|  |  |
| --- | --- |
| INTERNATIONAL TELECOMMUNICATION UNION | **STUDY GROUP 9** |
| **TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2009-2012 | **TD 119 (GEN/9)** |
| **English only****Original: English** |
| **Question(s):** | 12/9 | Geneva 14-18, January 2013 |
| **TEMPORARY DOCUMENT** |
| **Source:** | Editors P.3D-sam |
| **Title:** | Output - Draft New Recommendation P.3D-sam: *“Subjective assessment methods for 3D video quality”*  |

***TSB Note: This document represents the output from Q12/9 session during SG 9 meeting, 14-18, January 2013.***

----------------

<Recommendation No.>

Draft P.3D-sam: Subjective Assessment Methods for 3D Video Quality

AAP Summary

<to be added before Consent>

Summary

This Recommendation describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for three dimensional (3D) video applications such as 3D videoconferencing, and 3D cable television. These methods can be used for several different purposes including, but not limited to, selection of algorithms, ranking of system performance and evaluation of the quality level during a video connection. This Recommendation also outlines the characteristics of the source sequences to be used, like duration, kind of content, number of sequences, etc.

Details within this document are expected to change, based on experiments into how best to conduct 3DTV subjective tests.

Keywords

<Optional>

Introduction

Stereoscopic three-dimensional (3D) television attempts to emulate the response of the human binocular visual system to the relative depth perception of objects. This Recommendation applies to stereoscopic imaging that directs a different view of the same scene to each eye. The images of the objects depicted in the scene will have different relative positions in the left- and right-view. 3D television viewing does not perfectly recreate the real viewing experience, because the normal formulae for accommodation (i.e., focus) and vergence (i.e., eye angle) do not apply. A variety of displays produce this effect, including stereoscopic and autostereoscopic displays, using glasses with polarized lenses or shutters; and 2D televisions using complementary color anaglyphs and glasses with colored filters.

Assessment factors generally applied to monoscopic (two-dimensional, or 2D) television pictures can be applied to stereoscopic television systems. In addition, there are many factors unique to stereoscopic television systems. These include factors such as depth resolution, which is the spatial resolution in depth direction, depth motion (that is, whether motion or movements along depth direction), and visual comfort.

# 1 Scope

##

This Recommendation describes subjective assessment methods for 3D video quality in terms of light condition, viewing distance, comfort level, maximum crosstalk, display size, etc.

## 1.1 Applications

The applications for the subjective assessment methods for 3D video quality described in this Recommendation include, but are not limited to:

1. Obtaining perceptual 3D video quality

## 1.2 Limitations

# 2 References

Normative References

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

[ITU-T P.911] ITU-T Recommendation P.911 (1998), Subjective audiovisual quality assessment methods for multimedia applications.

[ITU-R BT.1788] ITU-R Recommendation BT.1788 (2007), Methodology for the subjective assessment of video quality in multimedia applications .

[ITU-R BT.500] ITU-R Recommendation BT.500-13 (01/2012), Methodology for the subjective assessment of the quality of television pictures.

[ITU-R BT.2021] ITU-R Recommendation BT.2021 (08/2012), Subjective methods for the assessment of stereoscopic 3DTV systems.

Informative References

# 3 Definitions

This recommendation uses the following terms defined elsewhere.

## 3.1 Terms defined elsewhere:

This Recommendation uses the following terms defined elsewhere:

**3.1.1**

**3.1.2**

## 3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

**3.2.1 Depth quality** refers to the ability of the system to deliver a sensation of depth. The presence of monocular cues (such as linear perspective, blur, gradients) convey some sensation of depth even in standard 2D images. Stereoscopic 3D images contain also disparity information which provides additional depth information and thus an enhanced sense of depth as compared to 2D.

**3.2.2 Double Stimulus:** a quality rating method where the subject is presented with two stimuli and then rates both stimuli in the context of the joint presentation (e.g., a rating that compares the quality of one sequence to the quality of the other sequence).

**3.2.3** **Picture Quality** is related to the quality of rendering of texture and motion, the level of visibility of visual artifacts and rendering details. It refers toPicture quality is mainly affected by technical parameters and errors introduced by, for example, encoding and/or transmission processes and/or 3D video formats conversions e.g. 2D plus depth to stereoscopic images. [editor’s note: definition and term will be discussed further and perhaps modified.]

**3.2.4** **Naturalness** refers to the perception of the stereoscopic image as being a truthful representation of reality (i.e. perceptual realism). The stereoscopic image may present different types of distortion which make it less natural. For example, stereoscopic objects are sometimes perceived as unnaturally large or small (puppet theatre effect), or they appear unnaturally thin (cardboard effect).

**3.2.5 Reference**: the source sequence. This is the highest quality version available of the audio, video, or audiovisual sequence.

**3.2.6** **Sense of presence** refers to “the subjective experience of being in one place or environment even when one is situated in another.”

**3.2.7 Single Stimulus**: a quality rating method where the subject is presented with one stimulus and then rates each stimulus in isolation (e.g., a viewer watches one video sequence, and then rates it).

**3.2.8 Visual Comfort** can be related to multi-symptoms, e.g. eye strain, dry eyes, double vision. Moreover, variation of Visual Comfort can be perceived as the sensation of visual impairment as well as the sense of vision difficulties when moving the fixation point from one area of the image to another area (due to the decoupling of accommodation and convergence). [editor’s note: definition and term will be discussed further and perhaps modified.]

**3.2.10 Visual discomfort** refers to the subjective sensation of discomfort that can be associated with the viewing of stereoscopic images. This issue is usually referred to as visual comfort. [editor’s note: this definition may need to be changed or deleted. Compare with ‘visual comfort’.]

**3.2.11** **Visual Experience** is defined as the overall quality of experience of the images in terms of immersion, perceived image quality as well as depth rendering considering the shape and the dimension of objects. [editor’s note: definition and term will be discussed further and perhaps modified.]

# 4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

# 5 Conventions

*None.*

# 6 Selection of 3D Source Content

In order to evaluate 3D visual experience and other terms defined in this Recommendation in various circumstances, the content should cover a wide range of 3D videos. In particular, 3D content with a variety of texture, depth and motion should be used for accurate assessment.

The 3D test sequences 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 sequences to eliminate a further source of variation.

The selection of the test material should be motivated by the experimental question addressed in the study. For example, the content of the test sequences (sport, drama, film, etc.) and their spatiotemporal and depth characteristics should be representative of the programmes delivered by the service under study.

## 6.1 Visual comfort

Excessive disparity/parallax causes visual discomfort possibly because it worsens the conflict between accommodation and vergence. Therefore, it has been suggested that to minimize the accommodation-vergence conflict, the disparities in the stereoscopic image should be small enough so that the perceived depths of objects fall within a “comfort zone”. [Editor’s Note: Further studies required. In particular, how to compute DoF. Subjective tests show the “comfort zone” is delimited by the Depth of Field (DoF) of the human eye. For typical 3DTV viewing conditions, DoF=±0.2 diopters is used as a common threshold.]

The selected stereoscopic test sequences content should be normally comfortable to watch. The visual comfort of stereoscopic images depends critically upon the disparity contained in the image. The visual comfort of stereoscopic images depends also upon the subject’s inter-pupilary distance and the viewing distance.

[Editor’s Note: Further studies required. In particular, how to compute DoF. To reduce bias due to the presence or feeling of visual discomfort on 3D source test material, it would be interesting to calculate the depth range of each scene in diopter unit to know whether the scene is in accordance with the comfort zone requirement. The depth range around the screen plane can be calculated in diopter unit knowing the maximum crossed and uncrossed horizontal disparities and the viewing conditions.]

## 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. Fair and 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 was 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.

**6.3 Optional Subjective Methods for 3D Reference Scene Selection:** **Visual Experience and Visual Comfort requirements**

To conduct a subjective test on 3D video, it is desirable to select a set of original scene contents (reference video) with a maximum visual comfort. Preferably, 3D original scene contents should have visual comfort similar to the 2D version of that content, for short viewing durations. In fact, the main goal of 3D video subjective tests is to evaluate the impact of 3D video technologies or image processing algorithms on viewers’ opinion in terms of visual experience, image quality and visual comfort. In case of visual discomfort issues with reference scene contents, visual discomfort may interact with the other quality scales and attenuate related results, resulting on much more difficulties to evaluate the interest of 3D technologies to guarantee an optimal visual experience. Therefore, to ensure a fair comparison of technologies as well as reliable and stable results, reference scene contents should be selected by using the following procedures:

Perform a subjective video quality experiment on the reference scenes (only). Two sessions will be performed: (1) rating visual comfort and (2) rating visual experience. Each session will include the 3D reference, and also the 2D version of that reference (e.g., left eye view).

Select reference sequences where

* The 3D version has the same level of visual comfort as the 2D version for a short viewing duration (e.g., no significant difference from a statistical point of view); and
* The 3D version has a higher (i.e., better) visual experience than the 2D version from a statistical point of view.

These 3D reference selection criteria can be obtained from previous subjective tests (e.g., performed by other laboratories). These 3D reference selection criteria can be gathered simultaneously with the target 3D experiment.

Preferably, select scene contents of various video complexities in terms of texture, motion and depth. In fact, to evaluate the impact of the scene content on results, it is important to select original test sequences of different depth levels as well as natural and synthetic ones.

## [Editor’s note: Further studies required on threshold values. 6.4 Discrepancies between left and right images

In stereo 3D systems, a binocular 3D image is formed by presenting the left and right image to their respective eye. If discrepancies arise between these two images, they can cause psychophysical stress, and in some cases 3D viewing can fail. For example, when shooting and displaying stereoscopic 3DTV programmes, there may be geometrical, optical, electrical or temporal asymmetries, such as size inconsistency, vertical shift, rotation error, and luminance or color levels between the left and right images. For the production of natural scene contents using two independent video cameras, the main issue is to guarantee the views asymmetries are under perceptual limits. Table 1 illustrates visibility thresholds obtained from subjective experiments using an impairment scale and for a viewing distance of 4.5 times the display height [1].

|  |  |  |
| --- | --- | --- |
| **Item** | **Description** | **Visibility threshold**  |
| Vertical disparity  | Vertical shift difference (local or global) | 0.4 % |
| Rotation | Rotation difference between the 2 views  | 0.25 deg. |
| Focal length  | Magnification difference  | 0.5 % |
| Black level  | Black level difference between the 2 views  | 3 % |
| White level  | White level difference between the 2 views  | 10 % |
| Colorimetry  | Colorimetry difference considering R, G and B signals  | 10 % |
| Temporal | Temporal asymmetry (shooting or visualization) | To be tested |

Table 1: Visibility thresholds related to left and right views asymmetries

[1] ChenWei, PhD Thesis report, Multidimensional characterization of quality of experience of stereoscopic 3D TV, October 2012.]

# 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 3D 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.

3D subjective experiments may measure opinions on three different perceptual scales:

* Visual experience
* Image quality
* Visual comfort

These perceptual scales must be rated independently. For example, the subject might watch all videos during one session and rate visual experience; then the subject might watch all videos a second time during a different session and rate visual comfort. Other perceptual scales may be of interest (e.g., perceived amount of depth).

This section describes the test methods, rating scales and allowable deviations. The method controls the sequence presentation. The rating scale controls way that people indicate their opinion of the sequences. A list of appropriate changes to the method follows.

## 7.1 Absolute Category Rating (ACR) Method

The Absolute Category Rating (ACR) method is a category judgment where the test sequences are presented one at a time and are rated independently on a category scale. ACR is a Single Stimulus Method. The subject observes one sequence, 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.

<<insert figure here>>

**Comments**

The ACR method produces a high number of ratings in a brief period of time.

ACR ratings confound the impact of the impairment with the influence of the content upon the subject (e.g., whether the subject likes or dislikes the production quality of the sequence).

## 7.2 Degradation Category Rating (DCR) 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 that reference sequence after being impaired by the systems under test. DCR is a Double Stimulus method.

In this case the 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.

<<insert figure here>>

**Comments**

The DCR method produces a fewer ratings than ACR in the same period of time (e.g., slightly more than one-half).

DCR ratings are minimally influenced by subject’s opinion of the content (e.g., whether the subject likes or dislikes the production quality). Thus, DCR is able to detect color impairments and skipping errors that the ACR method may miss.

DCR ratings may contain a slight bias. This occurs because the reference always appears first, and people know that the first sequence is the reference.

## 7.3 Comparison Category Rating (CCR) Method

The Comparison Category Rating (CCR) method is a method where the test sequences are presented in pairs. Two versions of the same stimuli are presented in a randomized order (e.g., reference shown first 50% and second 50% of the time). CCR is a Double Stimulus method. CCR may be used to compare source video with impaired video, or to compare two different impairments.

The subjects are asked to rate the impairment of the second stimulus in relation to the first stimulus. The following seven-level scale for rating the impairment should be used:

 -3 Much Worse

 -2 Worse

 -1 Slightly Worse

 0 The Same

 1 Slightly Better

 2 Better

 3 Much Better

The numbers may optionally be displayed on the scale.

During data analysis, the randomized order of presentation must be removed.

<<insert figure here>>

**Comments**

The CCR method produces a fewer ratings than ACR in the same period of time (e.g., slightly more than one-half).

CCR ratings are minimally influenced by subject’s opinion of the content (e.g., whether the subject likes or dislikes the production quality).

Test subjects will occasionally mistakenly swap their ratings when using the CCR scale (e.g., mark “Much Better” when intending to mark “Much Worse”). This is unavoidable due to human error. These unintentional score swapping events will introduce a type of error into the subjective data that is not present in ACR and DCR data.

The accuracy of CCR is influenced by the randomized presentation of stimuli one and two. For example, when comparing source and degraded video, if the source stimulus is presented first 90% of the time, then CCR will contain the same bias seen in the DCR method.

**7.4 Subjective Assessment Methodology for Video Quality (SAMVIQ)**

The SAMVIQ method defined in [ITU-R BT.1788] is commonly used to subjectively rate 2D video. The SAMVIQ method is also appropriate for use in measuring 3D subjective quality.

This method provides a global quality score for short display duration (10s–20s). It is inspired by the DSCQS (Double Stimulus using a Continuous Quality Scale) method. SAMVIQ is a multi stimulus method: several sequences to evaluate are directly accessible (e.g., played upon request).

SAMVIQ is able to discriminate between low quality as well as high quality video sequences. For this purpose, it combines subjective evaluation capabilities and the ability to discriminate near quality, using an implicit comparison process. The subject can compare each sequence under test with the reference one (i.e., 3D reference sequence without any treatment) and to the other versions of the 3D source. The SAMVIQ method includes a random access to play sequence files. Viewers can start or stop the evaluation, and give, change or keep the current score of each clip when they want. Additionally, they can replay sequences as often as they want.

The SAMVIQ quality evaluation method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each viewer moves a slider on a continuous scale graded from 0 to 100 annotated by 5 linearly spaced quality items (Excellent, good, fair, poor, bad). In the 3D case, three different perceptual scales are used: visual experience, image quality and visual comfort.

Each perceptual scale is rated during a different session.

**Comments:**

The main value of the SAMVIQ method for 3D video subjective quality assessment is to improve rating accuracy for viewers who have little experience viewing 3D content. Moreover, SAMVIQ increases the accuracy of results for each viewer (e.g., fewer judgment errors). This leads to more reliable results in terms of statistical analysis.

## 7.5 Acceptable Changes to the Methods

This section of the Recommendation is intended to be a living document. The methods and techniques described in this section cannot, by their very nature, account for the needs of every subjective experiment. It is expected that the experimenter may need to modify the test method to suit a particular experiment. Such modifications fall within the scope of this Recommendation.

The following acceptable changes have been evaluated systematically. Subjective tests that use these modifications are known to produce repeatable results.

### 7.5.1 Changes to Level Labels

Translating labels into a different languages does not result in a significant change to the MOS. Although the perceptual magnitude of the labels may change, the resulting MOS are not impacted.

An unlabeled scale may be used. For example, ends of the scale can be labeled with the symbols “+” and “-”.

A scale with numbers but no words may be used.

Numbers may be included or excluded at the preference of the experimenter.

Alternate wordings of the labels may be used when the rating labels do not meet the needs of the experimenter. One example is using the DCR method with the ACR labels. One example is using the ACR method with a listening-effort scale as mentioned in ITU-T Rec. P.800. An example specific to 3D, is when assessing visual fatigue and asking about focusing difficulty, to present the following five levels:

* None
* Mild
* Modest
* Bad
* Severe

### 7.5.2 ACR with Hiden Reference (ACR-HR)

An acceptable variant of the ACR method is ACR with Hiden Reference (ACR-HR). With ACR-HR, the experiment includes a reference version of each video segment, not as part of a pair, but as a freestanding stimulus for rating like any other. During the data analysis the ACR scores will be subtracted from the corresponding reference scores to obtain a DMOS. This procedure is known as “hidden reference removal.”

Differential viewer scores (DV) are calculated on a per subject per processed video sequence (PVS) basis. The appropriate hidden reference (REF) is used to calculate DV 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 of 5 indicates ‘Excellent’ quality and a DV 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) will generally be considered valid. Alternatively, a 2-point crushing function may be applied to prevent these individual ACR-HR viewer scores (DV) from unduly influencing the overall mean opinion score:

 crushed\_DV = (7\*DV)/(2+DV) when DV > 5.

**Comments**

ACR-HR will result in larger confidence intervals than ACR, CCR or DCR.

The ACR-HR method removes some of the influence of content from the ACR ratings, however to a lesser extent than CCR or DCR.

ACR-HR should not be used when the reference sequences are fair, poor or bad quality. The problem is that the range of DV ’excellent’ quality diminishes. For example, if the reference video quality is poor on the ACR scale, then DV must be 3 or greater.

## 7.6 Unacceptable Changes to the Methods

The following acceptable changes have been evaluated systematically. These modifications are not allowed.

### 7.6.1 Do Not Increase the Number of Levels

The number of levels should not be increased. Tests into the replicability and accuracy of subjective methods indicate that the accuracy of the resulting MOS does not improve. However, the method becomes more difficult for subjects.

Experiments that compare discrete scales (e.g., 5-point, 9-point, 11-point) with continuous scales (e.g., 100-point scales) all indicate that continuous scales contain more levels than can be differentiated by people. The continuous scales are treated by the subjects as if it were a discrete scale with fewer options (e.g., using five to nine levels).

Prohibited examples include changing ACR from a discrete 5-level scale to a discrete 9-level scale, a discrete 11-level scale, or a continuous scale.

# 8 Environment

## 8.1 Maximum display crosstalk

This issue is currently being investigated in J.3D-disp-req. P.3D-sam may refer to J.3D-disp-req and provide the maximum allowed display crosstalk rate.

## 8.2 Screen Brightness

For 3D displays that use eye glasses, the perceived brightness may be reduced due to the eye glasses. This aspect should be considered in setting the picture brightness for 3D subjective testing. All measurements including screen brightness measurement need to be carried out through glasses according to the 3D display technology.

## 8.3 Viewing distance and angle

[Editor’s note: This section needs further studies.]

In general, the viewing distance is about 3H (three times picture height) for TV environments. For PC monitors, 1H to 3H is recommended. For multimedia applications (e.g., mobile devices), 6H to 10H is recommended.

## 8.4 Viewing conditions

To optimize the 3D viewing environment, some additional details may be necessary, such as suggesting the optimal distance between the display and the back wall and the optimal viewing distance.

## 8.5 Color temperature of 3D displays

[Editor’s note: This section needs further studies.]

Most 3D monitors use LCD displays. Setting the 3D display to a certain color temperature may not be desirable, because such operations may result in a color shift. In general, factory settings may be used provided that such settings provide a natural color appearance.

# 9 Subjects

[Editor’s note: Further studies required on this topic. For 3D studies, the number of subjects used in the experiment are not different from those for 2D studies..]

# 10 Experiment design

If the material is known to contain excessive parallax, and thus known to be potentially uncomfortable, then the duration should be limited.

## 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 other factors such as the total quality range of the test conditions, the experience and expectations of the assessors, etc. 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 above include a “reference” sequence, whenever available, as part of the test sequences set. The “reference” is usually a version of the test sequence that has not undergone any processing (i.e. the original source sequence). The experimental plan might include also the monoscopic version of the “reference” (i.e. only one view of the original source sequence); for example in visual comfort studies it might be useful to use the visual comfort of the monoscopic reference as the baseline.

# 11 Experiment implementation

<Topics for this section include the each subject’s participation in the experiment, which typically consists of the following five stages: (1) Informed consent, (2) Pre-screening of subjects, (3) Instructions and training, (4) Voting sessions, and (5) optionally, questionnaire and/or interview. This section should also contain implementation specific details of the voting sessions (e.g., computer controlled stimuli playback versus video tape / video disc playback of stimuli at fixed intervals with paper ballots).>

Viewer instructions must include guidelines on how to react when subjects feel fatigue or discomfort. See [ITU-T Rec. J.3D-fatigue].

## 11.1 Informed consent

<<tbd>>

## 11.2 Viewer screening

Some populations are less able to perceive 3D content. Additionally, some people are entirely unable to perceive 3D content (e.g., due to blindness in one eye).

In addition to the conventional visual acuity and color vision test, 3D acuity testing should be performed for the viewer. Therefore, acceptable testing procedures should be provided in the new Recommendation.

Other topics: color blind, eye vision, inter-pupillary distance, etc

### 11.2.1 Eye vision test

<<tbd>>

### 11.2.2 Color blindness test

<<tbd>>

### 11.2.3 Stereoscopic acuity test

Tentatively, a maximum angle of stereopsis of 140 seconds is recommended. [editor’s note: this level can be changed later as more information regarding the necessary threshold is obtained.]

### 11.2.1 Inter-pupliary distance

When autostereoscopic monitors are used, inter-pupliary distance is a critical factor. This information should be recorded for each subject. Most autostereoscopic monitors are designed for a fixed inter-pupilary distance, and subjects who deviate to a large extent from that fixed inter-pupilary distance may experience increased crosstalk.

## 11.3 Instructions and training

Instruction should be tailored to dimension (e.g. depth quality, comfort, etc.) under investigation.

Ethical guidelines are critical, since participants might experience visual discomfort. The subjects must be informed of any possible negative 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).

## 11.4 Voting sessions

<<tbd>>

## 11.5 Questionnaire or interview

<<tbd>>

# 12 Data analysis

The results should be reported along with the details of the experimental set-up. For each combination of the test variables, the mean value and the standard deviation of the statistical distribution of the assessment grades should be given. [ Editor’s note: Some items can be mandatory while others need to be reported whenever possible. The method to calculate these statistical values is described in Recommendation ITU-R BT.500.

Perception of 3D contents depends on the shooting parameters and resulting horizontal disparities as well as on the viewing environment. For instance, the perception of the same 3D content for different visualization conditions (viewing distance, screen size, image definition, etc.) won’t provide necessarily the same subjective results and the same level of visual comfort. In order to provide reliable results analysis, it is essential to provide experimental parameters presented in table 1. These parameters could be taken into account for results comparison between laboratories as well as for publication issues.

|  |  |
| --- | --- |
| **Experimental parameter** | **Parameter unit** |
| Maximum crossed and uncrossed disparities for each 3D scene content | In pixels or percentage of the display width |
| Image definition of the display | Number of lines x number of rows |
| 3D video format | Side-by-side, Frame Packing, Top-Bottom, etc. |
| 3D rendering technology | Active shutters, line interleaved display using polarized glasses, autostereoscopic display, etc. |
| Viewing distance | Meter |
| Display size (diagonal and/or width and height) | Meter |
| Maximum luminance on the screen through glasses | cd/m2  |
| Crosstalk level | Percentage of the maximum luminance through glasses |

Table 1: experimental parameters needed for results presentation]

Appendix I

General considerations on 3D video quality

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

The following information should be considered in the development of P.3D-sam (Note: when P.3D-sam is finalized, part of this text will be moved to the appropriate normative section when/if deemed appropriate. The rest will be deleted.)

**1. Quality requirements for 3D TV systems – assessment factors**

Quality factors generally applied to monoscopic television pictures, such as resolution, color rendition, motion portrayal, overall quality, sharpness, depth, etc., could be applied to 3D television systems. In addition, there are many factors peculiar to 3D television systems. Some of them are discussed below.

* The goal of 3DTV viewing experience should be to create the illusion of a real environment. which can be watched for an indefinite period of time by the audience with normal visual acuity;
* The quality of the 3DTV service should be established by two principal parameters sensation of reality (SR) and comfort or ease of viewing; (EV). . These need to be established by subjective evaluations.
* The sensation of reality a viewer sees depends on the combination of quality factors such as resolution, sharpness, color-fidelity, and a group of quality factors termed “depth cues”. The quality factors define the potential for the sensation of reality, but the actual sensation of reality achieved depends on the combination of the quality factors and the scene content itself. The scene content must 'exploit' the quality factors for their effects to be seen.
* Quality factors generally applied to monoscopic television pictures, of course, could be applied to 3D television systems. In addition, there would be many factors peculiar to 3D television systems. Some of them are listed below
	1. **Depth resolution:** spatial resolution in depth direction. Coarse resolution in depth direction may reduce picture quality in 3D television.
	2. **Depth motion:** a factor related to whether motion or movement along depth direction is reproduced smoothly.
	3. **Size distortion “Puppet theatre effect”:** the reproduction magnification ratio of an object at the shooting distance (the perceived size) varies with the imaging and display conditions. The resulting distortion in size may make an object be perceived as unnaturally small.
	4. **Depth distortion “Cardboard effect”:** the imaging and display conditions may reduce the reproduction magnification ratio of depth directions and distort the perception of objects with visually imperceptible thickness. The 3-D positions of stereoscopic objects are perceived stereoscopically but they appear unnaturally thin.
	5. **The frame effect**: 3D pictures appear highly unnatural when objects positioned in front of the screen approach the screen frame. This unnatural effect is called “the frame effect”. The effect is generally reduced with a larger screen, because observers are less conscious of the existence of the frame when the screen is larger.
* One of the quality factors in the group of depth cues is **binocular disparity** - the difference between the pictures seen by the left and right eyes**.** This can be the strongest depth cue for close objects and has a strong influence on the “potency” of the image. Other depth cues include occlusion (objects hidden behind other objects), relative sizes of known objects, vanishing point perception, and others. Depth cues, except for binocular disparity, are provided in monocular planar images (SDTV, HDTV, and UHDTV).Viewers may not feel comfortable viewing left and right images that have size, verticality, inclination, and brightness differences. Cross-talk between the left and right images may also have an impact on viewing comfort.
* 3DTV systems are based on the (additional) provision of binocular disparity - the simultaneous provision of point source left and right eye pictures. Basic technologies (3D) provide two images only, regardless of viewer's head position. Other more developed systems (multi-view) provide a greater number of images, but with the same purpose of providing each eye with separate point source images, but which can change depending on the viewer's head position
* All 3DTV systems displayed on a screen in a single plane (such as a television screen) have limitations for a number of reasons. One of them is the potential conflict between convergence (the object that they eyes point themselves towards) and accommodation (the point on which the lens of the eye focuses) which the two signals gives rise to. The human eye focuses on an object according to the distance to that object. At the same time, we also control the convergence point (gaze point) on the object. Therefore, there is no inconsistency between accommodation and convergence in our everyday life. However when viewing 3D images, the focus point (accommodation) must always be fixed on the screen, independent of the convergence point which is derived from the disparity of the signals. Otherwise, the observer cannot focus clearly. Thus, an inconsistency between accommodation and convergence is introduced in 3D systems Optimizing 3D systems is the process of minimizing the effects of the limitations.
* Attention should be paid to the distribution of parallaxes in the stereoscopic images. From the correlations between psychological factors and the parallax distribution, we can grasp the essential characteristics of stereoscopic images, e.g., the sense of presence they convey and their ease of viewing. The parallax distribution of stereoscopic images is discontinuous during scene-change frames, where the scene depth and perceived convergence distance change. We need to evaluate how these changes affect the visual discomfort experienced during viewing of stereoscopic images.
* Visual fatigue caused by parallax 3DTV viewing: visual fatigue caused by viewing stereoscopic motion images is a particular safety concern. Viewers’ repeated adaptation to the discrepancy between eye convergence and accommodation causes a decline of their visual functions and results in visual fatigue.
* Visual functions vary greatly from person to person, so it is essential to understand that there are individual differences before subjective assessment begins. For instance, there are limits to the binocular parallax of left and right images which a person can fuse into one image; when the parallax exceeds these limits, a double image is perceived. In this situation, depth perception collapses and viewing becomes extremely uncomfortable. For this reason, it is necessary to know the range of binocular parallax over which two images can be fused into one. However, individual differences are vast and will necessitate a study of the stereopsis function of many people.
* Audio systems also play a part in creating a sense of reality for the viewer, and should be arranged so that both vision and sound work together to heighten reality

**2. Assessment methods**

The methods described in Recommendations ITU-R BT.500 and ITU-T P910 could be applied for the evaluation of the general picture quality of 3D TV systems as well as sharpness and depth.. When a reference image is available, double-stimulus continuous quality-scale or double-stimulus impairment scale methods can be used. When no reference is available, the single-stimulus categorical judgement method can be used, for example, to identify the merits of 3D systems. Evaluation methods for the assessment of particular factors of 3D television systems require further study.

**3. Viewing environment and conditions**

The effect of the viewing environment is fundamental on the perception of depth and to the quality of the overall viewing experience. The following situations should be considered:

* Studio/laboratory environment
* Home environment

In particular, in conjunction with viewing distance, picture size and subtended viewing angle play a role in the three-dimensional effect as perceived by the viewer.

Two major factors peculiar to 3D display should be taken into consideration: the display frame effect and inconsistency between accommodation and convergence,

It is generally said that the minimum value for depth of field of the human eye is ±0.3 D, where diopter (D) is the reciprocal value of distance (m). This means that we can perceive the image without defocusing when the object is located within ±0.3 D. When viewing 3D television, the accommodation point is fixed on the screen, and therefore 3D pictures should preferably be displayed within this range. Since ordinary television programs include images at infinite distance, the desirable range of depth to be displayed with 3D systems is considered to be within 0 to 0.6 D. Therefore, 0.3 D, i.e. 3.3 m, is considered to be the optimum viewing distance.

Camera parameters (camera separation, camera convergence angle, focal length of lens), resolution of the system and the frame effect should be taken into account in determining viewing conditions (screen size). In the case of HDTV when watching at the standard viewing distance of 3 *H* (*H* denotes picture height), the viewing distance of 3.3 m corresponds to a 90-inch screen. In the case of standard definition television (SDTV) when watching at the standard viewing distance of 6 *H*, this distance corresponds to a 36‑inch screen. A subjective assessment of the relationship between screen size and depth perception was carried out with 3D HDTV system, and the results showed that the most natural depth perception was obtained with a screen size of 120 inches, which corresponds to viewing distance of 2.2 *H*.

The effective viewing angle should allow 20% angular rotation of head movement in the horizontal plane.

**4. Observers**

Observers should have normal acuity (see Recommendation ITU-R BT.500 or ITU-T P.910). In addition, they should have normal stereopsis, which has to be checked using special binocular vision test materials.

**5 Test materials**

Test materials should contain still and motion sequences of natural scenes.

The 3-D effects obtained from stereoscopic pictures depend largely on the shooting conditions, such as camera separation, camera convergence angle and focal length of the lens.

Appendix II

General considerations on 3D video quality

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

The following information should be considered in the development of P.3D-sam (Note: when P.3D-sam is finalized, part of this text will be moved to the appropriate normative section when/if deemed appropriate and the rest will be deleted. )

# Abstract

Proposal of draft new Recommendation on methods for the subjective assessment of stereoscopic three-dimensional television (3DTV) systems [J.3DTV-sma].

# Scope

This Recommendation provides methodologies for the assessment of stereoscopic television systems including general test methods, the grading scales and the viewing conditions.

# Table of Contents

[Abstract 1](#_Toc308442164)

[Scope 1](#_Toc308442165)

[1. Introduction 1](#_Toc308442166)

[2. General viewing conditions 2](#_Toc308442167)

[3 Test material 3](#_Toc308442168)

[3.1 Visual comfort limits 3](#_Toc308442169)

[4 Experimental apparatus 4](#_Toc308442170)

[5 Observers 4](#_Toc308442171)

[5.1 Sample size 4](#_Toc308442172)

[5.2 Screening 4](#_Toc308442173)

[6 Instruction to observers 4](#_Toc308442174)

[7 Session duration 4](#_Toc308442175)

[8 Subjective methodologies 5](#_Toc308442176)

[9 Use of reference test sequences 5](#_Toc308442177)

[10 Statistical analysis and viewers’ rejection criteria 5](#_Toc308442178)

# 1. Introduction

Stereoscopic three-dimensional television exploits the characteristics of the human binocular visual system by recreating the conditions that bring about the perception of the relative depth of objects in the visual scene. The main requirement of current stereoscopic imaging is the capture of at least two views of the same scene from two horizontally aligned cameras. The images of the objects depicted in the scene will have different relative positions in the left- and right-view. This difference in relative positions in the two views is typically called parallax or disparity, and is usually expressed in pixels or physical distances (a percentage of screen width). It should be noted that parallax should not be confused with angular (retinal) disparity. In fact, the same parallax information would produce different angular (retinal) disparities with different viewing distances. The magnitude and direction of the perception of depth is based on the magnitude and direction of the retinal disparities elicited by the stereoscopic image.

Assessment factors generally applied to monoscopic television pictures, such as resolution, colour rendition, motion portrayal, overall quality, sharpness, etc. could be applied to stereoscopic television systems. In addition, there would be many factors peculiar to stereoscopic television systems. These might include factors such as depth resolution, which is the spatial resolution in depth direction, depth motion, that is, whether motion or movements along depth direction is reproduced smoothly and spatial distortions. Two well-known examples of the latter are the “*puppet theatre effect*, i.e. when objects areperceived as unnaturally large or small, and the *cardboard effect*, i.e. when objects are perceived stereoscopically but they appear unnaturally thin.

We can identify three basic perceptual dimensions which collectively affect the quality of experience provided by a stereoscopic system: *picture quality*, *depth quality*, and *visual comfort*. Some researchers have argued that the psychological impact of stereoscopic imaging technologies could be better measured in terms of more general concepts such as *naturalness* and *sense of presence.*

*Picture quality* refers the perceived quality of the picture provided by the system. This is a main determinant of the performance of any video system. Picture quality is mainly affected by technical parameters and errors introduced by, for example, encoding and/or transmission processes.

*Depth quality* refers to the ability of the system to deliver an enhanced sensation of depth. The presence of monocular cues, such as linear perspective, blur, gradients, etc., convey some sensation of depth even in standard 2D images. However, stereoscopic 3D images contain also disparity information which provides additional depth information and thus an enhanced sense of depth as compared to 2D.

*Visual (dis)comfort* refers to the subjective sensation of (dis)comfort that can be associated with the viewing of stereoscopic images. Improperly captured or improperly displayed stereoscopic images could be a serious source of discomfort.

*Naturalness* refers to *the perception of the stereoscopic image as being a truthful representation of reality (i.e. perceptual realism)*. The stereoscopic image may present different types of distortion which make it less natural. For example, stereoscopic objects are sometimes perceived as unnaturally large or small (puppet theatre effect), or they appear unnaturally thin (cardboard effect).

 *Sense of presence* refers to “the subjective experience of being in one place or environment even when one is situated in another” .

This Recommendation presents information regarding methods and procedures for the assessment of the three primary dimensions: picture quality, depth quality, and visual comfort, outlined above. Methodologies for the assessment of naturalness and sense of presence will be added at a later stage.

# 2. General viewing conditions

The viewing conditions (including screen luminance, contrast, background illumination, viewing distance, etc.) should generally match those used for 2D as described in the proposed preliminary draft new Recommendation ITU‑R BT.[GVC] “General viewing conditions for subjective assessment of quality of television pictures”, which is discussed in ITU-R WP6C. This Recommendation specifies two possible criteria for the selection of the viewing distance. The Design Viewing Distance (DVD) is to be selected. The DVD for a digital system is the distance at which two adjacent pixels subtend an angle of 1 arc-min at the viewer’s eye; and the horizontal design viewing angle as the angle under which an image is seen at its optimal viewing distance.

For example, when expressed in multiples of the picture’s height, the DVD for the 1280 x 720 image resolution system is 4.8H; and that for the 1920 x 1080 family HDTV image resolution system is 3.1H (static images).

For illustrative purposes, Table I reports the design viewing distance in metres for a representative sample of TV set diagonal sizes.

TABLE I - Design viewing distance in meters for various TV set diagonal sizes

|  |  |  |
| --- | --- | --- |
| Diagonal size (inches) | 1920 × 1080 image system | 1280 × 720 image system |
| Design viewing distance (metres) | Design viewing distance (metres) |
| 32 | 1.24 | 1.88 |
| 42 | 1.62 | 2.47 |
| 52 | 2.01 | 3.06 |
| 62 | 2.39 | 3.64 |
| 72 | 2.78 | 4.23 |
| 82 | 3.17 | 4.82 |
| 92 | 3.55 | 5.41 |
| 102 | 3.94 | 5.99 |

It should be noted since two adjacent pixels subtend an angle of 1 arc-min at the viewer’s eye, then at design viewing distance the smallest angular (retinal) disparity that can be represented by the system (i.e. depth resolution) is equal to 1 arc-min (or equivalently 60 arc-sec). This value is about twice the human disparity threshold, which is about 30 arc-sec. Therefore, most viewers should have no difficulty resolving the smallest disparity represented by the 3D system. (This is true for all systems in Table I when presented at the design viewing distance).

# 3 Test material

The selection of the test material should be motivated by the experimental question addressed in the study: e.g. the content of the test sequences (sport, drama, film, etc.) and their spatiotemporal characteristics should be representative of the programmes delivered by the service under study.

In addition, the selected stereoscopic test sequences content should also be normally comfortable to watch. The visual comfort of stereoscopic images depends critically upon the disparity contained in the image. Accordingly, care should be taken to ensure that the disparity does not exceed the limits outlined in the following section, unless the study is specifically aimed at measuring visual comfort. Moreover, whenever possible the statistics: mean, standard deviation, and range (min/max), of the disparity (screen parallax in pixel) distribution of the test sequences should be measured and reported.

## 3.1 Visual comfort limits

Excessive disparity/parallax causes visual discomfort possibly because it worsens the conflict between accommodation and vergence. Therefore, it has been suggested that to minimize the accommodation-vergence conflict, the disparities in the stereoscopic image should be small enough so that the perceived depths of objects fall within a “comfort zone”. Several limits have been proposed. One approach uses a measure of the screen parallax, expressed as a percentage of the horizontal screen size, to specify the limits of comfortable viewing. Values of 1% for crossed/negative disparities and 2% for uncrossed/positive disparities (for a total value of about 3%) have been suggested. According to another approach, the comfort zone is delimited by the depth of field of the eye. For the viewing conditions typical of television broadcast, researchers have assumed a depth of field between ±0.2D (diopters) and ±0.3D (diopters). For a 1920×1080 HDTV image resolution system watched from the design viewing distance of 3.1H, these values correspond approximately to ±2% and ±3% of screen parallax.

Recall that at the design viewing distance two adjacent pixels subtend an angle of 1 arc-min at the viewer’s eye. Thus, 60 pixels correspond to 1 degree of visual angle. This allows us to easily specify the comfort limits in terms of retinal disparity (for an average viewer). For example, for 1920×1080 HDTV image resolution systems, 1% (~19.2 pixels) corresponds approximately to 20 arc-min, 2% to ~40 arc-min and 3% to ~60 arc-min.

It should be noted that even though at the design viewing distance two adjacent pixels always subtend an angle of 1 arc-min, the physical separation (e.g. in mm) between those pixels increases with larger displays (the number of pixels remains the same, but the physical size of the screen increases). Therefore, the higher limits (e.g. ±3%) could result in larger displays in a physical distance between corresponding points (i.e. the parallax of the two views in mm) that exceed the interpupillary distance of the average viewer (~63-65 mm). This could result in increasing discomfort.

In general, since studies using stereoscopic test sequences could elicit some degree of visual discomfort, it is recommended to use, whenever possible, test material whose disparity does not exceed the lower limits, albeit occasional excursions above these limits might be allowed.

# 4 Experimental apparatus

The experimental apparatus (video server, display, etc.) should be capable of displaying full resolution HD test sequences, for example using an HDMI frame-packing format (unfortunately for now this is only possible for 720p target). This would allow greater flexibility in the range of studies that be carried out.

As of today, there is no reference display for 3DTV assessment. The display should exhibit very low cross-talk (ideally below human threshold) and capable of receiving a variety of input formats (without having to manually change settings).

# 5 Observers

## 5.1 Sample size

Sample size considerations for 3D studies are not different from those for 2D studies.

## 5.2 Screening

Observers should be screened for visual acuity, colour, and stereoscopic vision. The latter could be assessed using clinical tests, such as Randot, Titmus, or Frisby stereo tests. These clinical tests usually measures retinal disparities from 20 to 400 arc-sec. The test materials should be similar as for stereoscopic TV, given in ITU-R Recommendation..

# 6 Instruction to observers

Instruction should be tailored to dimension (e.g. depth quality, comfort, etc.) under investigation. Furthermore, ethical guidelines are more stringent than those typically used in image quality assessment since participants might experience visual discomfort. In general, these studies require more care in informing the participant of the motivations of the study as well as any possible negative resulting from exposure to the stimuli used in the study.

# 7 Session duration

If the viewing material is deemed comfortable, then the session duration might be as long as that used for 2D studies (i.e. ~20-40 minutes intermixed with breaks). If the material is known to contain excessive parallax, and thus known to be potentially uncomfortable, then the duration should be limited.

# 8 Subjective methodologies

Many of the standard methods outlined in Recommendation ITU‑R BT.500 could be used, occasionally in a slightly modified form (e.g. different scales), for the assessment of stereoscopic systems. A few methods that could be used for the assessment of picture quality, depth quality, and visual comfort are the following:

* Single – stimulus (SS) methods, sequence duration 8-10 sec.
* Double stimulus continuous quality scale (DSCQS) method, sequence duration 8-10 sec.
* Stimulus-comparison (SC) methods, sequence duration 8-10 sec.
* Single stimulus continuous quality evaluation (SSCQE) method, sequence duration 5 sec.

# 9 Use of reference test sequences

All the methods listed in Section 8 should include a “reference” sequence, whenever available, as part of the test sequences set. The “reference” is usually a version of the test sequence that has not undergone any processing (i.e. the original source sequence). For the 3DTV studies, the main “reference” is the original unprocessed stereoscopic sequence. The experimental plan might include also the monoscopic version of the “reference” (i.e. only one view of the original source sequence); for example in visual comfort studies it might be useful to use the visual comfort of the monoscopic reference as the baseline.

# 10 Statistical analysis and viewers’ rejection criteria

The statistical analyses and the viewers’ rejection criteria should be the same as for 2D studies.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Appendix III**

# Points of investigation for developing P.3D-sam

The points of investigation that are specific to subjective quality assessment of stereoscopic 3D video include:

* Repeatability of a given test methodology:
	+ This is a very crucial point for any subjective testing methodology. Empirical data are needed to prove that a given methodology can produce repeatable and reproducible data.
	+ Repetition of the same experiment (same test set with same methodology) can provide such empirical evidence
* Ability to separately assess the different basic perceptual attributes related to 3D quality (picture quality, viewing comfort and depth quality). An analogy can be made to audio-visual quality where cross-modal interaction between audio and video has been documented. In the same way, the question is whether subjects are able to assess independently visual quality, depth quality and visual comfort. If not, then is it relevant to ask them to judge these separate attributes? See also the point on “role of instructions”.
* Necessity to use anchors (2D and 3D anchors) in the test stimuli:
	+ The potential of 3D lies in the increased quality of experience compared to 2D. Viewers will only embrace 3D if it provides a better viewing experience than 2D. The underlying question is whether or not subjects are more able to judge 3D quality if they are asked to compare it to 2D, instead of simply judging a 3D stimulus on its own (or even in comparison to some 3D reference).
	+ With the hypothesis that subjects know more easily how to judge a 2D video stimulus, one adaptation of the 2D methodologies could make explicit reference to a 2D version of the stimulus. Explicit comparison can be made in the stimulus presentation and/or in the rating scale
* Viewing conditions (e.g., viewing angle):
	+ Currently 3 simultaneous viewers are allowed in front of a 2D HDTV screen in a subjective test. Because of the increase of crosstalk with viewing angle (angular position), this number may need modification (e.g., is a maximum of 1 or 2 viewers a more appropriate number for 3D tests?)
* Display characteristics:
	+ What is the influence of stereoscopic display characteristics (mainly crosstalk level/characteristics) on quality judgment
	+ Method to characterize and select a stereoscopic display for conducting subjective experiments (e.g., maximum crosstalk =< crosstalk threshold)
* Sequence duration:
	+ Short (10-sec) videos have been traditionally used in 2D video subjective testing with overall rating to avoid problems of recency effects. Literature has shown that subjects can confidently provide a judgment of image quality for this range of duration.
	+ The underlying question is whether or not such a short video duration is suitable to assess visual comfort and depth quality. Some works, without providing empirical data but only survey, have suggested that longer duration may be needed.
	+ Alternatively to the use of longer duration, test designs using stimulus repetition may provide a different path of investigation.
* Role of instructions and more elaborated practice session: These two points may need more emphasis in case of 3D than in 2D.
	+ Most subjects are not well experienced with viewing of 3D content. Most of them have viewed maybe a few 3D movies but experience is far from comparable to exposure to 2DTV. As a consequence, subjects may not well understand how they should judge the 3 basic perceptual attributes for two reasons:
		- Firstly, they may not well understand the meaning of the attribute to judge.
		- Secondly, they may not know if they need to consider this attribute alone or not. For example, in judging visual quality, should the perception of depth (depth quality) be taken into account? Should visual comfort be taken into account?
	+ Clear definition of depth quality and visual comfort:
		- Depth quality: from experience, this is usually the most difficult attribute to be judged. As viewers are not so experienced with viewing of 3D content, they usually find it difficult to know how to provide a judgment.
		- Visual comfort: although there is a natural sense in knowing what is and is not comfortable viewing, precise description of symptoms may be necessary.
* Use of additional questionnaires (besides the quality rating):
	+ Use of ad-hoc additional questionnaires (similar to simulator sickness questionnaire) should also be investigated to gain more understanding in how people judge 3D and react to it.
	+ Which questions are relevant in which context? When should these questions be asked?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_