|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |
|  | |
| Working Document toward a Preliminary draft new Recommendation ITU-R BT.[3DTV SubMEth] | |
| Subjective Methods for the Assessment of Stereoscopic Three-Dimensional Television (3DTV) systems | |

##### Summary

This document considers subjective methods for the assessment of stereoscopic (i.e., only two views assumed to be delivered with glasses based technology) systems. The existing recommendation concerning subjective methods for the evaluation of stereoscopic image system: ITU-R Rec. BT. 1438 "Subjective assessment of stereoscopic television pictures", is limited in scope, limited in the range of methodologies considered, and outdated with respect to important technical parameters. Current developments suggest the opportunity for a more focused recommendation built around the image formats described in Rec. ITU-R BT.1543 and Rec. ITU-R BT.709. This document describes the perceptual dimensions that could affect the ‘quality of experience’ of stereoscopic images, and outlines several methodologies, including viewing condition, video material selection, and apparatus that could be used to investigate these perceptual dimensions.

WOrking Document toward a Preliminary draft new RECOMMENDATION ITU-R BT.[3DTV SubMEth]

Subjective Methods for the Assessment of Stereoscopic Three-Dimensional Television (3DTV) systems

##### Scope

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

The ITU Radiocommunication Assembly,

considering

a) that a large amount of information has been collected about the methods used in various laboratories for the assessment of critical performance characteristics of three-dimensional television (3DTV) systems;

b) that examination of these methods shows that there exists a considerable measure of agreement between the different laboratories about a number of aspects of the tests;

c) that the adoption of standardized methods is of importance in the exchange of information between various laboratories;

d) that the introduction of three-dimensional television (3DTV) services might require the development of new image formats, image processing, and transmission techniques whose performance will need to be evaluated though subjective methodologies;

recommends

**1** that the general methods of test, the grading scales and the viewing conditions for the assessment of picture quality, described in the following Annex should be used for laboratory experiments and whenever possible for operational assessments;

Annex 1

# Assessment (Perceptual) Dimensions

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 it usually expressed in pixels or physical distances (a percentage of screen width). It should be noted that parallax should 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 movement along depth direction is reproduced smoothly and spatial distortions. Two well known examples of the latter are the “p*uppet theatre effect*, i.e., when objects areperceived as unnaturally large or small, and the *cardboard effect*, i.e., when objects are perceived 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.*

*Primary Perceptual dimensions*

*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, conveys 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.

**Additional perceptual dimensions**

*Naturalness* refers to *the perception of the stereoscopic image as being a truthful representation of reality (i.e., perceptual realism),* (Seuntiëns 2006). 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*” (Witmer and Singer, 1998).

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 Rec. BT. [GVC...not yet approved!] “General viewing conditions for subjective assessment of quality of television pictures”.

The reason why we would like to keep (as much as possible) the same parameters is twofold. First, in practice users will watch 3DTV with the same displays and viewing conditions as 2D. Secondly, the performance of 3DTV will often need to be measured in relation (i.e., “as compared”) to standard HDTV. In such cases, it might not be possible to keep all parameters the same. Which one should be changed, if any, should be matter of investigation, and the purpose of the experiment.

Rec.BT. [GVC] 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 1280xx720 (Rec. ITU-R BT.1543) image resolution system is 4.8H (static images…should be 4.7!); and that for the 1920x1080 family (Rec. ITU-R BT.709) HDTV image resolution system is 3.1H (static images).

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

|  |  |  |
| --- | --- | --- |
| TABLE I – **Design viewing distance in meters for various TV set diagonal sizes** | | |
|  | **1920 × 1080 Image system** | **1280 × 720 Image system** |
| **Diagonal Size (inches)** | **Design Viewing Distance (meters)** | **Design Viewing Distance (meters)** |
| 32 | 1.2 | 1.9 |
| 42 | 1.6 | 2.5 |
| 52 | 2.0 | 3.1 |
| 62 | 2.4 | 3.7 |
| 72 | 2.8 | 4.3 |
| 82 | 3.2 | 4.9 |
| 92 | 3.6 | 5.5 |
| 102 | 3.9 | 6.1 |

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 [some reference here]. 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 1920x1080 (Rec. ITU-R BT.709) 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 1920x1080 (Rec. ITU-R BT.709) 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. Figure 1 below shows an example of how it would be possible to test different 3DTV transmission formats using a simple single stimulus methodology.

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.

### 5.1 Screening

Observers should be screened for visual acuity, color, 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.

**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 Rec. 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 presented in Tables IIa, IIb, and IIc, respectively.

# 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.

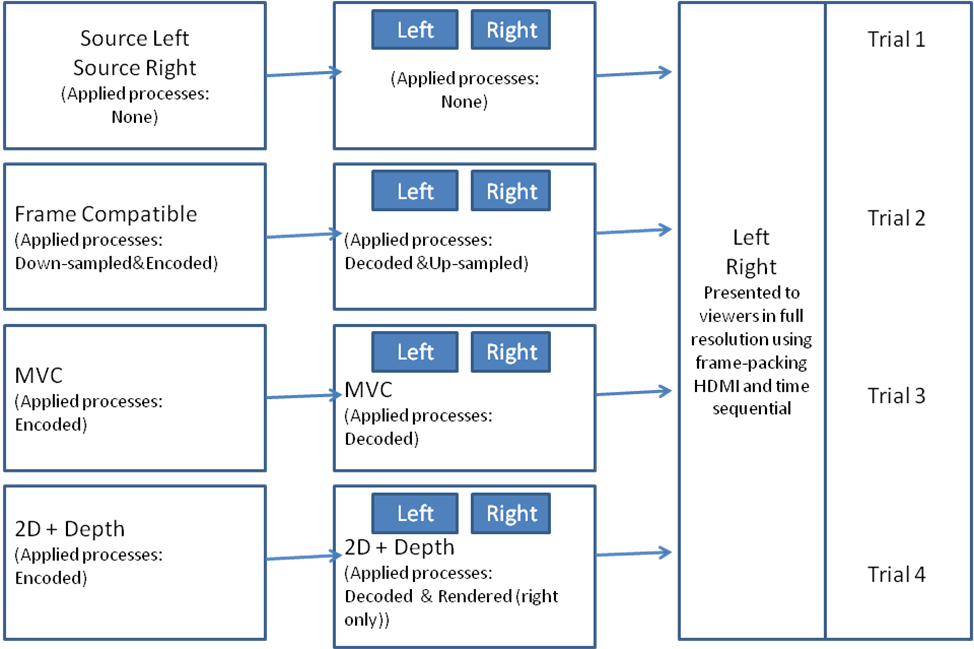


Figure 1

|  |  |  |  |
| --- | --- | --- | --- |
| Table IIa – Subjective Method for the Assessment of Picture Quality | | | |
|  |  | Discrete Scale | Continuous Scale |
| Single-stimulus (SS) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.1.  Sequence Duration: 8-10 sec |  | 5.Excellent  4.Good  3.Fair  2.Poor  1.Bad |  |
| Double stimulus continuous quality scale (DSCQS) method as described in Recommendation ITU R BT.500, Annex 1, § 5.  Sequence Duration: 8-10 sec |  |  |  |
|  |  |  |  |
| Stimulus-comparison (SC) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.2.  Sequence Duration: 8-10 sec |  | -3.Much worse  -2.Worse  -1.Slightly worse  0.The same  1.Slightly better  2.Better  3.Much better |  |
| Single stimulus continuous quality evaluation (SSCQE) method as described in Recommendation ITU R BT.500, Annex 1, § 6.3.  Sequence Duration: 5 minutes |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table IIb – Subjective Method for the Assessment of Depth Quality | | | |
|  |  | Discrete Scale | Continuous Scale |
| Single-stimulus (SS) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.1.  Sequence Duration: 8-10 sec |  | 5.Excellent  4.Good  3.Fair  2.Poor  1.Bad |  |
| Double stimulus continuous quality scale (DSCQS) method as described in Recommendation ITU R BT.500, Annex 1, § 5.  Sequence Duration: 8-10 sec |  |  |  |
|  |  |  |  |
| Stimulus-comparison (SC) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.2.  Sequence Duration: 8-10 sec |  | -3.Much worse  -2.Worse  -1.Slightly worse  0.The same  1.Slightly better  2.Better  3.Much better |  |
| Single stimulus continuous quality evaluation (SSCQE) method as described in Recommendation ITU R BT.500, Annex 1, § 6.3.  Sequence Duration: 5 minutes |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table IIc – Subjective Method for the Assessment of Visual Comfort | | | |
|  |  | Discrete Scale | Continuous Scale |
| Single-stimulus (SS) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.1.  Sequence Duration: 8-10 sec |  | 5.Very comfortable  4.Comfortable  3.Mildly Uncomfortable  2. Uncomfortable  1.Extremely Uncomfortable |  |
| Double stimulus continuous quality scale (DSCQS) method as described in Recommendation ITU R BT.500, Annex 1, § 5.  Sequence Duration: 8-10 sec |  |  |  |
|  |  |  |  |
| Stimulus-comparison (SC) methods as described in Recommendation ITU R BT.500, Annex 1, § 6.2.  Sequence Duration: 8-10 sec |  | -3.Much worse  -2.Worse  -1.Slightly worse  0.The same  1.Slightly better  2.Better  3.Much better |  |
| Single stimulus continuous quality evaluation (SSCQE) method as described in Recommendation ITU R BT.500, Annex 1, § 6.3.  Sequence Duration: 5 minutes |  |  |  |