

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.910

(07/2022)

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

Audiovisual quality in multimedia services

Subjective video quality assessment methods for multimedia applications

Recommendation ITU-T P.910

ITU-T

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.910

Subjective video quality assessment methods for multimedia applications

Summary

Recommendation ITU-T P.910 describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for multimedia applications, such as videoconferencing, storage and retrieval applications, as well as telemedical applications. These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of audiovisual system performance and evaluation of the quality level during an audiovisual connection. Recommendation ITU-T P.910 also outlines the characteristics, like duration, kind of content and number of sequences, of the source sequences to be used.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.910	1996-08-30	12	11.1002/1000/3641
2.0	ITU-T P.910	1999-09-30	12	11.1002/1000/4751
3.0	ITU-T P.910	2008-04-06	9	11.1002/1000/9317
4.0	ITU-T P.910	2021-11-29	12	11.1002/1000/14828
5.0	ITU-T P.910	2022-07-29	12	11.1002/1000/15005

Keywords

Distribution quality video, Internet video, multimedia applications, subjective assessment, video quality.

* 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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2022

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

	Page
1 Scope	1
2 References.....	1
3 Definitions	2
3.1 Terms defined elsewhere	2
3.2 Terms defined in this Recommendation.....	2
4 Abbreviations and acronyms	3
5 Conventions	4
6 Source signal.....	4
6.1 Recording environment	4
6.2 Recording system	4
6.3 Scene characteristics.....	5
7 Test methods and experimental design.....	9
7.1 Absolute category rating.....	9
7.2 Absolute category rating with hidden reference.....	10
7.3 Degradation category rating	11
7.4 Pair comparison method	12
7.5 Comparison of the methods.....	13
7.6 Reference conditions	13
7.7 Experimental design	14
8 Evaluation procedures	14
8.1 Viewing conditions.....	14
8.2 Processing and playback system	15
8.3 Viewers.....	15
8.4 Instructions to viewers and training session.....	16
9 Statistical analysis and reporting of results	16
Annex A – Details related to the characterization of the test sequences	18
A.1 Sobel filter	18
A.2 Definitions of EOTF and OETF functions	18
A.3 How to use spatial information and temporal information for test sequence selection	18
A.4 Examples	19
Annex B – Additional evaluative scales	20
B.1 Rating scales.....	20
B.2 Additional rating dimensions	21
Annex C – Simultaneous presentation of sequence pairs	23
C.1 Introduction	23
C.2 Synchronization.....	23

	Page
C.3 Viewing conditions.....	23
C.4 Presentations.....	23
Annex D – Video classes and their attributes	24
Annex E – An advanced data analysis technique for tests under challenging conditions	25
Appendix I – Test sequences	27
Appendix II – Instructions for viewing tests.....	28
II.1 Absolute category rating and absolute category rating with hidden reference	28
II.2 Degradation category rating	28
II.3 Pair comparison	28
Appendix III – The simultaneous double stimulus for a continuous evaluation	30
III.1 Test procedure	30
III.2 The training phase	30
III.3 Test protocol features	30
III.4 Data processing	31
III.5 Reliability of the subjects	34
Appendix IV – Object-based evaluation.....	36
Appendix V – An additional evaluative scale for degradation category rating.....	38
Appendix VI – Reference code for Annex E	39
Bibliography.....	45

Recommendation ITU-T P.910

Subjective video quality assessment methods for multimedia applications

1 Scope

This Recommendation describes non-interactive subjective assessment methods for evaluating the quality of digital video images coded at bit rates specified in classes TV3, MM4, MM5 and MM6 for applications such as videotelephony, videoconferencing, storage and retrieval applications. The methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of video system performance and evaluation of the quality level during a video connection.

NOTE – The classes are specified in Table D.2.

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 J.61] | Recommendation ITU-T J.61 (1988), <i>Transmission performance of television circuits designed for use in international connections.</i> |
| [ITU-T P.800] | Recommendation ITU-T P.800 (1996), <i>Methods for subjective determination of transmission quality.</i> |
| [ITU-T P.930] | Recommendation ITU-T P.930 (1996), <i>Principles of a reference impairment system for video.</i> |
| [ITU-R BT.500-14] | Recommendation ITU-R BT.500-14 (2019), <i>Methodologies for the subjective assessment of the quality of television images.</i> |
| [ITU-R BT.601-7] | Recommendation ITU-R BT.601-7 (2011), <i>Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios.</i> |
| [ITU-R BT.814-4] | Recommendation ITU-R BT.814-4 (2018), <i>Specifications of PLUGE test signals and alignment procedures for setting of brightness and contrast of displays.</i> |
| [ITU-R BT.1886] | Recommendation ITU-R BT.1886 (2011), <i>Reference electro-optical transfer function for flat panel displays used in HDTV studio production.</i> |
| [ITU-R BT.2100] | Recommendation ITU-R BT.2100-2 (2018), <i>Image parameter values for high dynamic range television for use in production and international programme exchange.</i> |
| [IEC TR 60268-13] | IEC TR 60268-13:1998, <i>Sound system equipment – Part 13: Listening tests on loudspeakers.</i> |

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 explicit reference; source reference: The condition used by the assessors as reference to express their opinion, when the degradation category rating method is used. This reference is displayed first within each pair of sequences. Usually, the format of the explicit reference is the format used at the input of the codecs under test (e.g., [ITU-R BT.601-7], common intermediate format, quarter common intermediate format or standard intermediate format).

NOTE – In the body of this Recommendation, the words "explicit" and "source" are omitted whenever the context makes clear the meaning of "reference".

3.2.2 gamma: A parameter that quantifies the discrimination between the grey level steps on a visual display. The relation between the screen luminance and the input signal voltage is non-linear, with the voltage raised to an exponent gamma. To compensate for this non-linearity, a correction factor that is an inverse function of gamma is generally applied in the camera. Gamma also has an impact on colour rendition.

3.2.3 implicit reference: The condition used by the assessors as reference to express their opinion on the test material, when the absolute category rating method is used. If the implicit reference is suggested by the experimenter, it must be well known to all the assessors (e.g., conventional television (TV) systems or reality).

3.2.4 optimization test: Subjective test that is typically carried out during either the development or the standardization of a new algorithm or system. The goal of such a test is to evaluate the performance of new tools in order to optimize the algorithms or the systems that are under study.

3.2.5 qualification test: Subjective test that is typically carried out in order to compare the performance of commercial systems or equipment. These tests must be carried out under test conditions that are as representative as possible of the real conditions of use.

3.2.6 reference conditions: Dummy conditions added to test conditions in order to anchor the evaluations coming from different experiments.

3.2.7 reliability of a subjective test:

- a) intra-individual ("within subject") reliability – the agreement between a certain subject's repeated ratings of the same test condition;
- b) inter-individual ("between subjects") reliability – the agreement between different subjects' ratings of the same test condition.

3.2.8 replication: Repetition of the same circuit condition (with the same source material) for the same subject.

3.2.9 spatial information (SI); spatial perceptual information: A measure that generally indicates the amount of spatial detail in a picture. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor is it associated with the information defined in communication theory.

NOTE – See clause 6.3.2 for the equation for SI.

3.2.10 temporal information (TI); temporal perceptual information: A measure that generally indicates the amount of temporal changes of a video sequence. It is usually higher for high motion

sequences. It is not meant to be a measure of entropy nor associated with the information defined in communication theory.

NOTE – See clause 6.3.3 for the equation for TI.

3.2.11 transparency; fidelity: A concept describing the performance of a codec or a system in relation to an ideal transmission system without any degradation.

Two types of transparency can be distinguished as follows:

The first type describes how well the processed signal conforms to the input signal or ideal signal using a mathematical criterion. If there is no difference, the system is fully transparent. The second type describes how well the processed signal conforms to the input signal or ideal signal for a human observer. If no difference can be perceived under any experimental condition, the system is perceptually transparent. The term "transparent" without explicit reference to a criterion will be used for systems that are perceptually transparent.

3.2.12 validity of a subjective test: Agreement between the mean value of ratings obtained in a test and the true value which the test purports to measure.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

2D	two-Dimensional
%GOB	percentage of Good Or Better
%POW	percentage of Poor Or Worse
ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
CI	Confidence Interval
CIF	Common Intermediate Format
CRT	Cathode Ray Tube
DCR	Degradation Category Rating
DMOS	Differential Mean Opinion Score
DV	Differential Viewer
EOTF	Electro-Optical Transfer Function
HDR	High Dynamic Range
HLG	Hybrid Log Gamma
MOS	Mean Opinion Score
OBE	Object-Based Evaluation
OETF	Opto-Electrical Transfer Function
PC	Pair Comparison
PQ	Perceptual Quantizer
PVS	Processed Video Sequence
QCIF	Quarter Common Intermediate Format
SD	Standard Deviation
SDR	Standard Dynamic Range

SDSCE	Simultaneous Double Stimulus for a Continuous Evaluation
S/N	Signal-to-Noise
SI	Spatial Information
SIF	Standard Intermediate Format
SP	Simultaneous Presentation
SSCQE	Single Stimulus Continuous Quality Evaluation
TC	Test Condition
TI	Temporal Information
TP	Test Presentation
TV	Television
VO	Virtual Object
VS	Video Segment
VTR	Video Tape Recorder

5 Conventions

None.

6 Source signal

In order to control the characteristics of the source signal, the test sequences should be established 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 sequences to eliminate a further source of variation.

6.1 Recording environment

Lighting source(s) (bulbs or fluorescent lamps) can be placed above or on the side of the camera. When placing the lights, recognize that overhead is more typical of office lighting, and should be used with scenes that portray the business environment. Studio lights and other non-typical sources should be avoided.

The lighting conditions of the room in the field of view could vary from 100 lx to ~10 000 lx for indoor use. The variation (alternating current frequency) of the light (fluorescent lighting) must be taken into account because this may cause a flicker in the recorded video sequence.

Lighting conditions, wall colours, surface reflectance, etc., should be carefully controlled and reported.

6.2 Recording system

6.2.1 Camera

Picture sequences should be recorded by a high-quality charge-coupled device camera.

The signal-to-noise (S/N) ratio of the input video signal can strongly affect the performance of the codec.

The following points should be used to specify the video input:

- the dynamic range of the *YUV* signals;
- the gamma correction factor (should be 0.45);

- the bandwidth or slopes of the filters;
- the sensitivity of the camera at very low lighting conditions and the characteristics of an automatic gain control, if used.

The weighted S/N ratio should be measured according to clause C.3.2.1 of [ITU-T J.61]. The weighted S/N ratio should be greater than 45 dB r.m.s.

The instability or the jitters of the clock signals could cause noise effects. A minimum stability of 0.5 ppm is required for the camera clocking device.

Either fixed or variable focal length systems can be used. For desktop terminals, a focal depth from 30 cm to 120 cm is reasonable, while for multi-user systems a focal depth from 50 cm to infinity might be more appropriate. To support the variation of illuminance in the recording room, either an adjustable iris or neutral density filters should be used. The camera should have an automatic white balance so that adaptation to the colour temperature of the light source can be accomplished. The correction of white temperature can range from 2 700 K (indoor use with electrical bulb) to 6 500 K (daylight temperature with clouded sky).

6.2.2 Video signal and storage format

Video source signals provided by the camera should be sampled in conformity with [ITU-R BT.601-7]. In order to avoid distortion of the source signal, it should be stored in digital format, e.g., on computer or D1 4:2:2 tape format.

6.3 Scene characteristics

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

Details of the characterization of the test sequences and examples of suitable test scenes are given in Annex A and in Appendices I and II.

The number of sequences should be established according to the experimental design. In order to avoid boring observers and to achieve a minimum reliability of the results, at least four different types of scene (i.e., different subject matter) should be chosen for the sequences.

Clause 6.3.1 shows how to pre- and post-process video frames for SI/TI calculation. Clauses 6.3.2 and 6.3.3 present methods for quantifying the SI and TI of test scenes. These methods for evaluating the SI and TI of test scenes are applicable to video quality testing both now and in the future. The location of the video scene within the spatiotemporal matrix is important because the quality of a transmitted video scene (especially after passing through a low bit-rate codec) is often highly dependent on this location. The SI and TI measures presented here can be used to ensure appropriate coverage of the spatiotemporal plane.

The SI and TI measures given in clauses 6.3.2 and 6.3.3 are single valued for each frame over a complete test sequence. This results in a time series of values that will generally vary to some degree. The variability itself may be usefully studied, e.g., with plots of spatiotemporal information on a frame-by-frame basis. The use of information distributions over a test sequence also permits better assessment of scenes with scene cuts. To aggregate SI and TI measures into one number each per sequence, aggregation measures given in clause 6.3.4 may be used.

6.3.1 Pre-processing of luma values and post-processing steps for SI/TI

For SI/TI calculation, the luma values for every frame shall be pre-processed according to the method described in this clause. This pre-processing ensures that the SI/TI calculation applies to contents with different colour range (limited vs. full), bit depth (e.g., 8 or 10 bits per plane) and optional high dynamic range (HDR) characteristics.

The pre- and post-processing pipeline is shown in Figure 1.

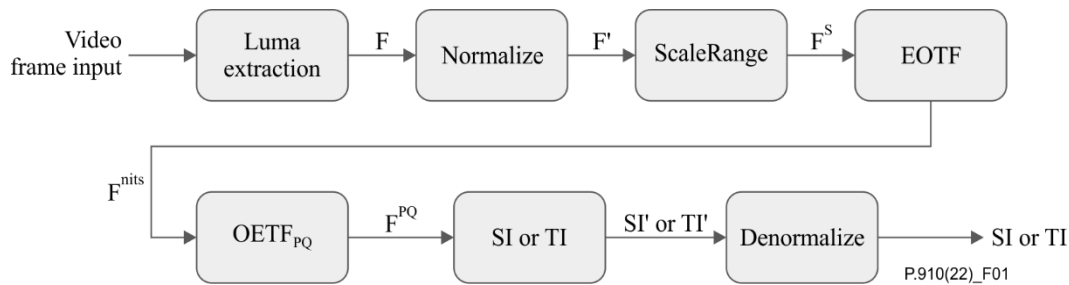


Figure 1 – SI/TI pre- and post-processing pipeline

The luma values for each video frame at index $n \in \{1, 2, \dots, N\}$ are a matrix F_n with its dimensions $I \times J$ corresponding to the resolution of the video frame.

A Python reference software implementing the processing pipeline and the following SI/TI calculations can be found at [b-siti-python].

6.3.1.1 Normalization

F_n shall be normalized to a range $[0, 1]$, resulting in F'_n . This normalization accounts for differences in original luma values for different bit depths (e.g., $[0, 255]$ for 8 bits per plane; $[0, 1023]$ for 10 bits per plane, etc.).

The normalization uses an inverse of the bit depth $b \in \{8, 10, 12, \dots\}$, yielding the function *Normalize*:

$$\text{Normalize}(x, b) = \frac{x}{2^b - 1}$$

$$F'_n = \text{Normalize}(F_n, b)$$

6.3.1.2 Range scaling

The function *ScaleRange* shall be applied to scale the normalized luma values F'_n to the full range for sequences with limited colour range (e.g., $[16, 235]$ for 8 bits per plane). Sequences with full colour range are not modified:

$$\text{ScaleRange}(x, \text{range}) = \begin{cases} \frac{x - \frac{16}{255}}{\frac{235}{255} - \frac{16}{255}} & \text{range} = \text{limited} \\ x & \text{otherwise} \end{cases}$$

$$F_n^S = \text{ScaleRange}(F'_n, \text{range})$$

6.3.1.3 Electro-optical transfer function

An electro-optical transfer function (EOTF) shall be applied to F_n^S in order to convert the luma values into luminance values in the physical domain F_n^{nlts} .

$$F_n^{\text{nlts}} = \text{EOTF}(F_n^S, \text{domain})$$

The choice and parameterization of the EOTF function depends on whether the sequence uses standard dynamic range (SDR) or has high dynamic range (HDR) characteristics. In the latter case, the luma values may be encoded in the hybrid-log-gamma (HLG) or perceptual quantizer (PQ) domain.

The EOTF is therefore chosen as follows:

$$EOTF(x, domain) = \begin{cases} EOTF_HLG, & domain = HLG \\ EOTF_SDR, & domain = SDR \\ x, & domain = PQ \end{cases}$$

For details on the EOTF functions EOTF_HLG and EOTF_SDR, see Annex A.

6.3.1.4 Opto-electronic transfer function

The values in the physical domain F_n^{nits} shall be converted into the PQ domain using the OETF_PQ function, resulting in F_n^{PQ} .

$$F_n^{PQ} = OETF_{PQ}(F_n^{nits})$$

For details on the OETF_PQ function, see Annex A.

F_n^{PQ} shall then be used to calculate SI and TI according to clauses 6.3.2 and 6.3.3, respectively.

6.3.1.5 Post-processing

The SI and TI values calculated according to clauses 6.3.2 and 6.3.3 shall be post-processed by applying the inverse *Normalize* function from clause 6.3.1.1, *Denormalize*:

$$Denormalize(x, b) = x * (2^b - 1)$$

This ensures that sequences with different bit depths can be compared on the same scale.

6.3.2 Spatial perceptual information measurement

The SI is based on the Sobel filter (see Annex A). The luminance values in each pre-processed video frame luminance plane F_n^{PQ} at time n shall be filtered with the Sobel filter, $Sobel(F_n^{PQ})$. The standard deviation (SD) over the pixels (σ_{space}) in each Sobel-filtered frame shall then be computed.

This operation shall be repeated for each frame in the video sequence and results in a time series of SI'_n of the scene.

This process for each frame can be represented in equation form as:

$$SI'_n = \sigma_{space}[Sobel(F_n^{PQ})]$$

To obtain the final values SI_n , the function *Denormalize* from clause 6.3.1.5 shall be applied to the SI'_n values.

$$SI_n = Denormalize(SI'_n, b)$$

6.3.3 Temporal information measurement

TI is based upon the motion difference feature, $M_n(i, j)$, that is the difference between the pixel values (of the luminance plane) at the same location in space but at successive times or frames. $M_n(i, j)$ as a function of time (n) is defined as:

$$M_n(i, j) = F_n^{PQ}(i, j) - F_{n-1}^{PQ}(i, j)$$

Here $F_n^{PQ}(i, j)$ is the pixel at the i th row and j th column of the n th frame in time.

The TI'_n measure is computed as the SD over space (σ_{space}) of $M_n(i, j)$ over all i and j , with $n \in \{2, 3, \dots, N\}$.

$$TI'_n = \sigma_{space}[M_n(i, j)]$$

Note that for the first frame, $n = 1$, no TI value exists per definition.

To obtain the final values TI_n , the function *Denormalize* from clause 6.3.1.5 shall be applied to the TI'_n values.

6.3.4 Aggregation of SI/TI scores

Multiple SI and TI values per sequence may be aggregated into single numbers for SI and TI, respectively, by applying appropriate statistical measures such as the minimum, maximum, median, average, or percentiles.

It is recommended to use the average as an aggregation measure:

$$SI = SI_{aggregated} = \frac{1}{N} \sum_1^N SI_n$$

$$TI = TI_{aggregated} = \frac{1}{N-1} \sum_2^N TI_n$$

Here, N is the number of frames and hence SI_n scores for the video sequence. Note that since the calculation for TI_n requires two frames each, it starts at $n = 2$, with a total of $N - 1$ values TI_n used in the aggregation.

NOTE – In the 2021 version of this Recommendation, the respective maximum value was recommended as aggregated score for SI and TI. In publications and other practical deployment, however, this was not consistently used. Thus, it is recommended to use the average. If the resulting SI and TI values are being compared to those provided in publications or with publicly available databases, deviations may stem from the previously recommended usage of the maximum.

Further statistical indicators such as the variance or standard deviation over SI_n or TI_n may be useful for determining the variation in complexity of a sequence.

Results from [b-Robitz] have shown that the average and median SI provide better correlation with compressibility compared to the minimum or maximum SI. For TI, the minimum and average provide better correlation than the maximum TI.

6.3.5 Usage of SI/TI with scene cuts

For SI and TI, aggregating individual values across scene cut boundaries may lead to different scene characteristics being averaged, possibly hiding the underlying information. For TI in particular, computation of the motion difference feature may yield large values at the scene boundaries, which, when aggregated, may distort aggregate results such as mean values.

For sequences that contain scene cuts, SI and TI should therefore not be computed across scene boundaries, instead treating the individual scenes from one sequence separately.

If separating a sequence is not possible, it may be necessary to compute SI and TI across scene boundaries. This may be the case for instance in a live feed, or when individual scenes would be too short to be useful for inclusion in subjective tests.

In case scene segmentation is not feasible through pre-processing, scene changes may be detected through occurrences of instantaneous, large TI values. These occurrences may be filtered out by means of comparing the difference between each TI_n value against a moving average of previous TI values. If that difference exceeds a certain threshold, a scene cut at frame n is likely and that TI value should be excluded. The threshold should be determined empirically. More information on this

method can be found in [b-Trioux]. This method may not work reliably with scene transitions involving crossfades.

6.3.6 Usage of SI/TI with different frame resolutions

SI and TI values must not be compared across sequences using different frame resolutions, because the value range of the Sobel filter – as well as its standard deviation – depend on the visual appearance of edges in the original sequence. Hence, sequences with higher resolution have intrinsically lower SI and vice-versa, which will also impact the TI scores.

To compensate for differences in aspect ratios, SI and TI values may be compared across sequences when the frame resolutions of those sequences do not differ by more than 10% in each dimension.

6.3.7 Usage of TI with different frame rates

TI values must not be compared across sequences using different frame rates since, for the same content, the motion difference between frames at shorter intervals (i.e., higher frame rates) is lower, resulting in a different TI value.

To analyse sequences with different frame rates, the sequences may be grouped into sets of equal frame rates (e.g., one set with 25 fps, one set with 60 fps) and analysed separately.

7 Test methods and experimental design

Measurement of the perceived quality of images 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 video sequence presented to the subjects in a test, and the magnitude and nature of the sensation caused by the stimulus.

A number of experimental methods have been validated for different purposes. Here three methods are recommended for applications using connections at bit rates specified in video classes TV3, MM4, MM5 and MM6, as specified in Table D.2. Further test methods are described in Appendices III and IV.

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.

7.1 Absolute category rating

The ACR method is a category judgement where the test sequences are presented one at a time and are rated independently on a category scale. (This method is also called the single stimulus method.)

The method specifies that after each presentation, the subjects are asked to evaluate the quality of the sequence shown.

The time pattern for the stimulus presentation can be illustrated by Figure 2. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤ 10 s. The presentation time may be reduced or increased according to the content of the test material.

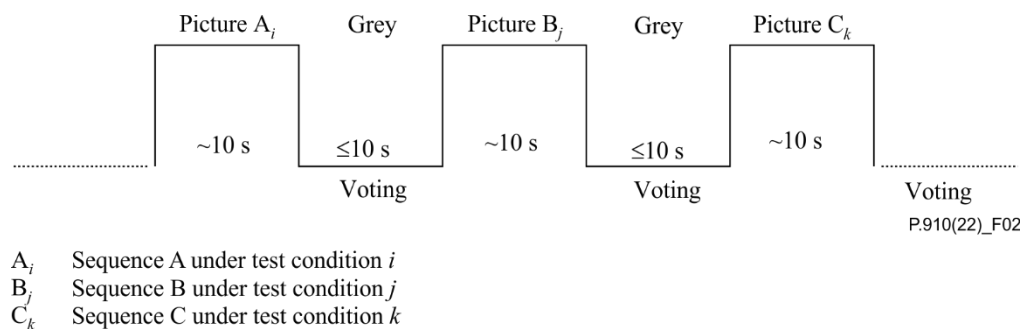


Figure 2 – Stimulus presentation in the absolute category rating method

The following five-level scale for rating overall quality should be used:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

If higher discriminative power is required, a nine-level scale may be used. Examples of suitable numerical or continuous scales are given in Annex B. Annex B also gives examples of rating dimensions other than overall quality. Such dimensions may be useful for obtaining more information on different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different.

For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

7.2 Absolute category rating with hidden reference

The absolute category rating with hidden reference (ACR-HR) method is a category judgement where the test sequences are presented one at a time and are rated independently on a category scale. The present test procedure must include a reference version of each test sequence shown as any other test stimulus. This is termed a hidden reference condition. During the data analysis, a differential mean opinion score (DMOS) of quality will be computed between each test sequence and its corresponding (hidden) reference. This procedure is known as "hidden reference".

The method specifies that, after each presentation, the subjects are asked to evaluate the quality of the sequence shown.

The time pattern for the stimulus presentation is illustrated in Figure 2. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤ 10 s. The presentation time may be reduced or increased according to the content of the test material.

The following five-level scale for rating overall quality should be used:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

Differential viewer (DV) scores are calculated on a per subject per processed video sequence (PVS) basis. The appropriate hidden reference (REF) is used to calculate DV(PVS) using the following formula:

$$DV(PVS) = V(PVS) - V(REF) + 5$$

where V is the viewer's ACR score. In using this formula, a DV(PVS) of 5 indicates excellent quality and a DV(PVS) of 1 indicates bad quality. Any DV values greater than 5 (i.e., where the processed sequence is rated better quality than its associated hidden reference sequence) are generally considered valid. Alternatively, a two-point crushing function may be applied to prevent these individual ACR-HR DV scores from unduly influencing the overall mean opinion score (MOS):

$$\text{crushed_DV} = (7 * DV) / (2 + DV) \quad \text{when } DV > 5$$

If higher discriminative power is required, a nine-level ACR scale may be used. Examples of suitable numerical or continuous scales are given in Annex B. Annex B also gives examples of rating dimensions other than overall quality. Such dimensions may be useful for obtaining more information about different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test, although the systems are clearly perceived as different.

For the ACR-HR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

The ACR-HR method should only be used with reference video that an expert in the field considers to be of good or excellent quality on the five-level scale specified above.

The ACR-HR method may not be suitable for analysing unusual impairments that occur in the first and last second of the video sequence. The viewer's unfamiliarity with the reference video sequence may cause an otherwise obvious impairment to be missed (e.g., if a sequence pauses immediately prior to the end, a viewer may not be able to determine whether this is intended content or a network error).

7.3 Degradation category rating

The DCR implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. (This method is also called the double stimulus impairment scale method.)

When reduced picture formats are used (e.g., common intermediate format (CIF), quarter common intermediate format (QCIF) or standard intermediate format (SIF)), it could be useful to display the reference and the test sequence simultaneously on the same monitor. Guidelines on this presentation procedure are discussed in Annex C.

NOTE – CIF is a picture format specified in [b-ITU-T H.261] for video phone: 352 lines × 288 pixels. QCIF is a picture format specified in [b-ITU-T H.261] for video phone: 176 lines × 144 pixels. SIF is a picture format specified in [b-ISO/IEC 11172] (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s.

The time pattern for the stimulus presentation is illustrated in Figure 3. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be ≤10 s. The presentation time may be reduced or increased according to the content of the test material.