content with a variety of spatial and temporal complexity, motion, and exploratory properties (in terms of focus of attention) should be used for accurate assessment.

The 360º videos should be selected according to the goal of the test and recorded on a digital storage system. When the experimenter is interested in comparing results from different laboratories, it is necessary to use a common set of source stimuli to eliminate a further source of variation.

The selection of the test material should be motivated by the experimental question addressed in the study. Examples of content types for 360º video experiments are provided in clause 4.2 of [ITU-R BT.2420-0].

This Recommendation covers the use of monoscopic 360º video content.

6.1 Source signals recordings

The source signal provides the reference stimuli and the input for the system under test.

The quality of the reference stimuli should be as high as possible. If available, pristine, uncompressed reference files shall be used in the target maximum resolution both in terms of frame rate and spatial resolution, with 4:2:2 or 4:4:4 chroma encoding in the luminance (Y)–blue luminance (U)–red luminance (V) (YUV) colour space at a minimum of 8 bits, or in the red–green–blue (RGB) colour space at a total of 24 or 32 bits. The resolution of the source videos should be at least 3 840 × 1 920 pixels.

It is noted that for viewing 360° videos, the HMD for testing provides further constraints on the resolution of the source videos. To retain the immersive characteristics and guarantee a precise perception of the quality of panoramic videos, the subjective quality assessment should be conducted with HMDs rather than plane screen monitors. Immersion requires that the 360°content can fill the entire FoV in HMD. Unlike the display of traditional two dimensional (2D) video, which can be presented in a per-pixel manner on the screen with fixed size by adding black pixels to the low resolution video or showing only part of the high resolution content, the 360° video must be presented in its entirety despite different resolutions [b-Zhang].

6.2 Spatial and temporal information

The selection of test scenes is an important issue. In particular, the spatial and temporal perceptual information of the scenes are critical parameters. These parameters play a crucial role in determining the amount of video compression that is possible, and consequently, the level of impairment that is suffered when the scene is transmitted over a fixed-rate digital transmission service channel. Relevant video test scenes must be chosen such that their spatial and temporal information is consistent with the video services that the digital transmission service channel is intended to provide. The set of test scenes should span the full range of spatial and temporal information of interest to users of the devices under test.

[ITU-T P.910] specifies simple metrics to estimate spatial information (SI) and temporal information (TI).

Existing measures for signal characterization are applied on planar representations of the 360º content (e.g., equirectangular or cube-map). The geometrical domain used to compute spatial and temporal indicators may have an influence in the characterization [b-DeSimone]. For example, the SI computed in equirectangular projection might take on a misleading perceptual characterization, due to the strong warping of the visual content around the poles. Similarly, considering the SI computed on the mosaicked cube-map planar images, vertical and horizontal edges corresponding to discontinuities at the borders between cube faces are taken into account, even if they are not features of the 360º content. To avoid this problem the computation of SI and TI can be done on each cube face separately considering the mean value across all faces as a measure of the spatial complexity of the entire 360° image.

Moreover, a single spatial complexity value might not be informative enough to characterize the entire 360º frames. Some content may show a significant variability in terms of SI and TI depending on the viewing direction (i.e., across cube faces), so the portion of 360º content attended by the user can have very different spatial complexity. To account for this variability, the variance of SI and TI over the cube faces can be considered. To select images having different characteristics, a suggested approach is to select images that have different variability across cube faces [b-DeSimone].

Finally, SI and TI can be calculated in the spherical domain. For each sampling point (m_i, n_i) on the equirectangular projected plane, it is first re-projected on to the sphere to get its corresponding longitude and latitude coordinates (φ_i, θ_i). The Sobel filter is then applied on the 3×3 window centred around the point on the sphere. Details of the computation of SI and TI in the spherical domain are provided in Appendix I.

How the spatial and temporal information is computed should be clearly reported.

6.3 Exploratory information

When viewing 360° videos, viewers explore the content (moving the head and eyes) according to the regions of interest in the presented videos. In the test, it should be ensured that the source sequences evoke a variety of different types of exploration behaviour.

The underlying property of videos is typically referred to as "saliency". With 360º content it can be obtained : a) from the positions of the head (and thus the centre of the FoV in the HMD) when it is referred to as "head saliency"; or b) from the positions of the eyes when using eye-trackers, when it is referred to as "head-eye saliency" [b-David]. In a test on audiovisual quality, source sequences that result in different saliency patterns should be used, from exploratory (nothing in the scene clearly catches the observers' attention) to focused contents (some objects stand out directing the observers' attention) [b-DeSimone].

For this purpose, subjective assessments by experts can be used, as well as objective measures for exploration or attention behaviours (e.g., similarity ring metric [b-ETSI TR 126 918] or entropy and inter-observer congruency [b-DeSimone]).

6.4 Comfort and simulator sickness symptoms

The content of the selected 360° sequences should be comfortable to watch and should not contain violent, sexual or disturbing content. Moreover, the selected sequences should not produce a high amount of simulator sickness in participants.

6.5 Duration of stimuli

The methods in this Recommendation are intended for stimuli that range from 10 s to 30 s in duration. Sequences of 10 s are recommended for assessing audiovisual quality.

Special attention should be paid when using sequences with time-varying properties (e.g., scene transitions). In addition, when dealing with non-uniform degradations (e.g., tile-based encoding), exploration patterns may change quality ratings between shorter and longer sequences.

6.6 Audio considerations

To evaluate video quality it is possible to use test stimuli either with or without audio. For video-only experiments, the missing audio should not be considered as an impairment.

When using non-uniform (e.g., tile-based encoding) degradations, audio coming from out of the region of interest (especially when using spatial audio) may influence audiovisual quality ratings.

7 Test methods

Measurement of the perceived quality of video requires the use of subjective scaling methods. The condition for such measurements to be meaningful is that there exists a relation between the physical characteristics of the stimulus, in this case the 360º video sequence presented to the subjects in a test, and the magnitude and nature of the sensation caused by the stimulus. The final choice of one of these methods for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed.

Subjective experiments with 360º video may measure opinions on different perceptual scales:

– audiovisual quality;
– simulator sickness symptoms;
– exploration behaviour;
– presence, emotion response and other factors may be measured, but are not covered in this Recommendation.

These perceptual scales must be rated independently. This clause describes the test methods and rating scales. The method controls the sequence presentation. The rating scale controls the way that people indicate their opinion of the sequences.

7.1 Test methods for audiovisual quality

These methods are appropriate for evaluating video quality in subjective experiments on 360º video.

7.1.1 Absolute category rating

The absolute category rating (ACR) method is a category judgement where the test sequences are presented one at a time and rated independently on a category scale. ACR is a single stimulus method. The subject observes one sequence and then has time to rate that sequence.

The ACR method uses the following five-level rating scale:

– 5 excellent;
– 4 good;
– 3 fair;
– 2 poor;
– 1 bad.

The numbers may optionally be displayed on the scale.

7.1.1.1 Comments

The ACR test method can be used when testing time is of relevance, since it produces a high number of ratings in a brief period of time [b-Singla].

7.1.2 Degradation category rating method

The degradation category rating (DCR) method presents sequences in pairs. The first stimulus presented in each pair is always the reference. The second stimulus is the same reference sequence after impairment by the systems under test. DCR is a double stimulus method. The DCR method is also known as the double stimulus impairment scale (DSIS) method. In this case, subjects are asked to rate the impairment of the second stimulus in relation to the reference. The following five-level scale for rating the impairment should be used:

– 5 imperceptible;
– 4 perceptible, but not annoying;
– 3 slightly annoying;
– 2 annoying;
– 1 very annoying.

The numbers may optionally be displayed on the scale.

7.1.2.1 Comments

The DCR method produces fewer ratings than ACR in the same period of time.

DSIS is statically more reliable than the ACR [b-Singla].

7.2 Test methods for simulator sickness symptoms

Simulator sickness is an undesirable phenomenon that is caused by the sensory conflict arising between the visual and vestibular systems. Simulator sickness is an accumulative factor, and therefore it should be assessed before and after each active viewing period. Additionally, experimenters may require measurement of simulator sickness at other moments in the session (e.g., periodically within the active period).

The following are appropriate to evaluate simulator sickness on 360º video.

7.2.1 Simulator sickness questionnaire

The simulator sickness questionnaire (SSQ) [b-Kennedy, 1993] is the recommended questionnaire for assessing simulator sickness symptoms. Subjects should complete the SSQ immediately before and after each active viewing period. Subjects must assess how much a symptom is affecting them at the moment they are being asked ("right now"), among the following:

– [GD] general discomfort;
– [FA] fatigue;
– [HE] headache;
– [ES] eyestrain;
– [DF] difficulty focusing;
– [IS] increased salivation;
– [SW] sweating;
– [NA] nausea;
– [CO] difficulty concentrating;
– [FH] fullness of head;
– [BV] blurred vision;
– [DO] dizzy (eyes open);
– [DC] dizzy (eyes closed);
– [VE] vertigo;
– [SA] stomach awareness;
– [BU] burping.

Each symptom must be rated in the following scale:

– 0 none;
– 1 slight;
– 2 moderate;
– 3 severe.

The numbers may optionally be displayed on the scale.

From the responses of the SSQ, four measurements are obtained:

– Nausea (N) = 9.54 (GD + IS + SW+ NA + CO + SA + BU)
– Oculomotor (O) = 7.58 (GD + FA + HE + ES + DF + CO + BV)
– Disorientation (D) = 13.92 (DF + NA + FH + BV + DO + DC + VE)
– Total Score (TS) = 3.74 (N/9.54 + O/7.58 + D/13.92)

7.2.2 Virtual reality sickness questionnaire

In the visualization of a 360° video, not all symptoms included in SSQ are equally prevalent. The previously described SSQ may be replaced by a shorter version of the SSQ called the virtual reality sickness questionnaire (VRSQ) [b-Kim] considering only the following symptoms: general discomfort, fatigue, eyestrain, difficulty focusing, headache, fullness of head, blurred vision, dizziness with eye closed and vertigo.

From the responses of VRSQ, the following scores can be extracted:

$$\text{Oculomotor (O)} = 100 \times \left(\frac{\text{GD+FA+ES+DF}}{12}\right)$$

$$\text{Disorientation (D)} = 100 \times \left(\frac{\text{FH+HE+BV+DC+VE}}{15}\right)$$

7.2.3 Vertigo score rating

If, according to the purpose of the specific experiment, simulator sickness needs to be assessed frequently (e.g., periodically within the active visualization period), it is recommended to use the single-scale question or fast self-report methods.

The vertigo score rating (VSR) [b-Pérez] is the recommended five-level scale for fast self-reporting of simulator sickness. Subjects should respond to the question, "Are you feeling any sickness or discomfort now?" according to the following scale:

– 5 no problem (no perceptible effect, natural feeling);

– 4 light effects (slight discomfort, but no sickness);

– 3 uncomfortable (moderate discomfort, but tolerable for a while);

– 2 unpleasant (strong discomfort or sickness, but can continue the test);

– 1 unbearable (strong discomfort or sickness, and want to stop test).

The numbers and the score explanation within parentheses may optionally be displayed on the scale.

7.3 Test methods for exploration behaviour

Exploration behaviour is tested by recording the head rotation position of the subject along the duration of the active viewing session. This recording is done by an application running in the subject HMD, normally the same application used to display the videos.

Head rotation position should be recorded at regular intervals with a frequency of at least 30 Hz, and it must be time referenced to the start of the presentation of each video sequence, so that it is possible to relate exploration behaviour with the content the subject was watching at each moment of time.

Similarly, eye movements can be recorded using eye trackers integrated in the HMDs [b-David].

7.4 Methods to collect the observers' scores

A relevant difference of 360º video visualization with respect to previous subjective evaluation methodologies is that subjects cannot use conventional scoring methods (e.g., paper, sliders) while they are wearing the HMD (the active viewing period). During the active viewing period, two possible procedures are recommended: a voting interface in the video player or verbal voting.

A voting interface is a simple virtual reality (VR) application which, after the playback of each of the sequences in the HMD, displays the scoring scale (ACR, DCR, etc.) and requests the response of the user via gaze or a handheld control. Subjects should be able to select the desired response in at most 5 s. Responses must be recorded by the interface for their processing.

With verbal voting, the scoring scale is displayed on the HMD for 5 s, and the subject is requested to verbally declare the score. In such a case, the test moderator should record the ratings of users.

Voting interface or verbal voting are recommended for evaluations performed within the active viewing period: audiovisual quality and VSR.

In the rest periods, i.e., where subjects are not wearing the HMD, conventional voting methods can be used (paper, computer applications, etc.). These methods are recommended for SSQ and RSSQ.

8 Environment and equipment

8.1 Test environment

A controlled environment should represent a non-distracting environment where a person would reasonably use the device under test. In this Recommendation, the test should be carried out in an environment without noise that can annoy or influence the observer when performing the test.

The observer should be seated on a swivel chair, in order to be able to freely rotate to explore the 360º video.

The test moderator should remain with subjects (in the same room without influencing the observer or in an adjoining room), due to concerns of simulator sickness, and ensure subjects halt the test when feeling symptoms, despite not finishing the session.

The environment must be documented.

8.2 Equipment

Any commercial HMD (tethered or untethered) can be used, provided that it has enough resolution and refresh rate to represent the content to be tested. A minimum resolution of 1 080 × 1 200 pixels per eye is required for tethered HMDs. For wireless or untethered HMDs, normally, a separate display device is required, such as a phone. The display resolution of a phone should be at least 2 560 × 1 440 pixels. A minimum refresh rate of 60 Hz is required. When possible, 90 Hz or higher is recommended.

9 Subjects

At least 28 subjects must be used for experiments conducted in a controlled environment. Details of the statistical analysis performed to obtain the minimum sample size are provided in Appendix II.

10 Experiment design

10.1 Inclusion of reference conditions within the experiment

The results of quality assessments often depend not only on the actual video quality, but also on factors such as the total quality range of the test conditions, and the experience and expectations of the assessors. In order to control some of these effects, a number of dummy test conditions can be added and used as references.

Some of the methods listed in clause 7.1 include a reference sequence, whenever available, as part of the test sequence set. The reference is usually a version of the test sequence that has not undergone any processing (i.e., the original source sequence).

10.2 Size of the experiment and subject fatigue

The size of an experiment is typically a compromise between the conditions of interest and the amount of time individual subjects can be expected to observe and rate stimuli.

Preferably, an experiment should be designed so that each subject's participation is limited to 1.5 h, of which no more than 50 min is spent rating stimuli, and no more than 25 min continuously. When larger experiments are required, frequent breaks and adequate compensation should be used to

counteract the negative impacts of fatigue and boredom. The number of times that each source stimulus is repeated also impacts subject fatigue. Among different possible test designs, preferably choose the one that minimizes the number of times a given source stimulus is shown.

11 Experiment implementation

11.1 Informed consent

Subjects should be informed of their rights and be given basic information about the experiment. It may be appropriate for subjects to sign an informed consent form. In some countries, this is a legal requirement for human testing.

Recommendations on the information that should be provided to the participants and a sample consent form are provided in Appendix III.

11.2 Viewer screening

Pre-screening procedures include tests of vision and colour blindness.

Prior to a session, the observers should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal colour vision. Concerning acuity, no errors on the 20/30 line of a standard eye chart [b-Snellen] should be made. The chart should be scaled for the test viewing distance and the acuity test performed at the same location from where the video images will be viewed (i.e., lean the eye chart against the monitor) and have subjects seated. Concerning colour, the [b-Ishihara] colour vision test should be passed.

Subjects who fail such screening should preferably be run through the experiment with no indication given that they failed the test. The data from such subjects should be discarded when a small number of subjects are used in the experiment. Data from such subjects may be retained when a large number of subjects is used.

11.3 Post-screening of subjects

Post-screening of subjects may or may not be appropriate depending upon the purpose of the experiment. The following subject screening methods may serve as reference: clause A1-2.3 of [ITU-R BT.500-14], Annex A of [ITU-T P.913], and questionnaires or interviews after the experiment to determine whether the subject understood the task. It should be noted that these screening methods might lead to different subject screening results.

When subjects are eliminated due to post-screening, it may be appropriate to present the data of screened subjects separately or to analyse the data both with and without the screened subjects.

The final report should include a detailed description of the screening methodology.

11.4 Instructions and training

Instruction should be tailored to the dimension (e.g., audiovisual quality or simulator sickness) under investigation. The instructions must tell subjects what to do when discomfort, dizziness or simulator sickness is experienced. A possible text for instructions to be given to the assessors is suggested in Appendix IV.

Ethical guidelines are critical, since participants might experience discomfort, dizziness, simulator sickness, etc. The subjects must be informed of any possible negative effect resulting from exposure to the stimuli used in the study. The subjects must be told that they can stop the test at any point, without negative consequence (e.g., the subject may leave the test chamber in the middle of the experiment and still be paid in full, if payment is foreseen).

Subjects must have a period of training in order to get familiar with the test methodology and software and with the kind of factors (e.g., audiovisual quality) they have to assess. The training phase is a

crucial part of this method, since subjects could misunderstand their task. Written instructions should be used to ensure that all subjects receive exactly the same information. The instructions should include explanations about what the subjects are going to see and hear, what they have to evaluate (e.g., difference in quality) and how to express their opinion.

Questions about the procedure and meaning of the instructions should be answered with care to avoid bias. Questions about the experiment and its goals should be answered after the final session.

After the instructions, a training session should be run. The training session is typically identical to the experiment sessions, yet short in duration. Stimuli in the training session should demonstrate the range and type of impairments to be assessed. Training should be performed using stimuli that do not otherwise appear in the experiment. In addition, the scores collected during the training session should be discarded and not considered for the data analysis.

The purpose of the training session is to: 1) familiarize subjects with the voting procedure and pace; 2) show subjects the full range of impairments present, thus stabilizing their votes; 3) encourage subjects to ask new questions about their task, in the context of the actual experiment; 4) if necessary, adjust the audio playback level, which will then remain constant during the test phase; 5) help the observers to put and adjust correctly the HMD (including focus and inter-pupillary distance). For a simple assessment of audiovisual quality in absolute terms, a small number (e.g., four to six) of stimuli in the training session may suffice. For more complicated tasks, the training session may need to contain a large number of stimuli.

11.5 Experiment sessions and breaks

Ideally, no session of active viewing of 360º video should last for more than 25 min, and in no case should the active viewing time exceed 50 min. At least, every 25 min, subjects should be asked to take a break of at least 15 min.

The stimuli should be presented in a pseudo-random order.

The pattern within each session (and the training session) is as follows: play sequence; pause to score; repeat. The specific pattern and timing of the experimental sessions depend upon the playback mechanism.

11.6 Questionnaire or interview

For some experiments, questionnaires or interviews may be desirable either before or after the subjective sessions. The goal of the questionnaire or interview is to supplement the information gained by the experiment. Examples include:

– demographics that may or may not influence the votes, such as age, gender, television watching habits, experience using VR devices;

– feedback from the subject after the sessions;

– quality of experience observations on deployed equipment used by the subject (i.e., service observations).

The disadvantage of the service observation method for many purposes is that little control is possible over the detailed characteristics of the system being tested. However, this method does afford a global appreciation of how the equipment performs in the real environment.

A possible questionnaire to be given to the assessors is suggested in Appendix V.

12 Data analysis

The results should be reported along with the details of the experimental setup. Clause 12 of [ITU-T P.800.2] specifies the minimum information that should accompany mean opinion score (MOS) values to enable them to be correctly interpreted.

For each combination of test variables, the MOS and standard deviation of the statistical distribution of the assessment grades should be given. Some items can be mandatory, while others need to be reported whenever possible. The calculation of these statistical values is described in [ITU-R BT.500-14]. [ITU-T P.800.2] provides additional information about MOSs.

12.1 Calculate mean opinion score or differential mean opinion score

After all subjects are run through an experiment, the ratings for each clip are averaged to compute either a MOS or a differential mean opinion score (DMOS).

Use of the term MOS indicates that the subject rated a stimulus in isolation. The following methods can produce MOS scores:

– ACR;

– absolute category rating with hidden reference (ACR-HR; using raw ACR scores).

Use of the term DMOS indicates that scores measure a change in quality between two versions of the same stimulus (e.g., a source video and its processed version). The following methods can produce DMOS scores:

– ACR-HR;

– DCR.

[ITU-T P.800.2] provides additional information about MOSs.

12.2 Analysis of exploration data

Recommendations on how to analyse and present exploration data are provided in Appendix VI.

13 Elements of subjective test reporting

Reports on subjective testing are more effective when descriptions of both mandatory and optional elements defining the test are included. A full description of all the elements of the subjective test supports the conclusions from the test.

The goal is that the reader can reproduce the experiment and, by following the specified procedure, be expected to reach the same conclusions.

13.1 Documenting the test design

The description of the test design needs to list the details of the stimuli (source reference circuits (SRCs), the impairments (HRCs), and the reasoning for choosing those stimuli and HRCs. Any details that are non-traditional need to be discussed thoroughly.

Definitions of the source stimuli must include the type or subject matter of the video and audio, signal format, number of clips, range of video coding complexities, mechanism used to obtain stimuli and quality of the original recordings. Impairment choices should flow from and support the goal of the test. As in the description of the SRCs, descripttions of the HRCs must include the type and number of HRCs, with sufficient technical details to enable the reader to reproduce these impairments (e.g., codec, bit-rate, encoding options or processing chain). The software or hardware used to process or record the PVSs should also be specified.

Central to the test is the HMD used by subjects. The specifications of the device (i.e., resolution, refresh rate, etc.) should be reported (see clause 8.2).

Specify the method used to record scores. If automated scoring is used, describe the device and software.

Identify the test method and rating scale. The report of the test should describe the test method type, including the type of stimuli (single, double, multiple) and the rating scale used. Any changes to the methods should be noted in the report.

13.2 Documenting subjective testing

The clause of the test report that specifies the subjective test situation must describe three elements: 1) the participants; (2) the environment; and (3) the mechanism used to present the stimuli. Furthermore, the report needs to include the time duration for the test sessions as well as the dates and times of the test.

The report needs to state the number of participants, as well as the distributions of their ages and genders. Preferably, the instructions to participants are included. If insufficient space exists, the subject instructions may be summarized.

The subjective test environment must be reported as well as whether the users were sitting in a swivel chair.

A description of the hardware and software used to present the stimuli is essential to the test report. Details on the hardware (e.g., HMD) help define any effect it may have on the results. Include a brief description of the program used to play the source stimuli. It is important to understand the post-processing of HRCs that was required to enable playback on the HMD.

13.3 Data analysis

The report should include the process used to calculate the MOS or DMOS as specified in clause 12. It is important to incorporate the minimum information from clause 12 of [ITU-T P.800.2]. Of particular importance are details of the methodology of the test when not using methods specified in ITU Recommendations or when modifying methods defined in ITU Recommendations.

13.4 Additional information

Any deviation from the methods specified in this Recommendation must be described in detail.

A test report can also contain design and results of pilot testing and pre-testing, as appropriate.

Appendix I

Spatial and temporal information measurement for 360º video in the spherical domain

(This appendix does not form an integral part of this Recommendation.)

This appendix introduces the measurement of spatial and temporal information for 360º video. Considering planar representations (equirectangular, cube-map, etc.) change the characterization of the 360º content because of warping, discontinuities, etc., spatial and temporal information measurement in 2D cannot represent the information subjects perceived using HMD. Here, calculation of the SI and TI specified in [ITU-T P.910] in the spherical domain is recommended.

Spatial information measurement

The SI is based on the Sobel filter in the spherical domain. The sphere can be sampled with longitude (ϕ) and latitude (θ). The longitude ϕ is in the range [−π, π], and latitude θ is in the range [−π/2, π/2]. For each sampling point (*m*, *n*) on the 2D plane F_n of size $M \times N$, it is first converted to the longitude and latitude (ϕ, θ) [b-Chen].

The Sobel filter is then applied on the 3 × 3 window centred around the point (ϕ, θ) on the sphere, which can be obtained according to the location of the centre (ϕ, θ) and the angle α_0, β_0 occupied by each point on the sphere:

$$\begin{bmatrix} (\phi - \alpha_0, \theta + \beta_0) & (\phi, \theta + \beta_0) & (\phi + \alpha_0, \theta + \beta_0) \\ (\phi - \alpha_0, \theta) & (\phi, \theta) & (\phi + \alpha_0, \theta) \\ (\phi - \alpha_0, \theta - \beta_0) & (\phi, \theta - \beta_0) & (\phi + \alpha_0, \theta - \beta_0) \end{bmatrix} \tag{I.1}$$

$$\alpha_0 = \frac{2\pi}{M}, \beta_0 = \frac{\pi}{N} \tag{I.2}$$

By applying the Sobel filter to all the spherical points ($\mathrm{Sobel_s}$) converted from frame F_n, the weighted standard deviation over all the pixels ($s^{\mathrm{w}}_{\mathrm{space}}$)is computed. The maximum value in the time series ($t_{\max}$) is chosen to represent the SI content of the video. This process is represented by Equation (I.3):

$$\mathrm{SI} = t_{\max}\left\{s^{\mathrm{w}}_{\mathrm{space}}[Sobel_s(F_n)]\right\} \tag{I.3}$$

where $s^{\mathrm{w}}_{\mathrm{space}}$ is calculated by:

$$s^{\mathrm{w}}_{\mathrm{space}}(X) = \sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{\left(X(m,n)\cdot w(m,n) - \mu^{\mathrm{w}}_{\mathrm{space}}(X)\right)}{\sum_{m=1}^{M}\sum_{n=1}^{N} w(m,n)}} \tag{I.4}$$

$$\mu^{\mathrm{w}}_{\mathrm{space}}(X) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(X(m,n)\cdot w(m,n)\right)}{\sum_{m=1}^{M}\sum_{n=1}^{N} w(m,n)} \tag{I.5}$$

where $\mu^{\mathrm{w}}_{\mathrm{space}}(X)$ is the mean. The calculation of weight $w(m,n)$, denoting the spherical area covered by each position on the 2D projection plane, is dependent on the projection format. Weight derivation for each projection format is discussed in [b-Ye].

Temporal information measurement

The measure of TI is based upon the motion difference feature M_n, which is the pixel difference between two subsequent frames, F_n and F_{n-1}.

$$M_n = F_n - F_{n-1} \tag{I.6}$$

The TI is computed as the maximum over time ($t_{\max}$) of the weighted standard deviation over space ($s^{\mathrm{w}}_{\mathrm{space}}$) of M_n over all the points:

$$\mathrm{TI} = t_{\max}\left\{s^{\mathrm{w}}_{\mathrm{space}}[M_n]\right\} \tag{I.7}$$

NOTE – As described in [ITU-T P.910], for relevant scenes, scene cuts can be either included or excluded for the temporal information measurement, resulting in two values.

Appendix II

Computation of sample size from statistical power

(This appendix does not form an integral part of this Recommendation.)

In hypothesis testing, statistical power refers to the probability of rejecting the null hypothesis, when the alternative hypothesis is true. It is inversely correlated to the concept of type II error (wrongly failing to reject the null hypothesis). For a probability β of witnessing a type II error, the statistical power is equal to $1 - \beta$. Statistical power depends on the type of hypothesis testing, on the effect size, and on the sample size.

The minimum sample size can be computed by fixing the desired statistical power, in order to be reasonably assured of correctly detecting an effect of a given size. Free software, such as G*Power [b-Faul], can be used to compute the minimum sample size for a given statistical test.

For the specific case of a statistical test aiming to determine whether one distortion leads to higher MOS scores with respect to another, where that the same subjects will rate both distortions, the Wilcoxon signed-rank statistical test with one tail might be the most appropriate. Assuming a type I error probability $\alpha = 0.05$, and an effect size of $r = 0.5$, and fixing the statistical power to be $1 - \beta = 0.8$, we obtain a minimum sample size of $n = 28$.

Appendix III

Recommendations for information sheet and consent form

(This appendix does not form an integral part of this Recommendation.)

General purpose of the information sheet

Information sheets should cover basic information about a subjective experiment that is going to be conducted. Subjects should be concisely informed about the purpose of the study and the importance of the data collected during the experiment. It is mandatory to deliver an information sheet to the subject before signing a consent form.

Content of an information sheet

1) Title of the study/project
2) Principal researcher/coordinator (name, email, job title)
3) Institution
4) Session structure (different parts and breaks) and expected duration
5) Location
6) Document identification (date, version, short ID etc.)
7) What is the purpose of this research study? How will the data be processed?
8) Who can take part in this study? (requirements) How many participants will there be?
9) Why should you consider joining this study as a research subject?
   - What kinds of benefit can you expect personally from taking part in this study?
   - What kinds of benefit to others can come out of this study?
10) Do you have to become a subject in this study? If you joined the study, can you change your mind and drop out before it ends?
11) Are there any risks involved? *(For some people, immersive 360 videos may give some temporary discomfort or nausea, which will go away shortly after finishing watching the video, but for most people there are no problems encountered. This test will not have an influence on your physical health. However, for some people it can lead to an epileptic seizure if they are confronted with certain visual stimuli. If during the test under any circumstances, if symptoms like dizziness, odd perception, eye or muscle twitches, shivering arms or legs, disorientation or confusion appear, please inform your supervisor immediately.)*
12) What exactly will be done to you if you agree to be a research subject in this study? (What is involved? What kind of personal data we need to collect?)
   - Please include such information:
     - *For testing if the head-mounted display is appropriately mounted on your head, the test supervisor may touch your head.*
     - *In the scope of the study, the given quality ratings, other given data on the questionnaires filled in and head-rotation data, which are recorded during the experiment.*
13) What will the researchers do to make sure that the information they will collect on you will not get into the wrong hands? (storage) Who will be responsible? Is it confidential/anonymous? Who gets to keep this document, once you sign it?
14) Will you get paid for taking part in this study? (Is there a reward/compensation for participation?)
15) Who is organizing and funding the research?
16) What is the legislation that this research project complies with?

Sample of consent form

Title of study / project:

_____

Participant details:

– First and surname: _____

– Passport/ID: _____

Principal researcher, email: _____

Institution: _____

By ticking/initialling each box you are consenting to this element of the study. It will be assumed that un-ticked/un-initialled boxes mean that you DO NOT consent to that part of the study and you may be deemed ineligible for the study.

No	Consent item	Please, tick or initial
1	I confirm that I have read and understood the participants information sheet __*(document ID of information sheet)* _____ for the above study. I have had the opportunity to consider the information and asked questions which have been answered satisfactorily.	□
2	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without being disadvantaged in any way.	□
3	I understand that people with known epileptic seizure attacks are not allowed to take part in this test. The ---------- *(institute name)* is not liable for any damage to any kind of visual aids caused by wearing a head-mounted display.	
4	I consent to the processing of my personal information for the purposes explained to me. I understand that such information will be handled in accordance with current data protection regulations.	□
5	I understand that my information may be subject to review by responsible individuals from the _____ *(institution name)* and/or regulators for monitoring and audit purposes.	□
6	I understand that confidentiality and anonymity will be maintained and the researcher will not identify me in any research output.	□
7	I agree to be contacted in the future by _____ *(institution name)* researchers who would like to invite me to participate in future studies of a similar nature	□
8	I agree that the research team may use my data for future research and understand that any use of identifiable data would be reviewed and approved by a research ethics committee. (In such cases, as with this project, data would not be identifiable in any report).	□
9	I consent to the anonymous use of the test scores obtained by the research team in scientific reports and journals.	□
10	I have the right to request to see a copy of the information _____ *(institution name)* hold about me and to request corrections or deletions of the information that is no longer required (if they do not make impossible to get the objectives of research). I can ask the _____ *(institution name)* to stop using my images at any time, in which case it will not be used in future publications, but may continue to appear in publications already in circulation.	□
11	I agree to take part in the above study.	□

Date Participant (signature)

Principal or designated researcher confirming statement

I have provided this research subject with information about the study, which I consider to be accurate and complete. The subject indicated that he or she understands fully the nature of the study, including risks and benefits, and the rights of a research subject. There has been no coercion or undue influence. I have witnessed the signature of this document by the subject.

Date Researcher (signature)

Appendix IV

Sample instructions

(This appendix does not form an integral part of this Recommendation.)

This appendix presents sample instructions to cover a two session experiment rating 360º video sequences on the ACR scale. However, an experiment could be done in one session or could require more than two sessions. Other modifications may be required.

Instructions for absolute category rating

The aim of this test is to evaluate the quality of 360º videos.

The test has a duration of approximately [*TBD depending on the test session*] min. It is divided into two sessions, where you will watch a series of 360º videos of different qualities. Each video has a duration of [*TBD depending on the test session*] s. Each session consists of [*TBD depending on the test session videos*]. We ask you to observe each video and, once it is finished, rate its overall quality using the rating interface. You should evaluate the quality of each 360º video using the following quality scale: 5. excellent; 4. very good; 3. fair; 2. poor; and 1. bad. After rating a video, the next one will automatically appear. During the test you will be seated in a swivel chair, so you can freely rotate to explore the whole 360º content.

Please, take into account that you may perceive different qualities in different parts of the video while exploring it. Please, consider the overall quality when providing your quality rating for the videos. In addition, some of the videos will present acquisition artefacts, such as stitching artefacts. Please, do not take these artefacts into account in your quality evaluation of the videos.

Before starting the formal test, you will do a preliminary perceptual test to check your vision (visual acuity, colour vision, etc.). Then, you will do a training session with some example videos to familiarize yourself with the evaluation method, the interface and to have a reference of the range of available qualities. Please, during this training session, do not hesitate to ask the experimenter to adjust the HMD (volume, camera focus, clean the screen and lenses, etc.) and any other question or doubt you may have to fully understand the test.

Before and after each session, we will ask you to complete a questionnaire about sickness and comfort. Also, we will ask you to have some minutes of rest between the two sessions.

Finally, if during the test you feel any persistent problems (headache, dizziness, etc.) do not hesitate to indicate it to the experimenter.

Thanks for participating in this test.

Instructions for degradation category rating

The aim of this test is to evaluate the quality of 360º videos.

The test has a duration of approximately [*TBD depending on the test session*] min. It is divided into two sessions, where you will watch a series of 360º videos of different qualities. Each video has a duration of [*TBD depending on the test session*] s.

Each session consists of [*TBD depending on the test session*] judgement trials. In each trial, you will be shown two versions of the same video clip in succession as follows.

– The first version is preceded by a message showing the letter A. This video clip is an example of the best quality possible for that video sequence. This example, which is called the reference sequence, is provided for information only and it is not to be rated. Observe it carefully in all of its details.

– The second version is preceded by a message showing the letter B. This video clip is called the test sequence. Your task is to rate the picture quality of this (and only this) second clip.

You are asked to evaluate the impairment of the Test sequence using the following quality scale: 5. imperceptible; 4. perceptible but not annoying; 3. slightly annoying; 2. annoying; and 1. very annoying. After rating a test sequence, the next reference sequence will automatically appear. During the test you will be seated in a swivel chair, so you can freely rotate to explore the whole 360º content.

Before starting the formal test, you will do a preliminary perceptual test to check your vision (visual acuity, colour vision, etc.). Then, you will do a training session with some example videos to have a reference of the range of available qualities and to familiarize with the evaluation method, the interface, etc. Please, during this training session, do not hesitate to ask the experimenter to adjust the HMD (volume, camera focus, clean the screen and lenses, etc.) and any other question or doubt you may have to fully understand the test.

Before and after each session, we will ask you to complete a questionnaire about sickness and comfort. Also, we will ask you to have some minutes of rest between the two sessions.

Finally, if during the test you feel any persistent problems (headache, dizziness, etc.) do not hesitate to indicate it to the experimenter.

Thanks for participating in this test.

Appendix V

Sample questionnaire for background data on subjects

(This appendix does not form an integral part of this Recommendation.)

This appendix presents a sample questionnaire to collect demographic data, background data, and feedback from observers.

To be filled in by the test subject **before** the tests:

1) Birth year:
2) Gender:
3) Profession/occupation:
4) Do you wear glasses/lenses?
5) What is your native language?
6) Experience using VR headsets (1−5, where 1: first time; 2: fewer than 5 times; 3: 5 to 20 times; 4: more than 20 times; 5: every day):
7) Previous test experience (1–5, where 1: none; 5: a lot):

To be filled in by the test subject **after** the test:

8) Do you think the experiment was easy or difficult? (1−5, where 5 is very difficult):
9) Do you think the degradations were typical of what you found in your earlier experience? (1−5, where 1: not at all; 5: very typical). You can also write a clear text comment.
10) Did you think the range of degradations was typical of what you found in your earlier experience? (1−5 where 1: not at all; 5: very typical. You can also write a clear text comment.
11) Did you use any particular part of the content in your assessments? (1: facial features; 2: movements; 3: the centre of the picture; 4: sharp edges in the image; 5: the whole picture. More than one answer can be provided, as well as a clear text comment.)
12) Do you think you were given sufficient and clear instructions before the experiment? (1−5, where 1: very unclear; 5: very clear). You can also write a clear text comment.
13) Was it difficult to concentrate on the task? (1: never; 2: at the end of the trial; 3: at the beginning of the trial; 4: periodically; 5: continuously). You can also write a clear text comment.
14) How do you think you sat during the trial? (1–5, where 1: very poor; 5: very good). You can also write a clear text comment.
15) Did you move your head a lot during the trial? (1–5, where 1: all the time; 5: never). You can also write a clear text comment.
16) Did you rotate on the swivel chair a lot during the trial? (1–5, where 1: all the time; 5: never). You can also write a clear text comment.
17) Were you disturbed by anything during the trial? (1–5, where 1: all the time; 5: never). You can also write a clear text comment.

Other comments:

Appendix VI

Analysis of exploration data

(This appendix does not form an integral part of this Recommendation.)

The proposed methods for evaluating head rotation data provide useful insights on the exploration behaviour of persons watching 360º videos.

Analysis of head rotation data

Since perception is possible during head movements, head rotation data can be modelled as trajectories from the data samples recorded by the HMD [b-David]. It is convenient to obtain data samples aligned between observers for each stimulus. If pre-processing over these samples is applied (e.g., down-sampling), it should be reported.

Analysis of eye tracking data

The gaze data provided by the eye tracker may need a pre-processing step to parse them into fixations and saccades, in particular to extract fixations, which are periods of reduced eye movements when scene perception is implied. The obtained fixations may allow the generation of saliency maps and scanpaths. For details on how to parse gaze data to fixations, see [b-David].

Report of exploration data results

A common way to represent eye and head tracking data is by means of saliency maps, which are computed by convolving each fixation or trajectory point (for all observers of one video) with a Gaussian to account for tracker precision, foveal perception and to reflect instantaneous saliency [b-David]. This convolution operation is done in a 2D sphere space (latitude and longitude coordinates), because an isotropic Gaussian on an equirectangular map would be anisotropically back projected on to a sphere. For head rotation data, other distributions can be used as well to account for eye exploration within the viewport when no eye-tracking data is available [b-Rai].

Videos containing saliency maps for each frame (or a group of them) can also be helpful [b-David].

Moreover, to obtain a general impression of which parts are explored more extensively and which less, the overall heatmap is a common approach [b-Fremerey]. In Figure VI.1, an overall heatmap is shown as an example. Here, all recorded data points of all subjects over all timestamps are shown, while the *y*-axis refers to the pitch and the *x*-axis to the yaw values. It is mostly helpful in more static sequences where no cuts or movements of persons are located. In turn, it is not as helpful in more dynamic contents, as the time component obviously gets lost using such type of evaluation. In Figure VI.1, it is apparent that the mountain, especially the waterfall, is very salient. People "scanned" the falls and the mountain with their head, where the publisher's logo also seems to be watched for quite a time, probably due to the fact that participants read it.

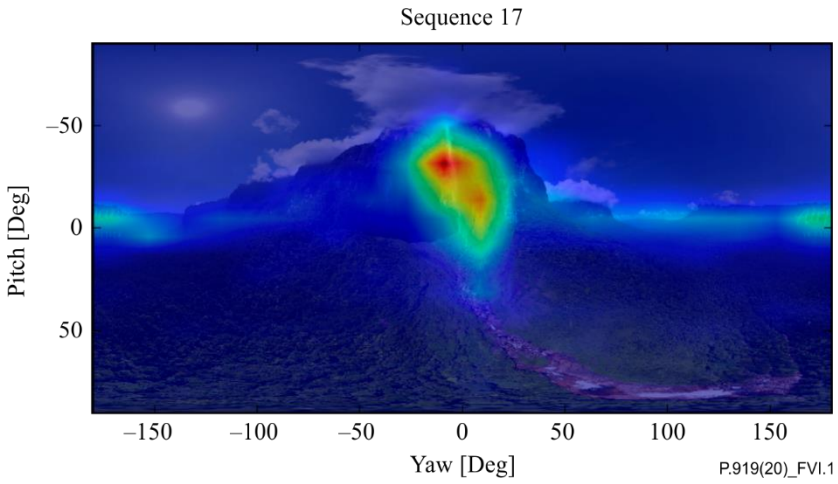

Figure VI.1 – Overall heatmap

When it is important to know which areas of a video are discovered or how long subjects spend watching these, the pitch and yaw values can be assigned to several different bins, resulting in the desired information. If the impact of several events included in the video over time is to be investigated, video heatmaps are one possibility.

For example, one method for evaluating head rotation data is shown in Figures VI.2 and VI.3 [b-Fremerey]. Here, the percentage of subjects that explored the respective distance between the minimum and maximum pitch/yaw value seen over all timestamps is displayed as categorized into several bins. The plots are generated by taking the maximum and minimum pitch or yaw value out of all recorded values, calculating the distance between the two data points and, based on that, assigning the subject to the respective area. Hence, this type of evaluation gives detailed information on how extensively subjects explored the 360° contents. Seen over all sequences, for the pitch direction, almost 90% of subjects felt most comfortable in just slightly pitching their head in a range of 100° – summed for both upper and lower areas. Only very few participants explored the video in the pitch direction beyond this area. For the yaw direction, the differences are more visible over all contents. Especially for content #20, nearly no one explored the whole video.

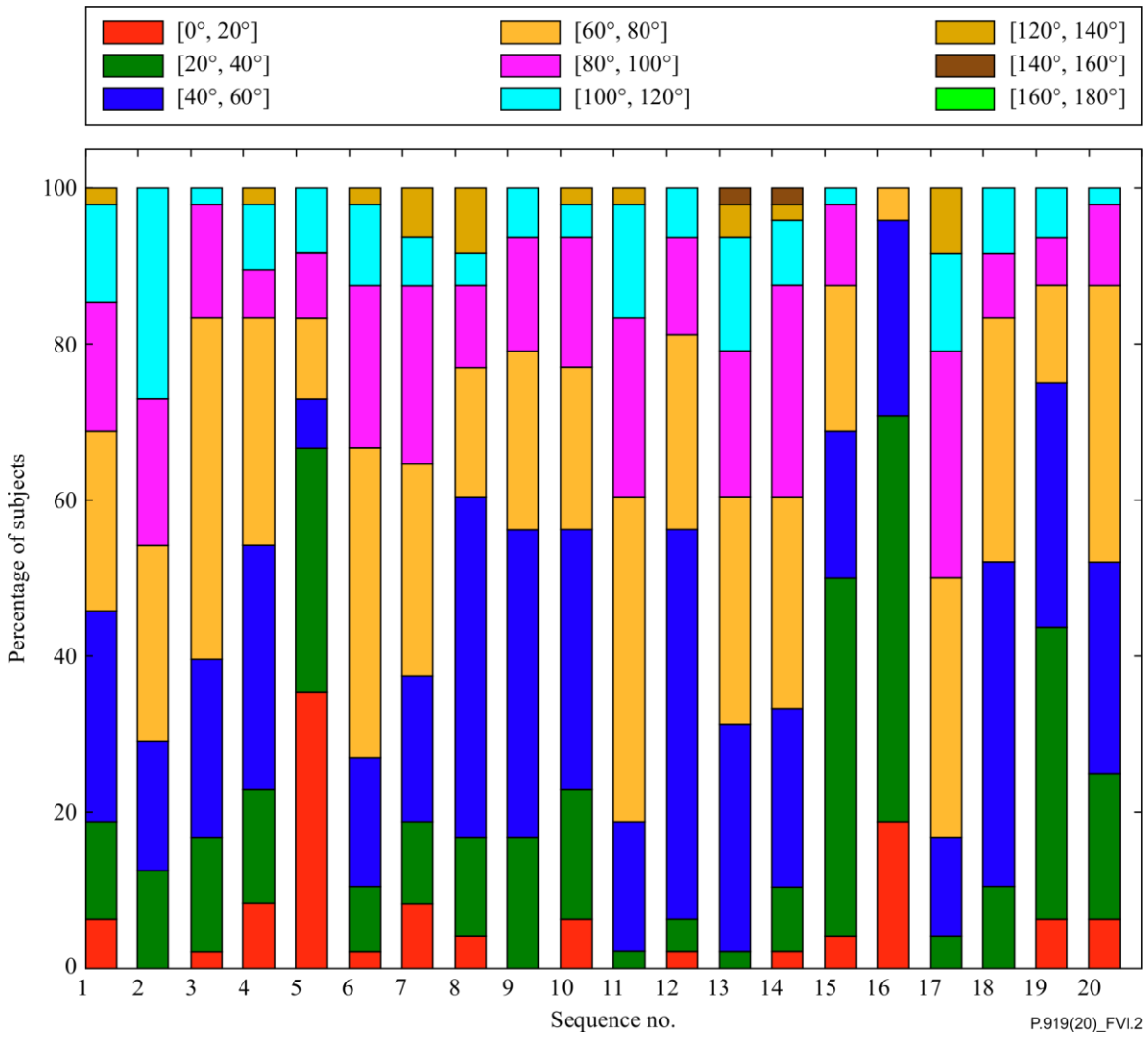

Figure VI.2 – Percentage of subjects who explored the respective pitch distance (min-max) per sequence

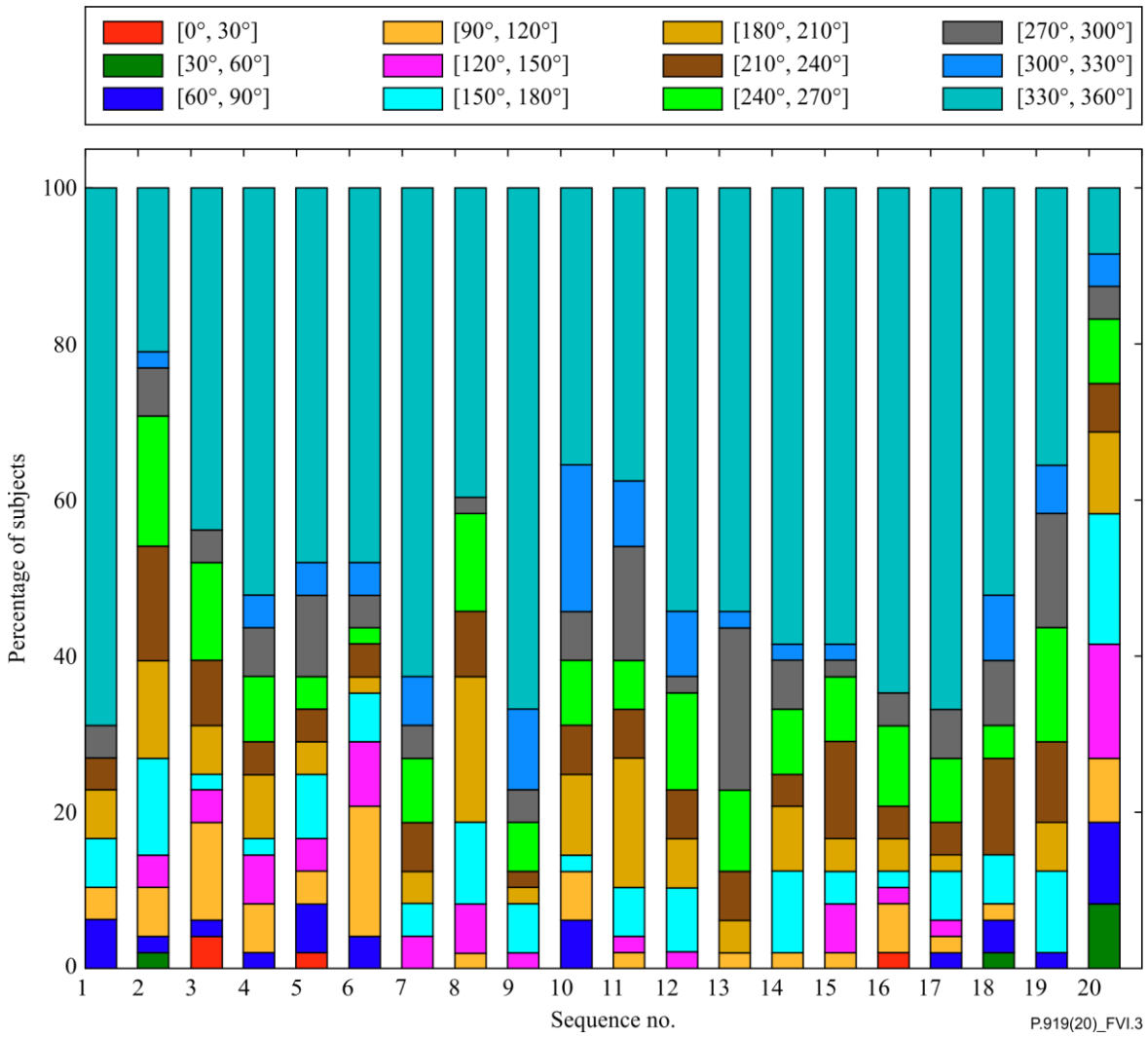

Figure VI.3 – Percentage of subjects who explored the respective yaw distance (min-max) per sequence

Another method is shown in Figures VI.4 and VI.5 [b-Fremerey], where the percentage of time watched in the respective yaw and pitch bins for each sequence summed up for all subjects is displayed. This kind of evaluation provides information on how much time users spent on specific parts of the video. It can be assumed that the areas where most of the time was spent also represent the most salient areas. Over all videos watched, the results are quite different, as the interesting contents are not always placed in the same areas. Nevertheless, the exploration behaviour can be generalized for the watched sequences. Subjects mostly did not keep watching the extreme yaw areas of the video for very long, i.e., from −150° to 150°. Almost half of the time, people kept watching the contents at or around the initial position, i.e., −30° to 30°. For pitch, roughly 90% of the time is spent on watching areas between −30° and 30°. In conclusion, if people turn their heads to discover content located at the upper or lower parts of the video, they do it for a short time and keep watching the video in a more comfortable or ordinary position afterwards.

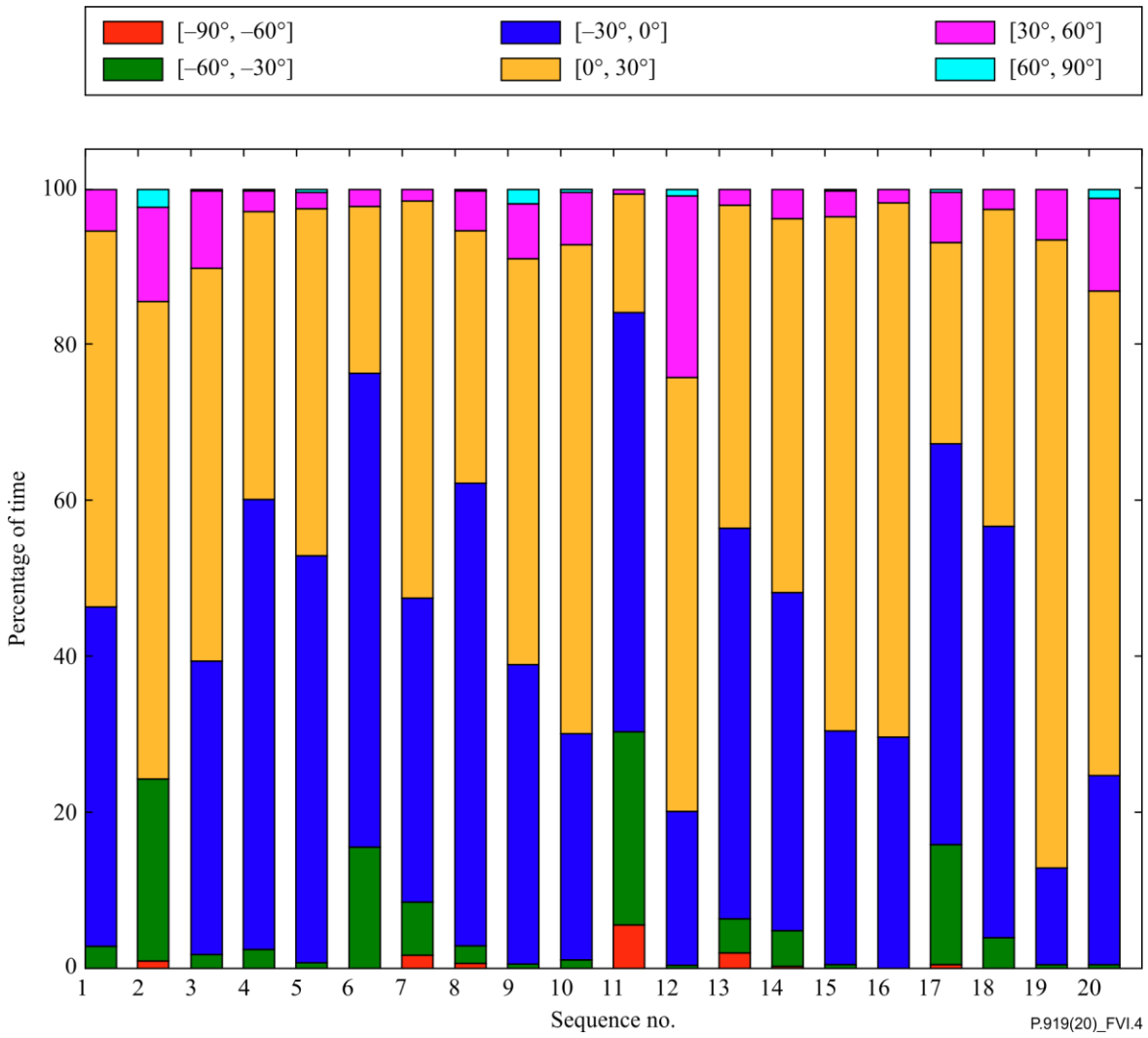

Figure VI.4 – Percentage of time watched in respective pitch bins per sequence

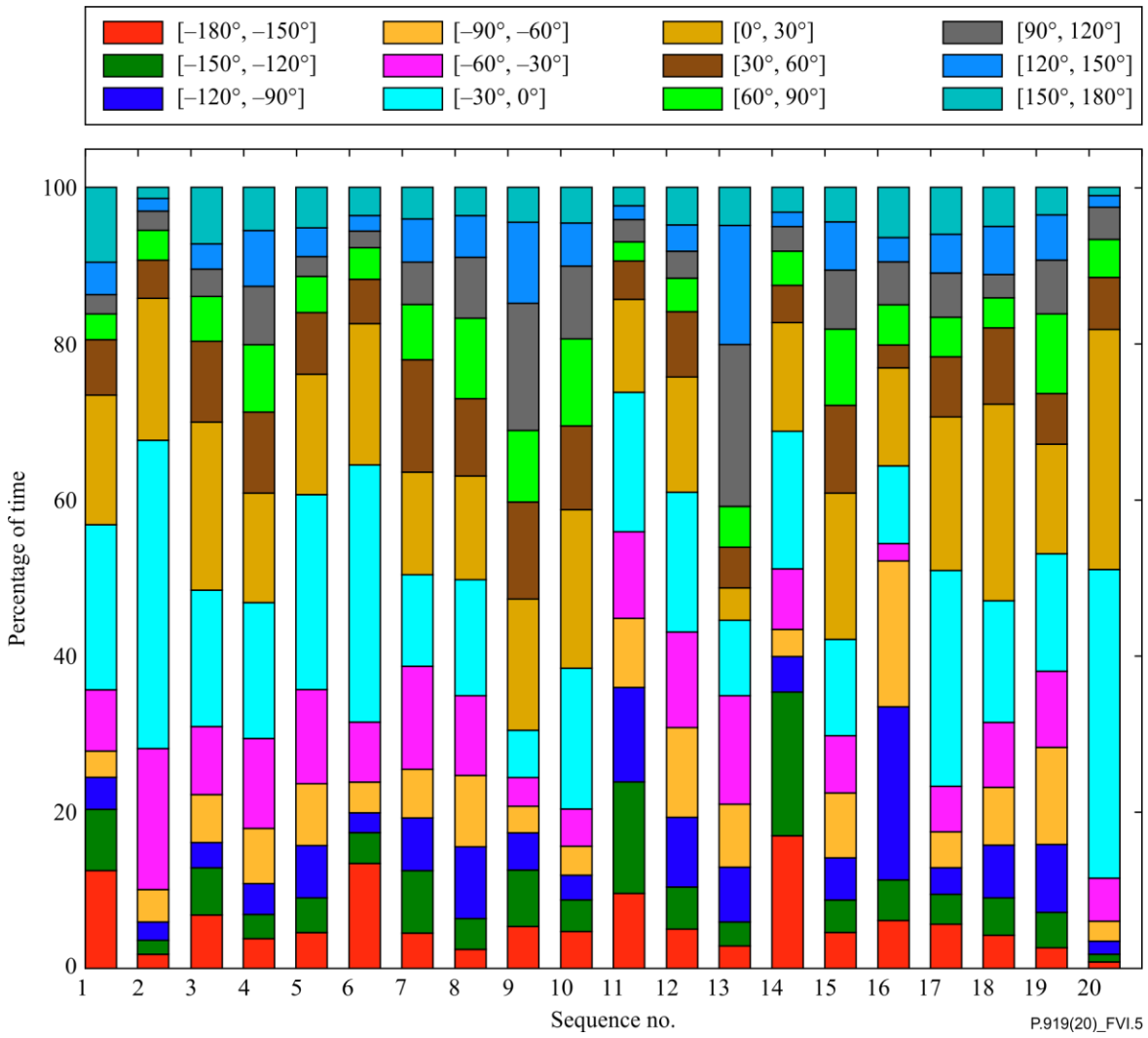

Figure VI.5 – Percentage of time watched in respective yaw bins per sequence

A similar analysis can be done for each sequence to show the distribution of fixations or head rotation samples as a function of longitude and latitude, as depicted in Figure VI.6 [b-David]. As an example, longitudinally (horizontally on the equirectangular projection) observers explore more the front part of the content (peak at 0º), and latitudinally (vertically) observers tend to explore the area around the equator (90º).

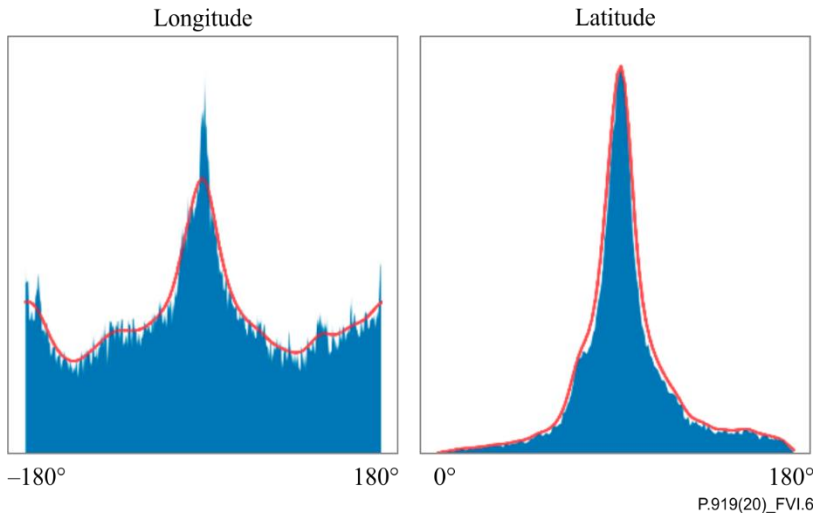

Figure VI.6 – Number of fixations or head trajectory samples

Another method for evaluating head rotation data is shown in Figure VI.7 [b-Fremerey]. Here, the angular yaw speed in degrees per second is colour coded for each subject over the duration of the video. Here, number of subjects is plotted on the *y*-axis against time on the *x*-axis. Using that method, it is possible to detect any event in the video leading to a behaviour change for some or even all subjects. The method has no loss of information included, except for colour coding the speed. One disadvantage of this type of evaluation is that sometimes it results in very noisy plots, where no specific information can be extracted. In Figure VI.7, it is apparent that starting from second 20, most people no longer move their heads.

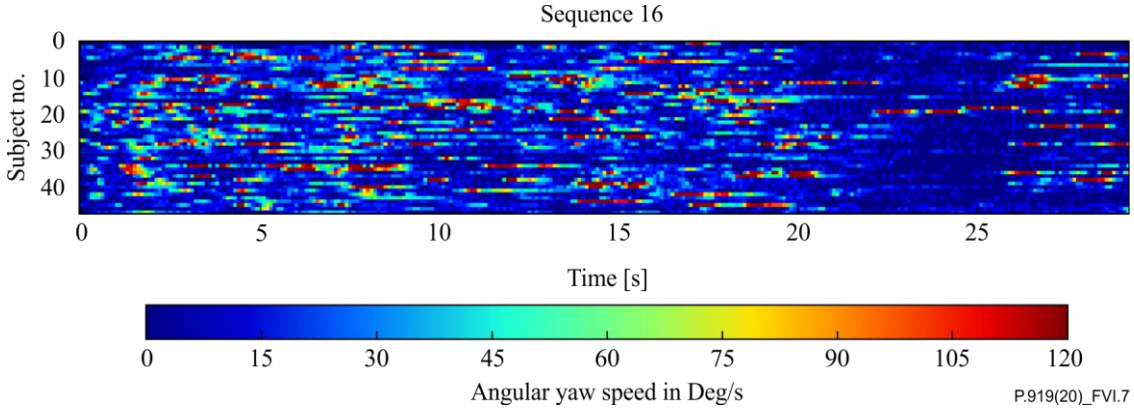

Figure VI.7 – Example of an angular yaw speed heatmap

Bibliography

[b-ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.

[b-ITU-R RS.1804] Recommendation ITU-R RS.1804 (2007), *Technical and operational characteristics of Earth exploration-satellite service (EESS) systems operating above 3 000 GHz*.

[b-ETSI TR 126 918] Technical Report ETSI TR 126 918 V16.0.0 (2020), *Universal mobile telecommunications system (UMTS); LTE; Virtual reality (VR) media services over 3GPP, 3GPP TR 26.918 version 16.0.0 Release 16*.

[b-Chen] Chen, S., Zhang, Y., Li, Y., Chen, Z., Wang, Z. (2018). Spherical structural similarity index for objective omnidirectional video quality assessment. In: Proc: *IEEE International Conference on Multimedia and Expo (ICME)*, 6 pp. New York, NY: Institute of Electrical and Electronics Engineers.

[b-David] David, E.J., Gutiérrez, J., Coutrot, A., Da Silva, M.P., Le Callet, P. (2018). A dataset of head and eye movements for 360° videos. In: *Proc. 9th ACM Multimedia Systems Conference*, pp. 432–437. New York, NY: Association for Computing Machinery.

[b-DeSimone] De Simone, F., Gutiérrez, J. Le Callet, P. (2019). Complexity measurement and characterization of 360-degree content- *Electron. Imag.*, **2019**(12), pp. 216-1-216–7.

[b-Faul] Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, **39**, pp. 175-191.

[b-Fremerey] Fremerey, S., Singla, A., Meseberg, K., Raake, A. (2018). AVtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD. In: *Proc. 9th ACM Multimedia Systems Conference (MMSys)*, pp. 403–408. New York, NY: Association for Computing Machinery.

[b-Ishihara] Ishihara (1917). *Ishihara plates*. For example, see [viewed 2020-12-11]: https://www.colour-blindness.com/colour-blindness-tests/ishihara-colour-test-plates/

[b-Kennedy] Kennedy, R.S., Lane, N.E., Berbaum, K.S.,Lilienthal, M.G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.*, **3**, pp. 203–220.

[b-Kim] Kim, H. K., Park, J., Choi, Y., Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Appl. Ergon.*, **69**, pp. 66–73.

[b-Pérez] Pérez, P., Oyaga, N., Ruiz, J.J., Villegas, A. (2018). Towards systematic analysis of cybersickness in high motion omnidirectional video. In: *Proc. 10th International Conference on Quality of Multimedia Experience (QoMEX)*, Cagliari, 2018. 3 pp. New York, NY: Institute of Electrical and Electronics Engineers.

[b-Rai] Rai, Y., Le Callet, P., Guillotel, P. (2017). Which saliency weighting for omni directional image quality assessment?,In: *Proc. 9th International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, 2017. 6 pp. New York, NY: Institute of Electrical and Electronics Engineers.

[b-Singla] Singla, A., Robitza, W., Raake, A. (2019). Comparison of subjective quality test methods for omnidirectional video quality evaluation. In: *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Kuala Lumpur, Malaysia, 2019. 6 pp. New York, NY: Institute of Electrical and Electronics Engineers.

[b-Snellen] Snellen, H. (1862). *Snellen eye chart*. For example, see [viewed 2020-12-09]: https://www.provisu.ch/images/PDF/Snellenchart_en.pdf

[b-Ye] Ye, Y., Alshina, E., Boyce, J. (2018). *JVET-E1003: Algorithm descriptions of projection format conversion and video quality metrics in 360Lib*. Geneva; Joint Video Exploration Team.

[b-Zhang] 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), pp. 461-473..

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2020