RECOMMENDATION ITU-R BT.1663

Expert viewing methods to assess the quality of systems for the digital display of large screen digital imagery[1] in theatres

(Question ITU-R 15/6)

(2003)

The ITU Radiocommunication Assembly,

considering

a) that ITU, as well as other international standardization bodies, is studying a new service called large screen digital imagery (LSDI);

b) that several applications will likely be identified for the LSDI service;

c) that it will be necessary to specify performance requirements and to check the suitability of technical solutions considered for each application with the performance requirements of that application;

d) that such checks will also necessarily involve subjective assessment tests under rigorous scientific conditions;

e) that different subjective test methodologies may offer different sensitivities for detecting certain kinds of impairments;

f) that the subjective assessment methods specified in Recommendation ITU-R BT.500 based on non-expert viewing are time-consuming and expensive, in view of the large number of scores that must be collected in order that systems may be differentiated in performance;

g) that a new method of subjective testing, based on the use of a small number of expert viewers, is now proposed that will provide comparable ability to differentiate the performance of different systems while allowing faster, less expensive procedures,

recommends

1 that the test method described in Annex 1, based on the use of expert viewers, should be used in the subjective assessment of LSDI solutions whenever time or budget constraints do not allow use of the non-expert viewing methods specified in Recommendation ITU-R BT.500, and where the sensitivity of the method in Annex 1 is sufficient to differentiate the systems being evaluated,

---

[1] Large screen digital imagery (LSDI) is a family of digital imagery systems applicable to programmes such as dramas, plays, sporting events, concerts, cultural events, etc., from capture to large screen presentation in high-resolution quality in appropriately equipped cinema theatres, halls and other venues.

further recommends

1 that, as part of the testing process, studies be conducted to verify the sensitivity of the test method described in Annex 1;

2 that, in order to improve this Recommendation, further studies be conducted on the processing of the results;

3 test managers and organizations are encouraged to make available to ITU administrations and Sector Members any test materials and test tools (e.g. computer programs to generate side-by-side or mirror image presentations) that are developed, in order to facilitate future testing by other organizations.

Annex 1

Expert viewing to assess the quality of systems for the digital display of LSDI pictures in theatres

1 Introduction

In past years, expert viewing often has been employed to perform a quick verification of the performance of a generic video process.

This Annex describes an expert viewing test method that will ensure consistency of results obtained in different laboratories, using a limited number of expert assessors.

2 Why a new method based on "expert viewing"?

It is useful to point out the advantages resulting from the application of the proposed methodology.

First, a formal subjective assessment test typically requires use of at least 15 observers, selected as "non-experts", requiring lengthy tests and a continuous search for new observers. This number of observers is necessary to achieve the sensitivity necessary so that the systems being tested may be confidently differentiated and ranked, or be confidently judged equivalent.

Second, by using non-expert observers, traditional tests may fail to reveal differences that, with protracted exposure, may become salient, even to non-experts.

Third, traditional assessments typically establish measures of quality (or differences in quality), but do not directly identify the artefacts or other physical manifestations that give rise to these measures.

The methodology proposed here tries to solve all three problems.

3 Definition of expert subjects

For the purpose of this Recommendation, an "expert viewer" is a person that knows the material used to perform the assessment, knows "what to look at" and may or may not be deeply informed on the details of the algorithm used to process the video material to be assessed. In any case, an "expert viewer" is a person with a long experience in the area of quality investigation, professionally engaged in the specific area addressed by the test. As an example, when organizing an "expert viewing" test session on LSDI material, experts in the production or post-production of film or in the production of high-quality video content should be selected (e.g. directors of photography, colour correctors, etc.); this selection has to be made considering the ability to make unique subjective judgements of LSDI image quality and compression artefacts.

4 Selection of the assessors

An expert viewing test is an assessment session based on the opinions of assessors, in which judgements are provided on visual quality and/or impairment visibility.

The basic group of experts is made of five to six subjects. This small number makes it easier to collect assessors, and to reach a faster decision.

According to the experiment needs, it is allowed to use more than one basic group of experts, grouped into a larger combined pool of experts (e.g. coming from different laboratories).

It is recognized that experts may tend to bias their scores when they test their own technology, therefore it should be avoided to include persons that were directly involved in the development of the system(s) under test.

All assessors should be screened for normal or corrected-to-normal visual acuity (Snellen Test) and normal colour vision (Ishihara Test).

5 Test material

Test materials should be selected to sample the range of production values and levels of difficulty foreseen in the real context in which the system(s) under test would be used. Selection should favour more challenging material without being unduly extreme. Ideally, 5-7 test sequences should be used.

The method to select material may vary also in relation to the application for which the system under test has been designed.

In this regard, no further indication is given here on rules for the selection of the test material, leaving the decision to the test designer in relation of the considerations above.

6 Viewing conditions

The viewing conditions, which shall be described fully in the test report, shall be in accordance with Table 1 and shall be kept constant during the test.

TABLE 1

Viewing conditions	Setting(s)	
	Minimum	Maximum
Screen size (m)	6	16
Viewing distance [1]	1.5 H	2 H
Projector luminance (centre screen, peak white)	10 ftL	14 ftL
Screen luminance (projector off)		<1/1 000 of projector luminance

[1] The "butterfly" presentation should be used when the viewing distance is closer to 1.5 H. If the "side-by-side" presentation is used, the viewing distance should be closer to 2 H value.

7 Methodology

7.1 Evaluation sessions

Each evaluation session (defined as the set of test sittings for a given group of observers) should consist of two phases (i.e. Phase I and Phase II).

7.1.1 Phase I

Phase I consists of a formal subjective test performed in a controlled environment (see § 6) which will permit valid, sensitive and repeatable test results. Here, the experts individually rate the material shown using the rating scale described below. Members of the panel are not permitted to discuss what they are seeing or to control the presentations. During this phase, the experts should NOT be aware of the coding scheme under test, or of the order of presentation of the material under test. The material under test will be randomized, so as to avoid any bias in the assessment.

7.1.1.1 Presentation of material

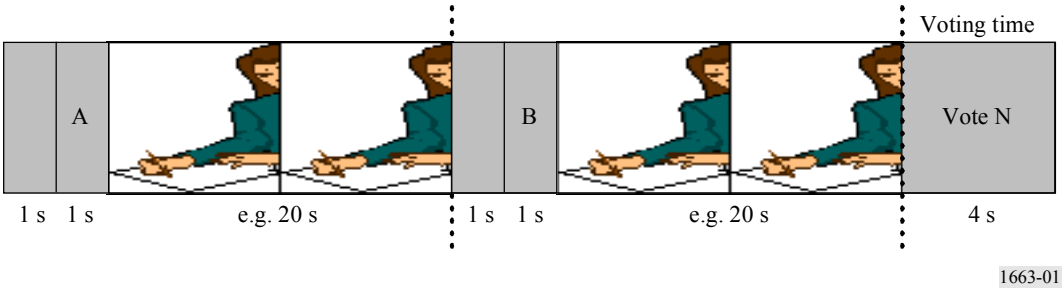
The presentation method combines elements of the simultaneous double stimulus for continuous evaluation (SDSCE) method (Recommendation ITU-R BT.500, § 6.4) and the double stimulus continuous quality scale (DSCQS) method (Recommendation ITU-R BT.500, § 5). For reference, it may be called the simultaneous double stimulus (SDS) method.

As with the SDSCE method, each trial will involve a split-screen presentation of material from two images. In most cases, one of the image sources will be the reference (i.e. source image), while the other is the test image; in other cases, both images will be drawn from the reference image. The reference shall be the source material presented transparently (i.e. not subjected to compression other than that implicit to the source recording medium). The test material shall be the source material processed through one of the systems under test. The bit-rate and/or quality level shall be as specified by the test design. Unlike the SDSCE method, observers will be unaware of the conditions represented by the two members of the image pair.

The split-screen presentation shall be done either using the traditional split screen without mirroring or by the butterfly technique, where the image on the right side of the screen is flipped horizontally. Because full-width images will be used, only half of each image can be displayed at a time. In each presentation, the same half of the image will be shown on each side of the display.

As with the DSCQS method, the image pair is presented twice in succession, once to allow familiarization and scrutiny and once to allow confirmation and rating. Each sequence will be 15-30 s in duration. Each sequence may be labelled at the beginning of each clip to assist assessors (see non-mirrored split screen example shown in Fig. 1).

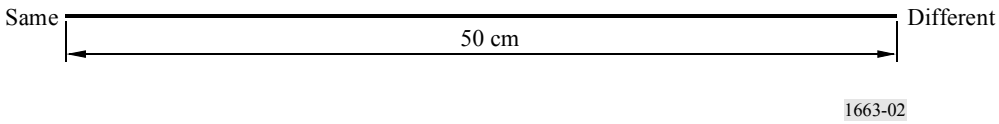
FIGURE 1



7.1.1.2 Judgement scale

The criterion for acceptability in LSDI applications is that the test (i.e. compressed) image be indistinguishable from the reference. Several commonly used scoring methods can be used to evaluate the systems under test. A suggested method is the stimulus comparison scales recommended in Recommendation ITU-R BT.500, § 6.2. A specific example scale is the non-categorical (continuous) SAME-DIFFERENT scale as described in Recommendation ITU-R BT.500, § 6.2.4.2:

FIGURE 2



7.1.1.3 Judgement session

The session, which may involve more than one sitting depending on the number of test conditions, shall involve two types of trials: test trials and check trials. In a test trial, one half of the display shows the reference while the other half shows the test. In a check trial, both halves show the reference. The purpose of the check trial is to assess a measure of judgement bias.

For each system tested, the following test trials are required for each test sequence:

TABLE 2

Left display panel	Right display panel
Left half reference	Left half test
Right half reference	Right half test
Left half test	Left half reference
Right half test	Right half reference

Preferably, there would be at least two repetitions of each of the cases above. For each system, the following check trials are required for each test sequence:

TABLE 3

Left panel	Right panel
Left half reference	Left half reference
Right half reference	Right half reference

Again, preferably there would be at least two repetitions of each of the cases above.

The test session should be divided into sittings not more than one hour in duration separated by 15 min rest periods. Test and check trials resulting from the combination of codec and test sequence should be distributed across sittings by pseudorandom assignment. It is more complex, but worthwhile, to impose some restriction on this process. For example, if there were four sittings, one might randomly assign each of the four test trials for a given codec and test sequence to a randomly determined position in one of the sittings. This approach has the benefit of ensuring that each system's test trials are distributed over the entire test session.

7.1.1.4 Processing of test scores

For a given test trial, the test score is the distance between the "SAME" endpoint of the scale and the mark made by the observer, expressed on a 0-100 scale. The results will be analysed in terms of mean opinion score (MOS), and the MOS will be used to establish rank ordering of the systems tested. Depending on the number of observations per system (observers × test sequences × repetitions), the data may be subjected to analysis of variance (ANOVA)[2]. Performance on check trials can be used to derive a baseline "chance" judgement difference.

---

[2] A total of 10-20 observations in the lowest-order condition of interest is sufficient for application of inferential statistical treatments, such as ANOVA.

7.1.2 Phase II

One of the main goals of Phase II is to refine the relative ranking of the results of Phase I, the precision and reliability of which may be reduced by a limited number of observers and/or judgement trials. A further, and important, objective is to elicit observations as to the characteristics upon which images are perceived to differ and upon which judgements in Phase I were based.

This part involves review by the expert panel of the material shown. Here, the experts are permitted to discuss the material as it is shown, to repeat part or all of the material as many times as necessary for review and/or demonstration, and to arrive at a consensus judgement and a description of what they see. "Trick Play", including the use of modes such as slow motion, single step and still frame, are permitted if requested by the expert viewers. These techniques will require some interaction with, and intervention by, the test manager.

7.1.2.1 Grouping the material under test

To properly perform the Phase II test, it is necessary to group the material under test by content, obtaining a so-called Basic Expert viewing Set (BES), i.e. all the coded sequences obtained from the same source sequence have to be grouped and then ordered in accordance with the ranking derived from Phase I.

The test material will be ordered from the lowest MOS value to the highest MOS value. There will be as many BESs as the number of sequences used for the test.

7.1.2.2 Basic expert viewing test sub-session

A basic expert viewing (BEV) test sub-session is a discussion session during which the experts examine all the material included in a BES; one task is to confirm or modify the ranking order that resulted from the Phase I formal test. Therefore, the relative visibility of differences has to be confirmed or modified.

7.1.2.3 Phase II plan

During Phase II, all the BEVs have to be carried out. The experts will be made aware that the presentation order is the result of the ranking of Phase I. The experts will not be aware of any relationship between proponent systems and ranking.

Phase II will be conducted as a group effort resulting in consensus opinions among the assessors.

Before Phase II begins, assessors will be instructed, possibly using a written text, to perform the following tasks:

– Look at the material in each BEV.

– Discuss the ranking of the material in each BEV; should the group disagree with the ranking, define a new ranking order.

– Comment on each case, providing detailed remarks on the nature of the differences seen, if any.

– Document their rankings, comments and observations.

It will be the responsibility of the test manager to collect all the comments from the groups and to check for discrepancies. While tests are under way, the results of Phases I and II from individual groups will be kept confidential to prevent influencing subsequent groups.

When possible, the test manager is authorized to identify discrepancies and to support resolution by further testing controversial rankings. The aim of this last step is to assure an overall consensus.

8 Report

The final report of the test will be the responsibility of the test manager.

In this report the following information will be provided:

- Results of Phase I (including tables of MOS, as well as the results of statistical analyses, if appropriate).
- Comments from the experts collected during Phase II.
- Comments on any re-evaluation of rankings.
- All relevant information on viewing conditions, input signal characteristics, signal processing, projector characteristics, projector set-up, chromaticity, viewer selection and test conditions.
- A full characterization of the performance of the display device (mean time between failures, etc.).
- Summary and conclusions.