RECOMMENDATION ITU-R BT.1788

Methodology for the subjective assessment of video quality in multimedia applications

(Question ITU-R 102/6)

(2007)

Scope

Digital broadcasting systems will permit the delivery of multimedia and data broadcasting applications comprising video, audio, still-picture, text and graphics. This Recommendation specifies non-interactive subjective assessment methods for evaluating the video quality of multimedia applications.

The ITU Radiocommunication Assembly,

considering

a) that digital broadcasting systems are being introduced in many countries;

b) that multimedia and data broadcasting services which comprise video, audio, still-picture, text, graphics, etc., have been introduced or are planned to be introduced using digital broadcasting systems;

c) that multimedia services will involve a broadcasting infrastructure characterized by the possible use of fixed and mobile receivers, fixed and variable frame rates, different image formats, advanced video codecs, packet loss, etc.;

d) that it will be necessary to specify performance requirements and to verify the suitability of technical solutions considered for each service with the performance requirements of that service;

e) that such verification will principally involve subjective assessment of video quality under controlled conditions;

f) that the subjective assessment methodologies specified in Recommendation ITU-R BT.500 may be used for multimedia applications;

g) that subjective assessment methodologies other than those specified in Recommendation ITU-R BT.500 may also be used;

h) that the adoption of standardized methods is of importance in the exchange of information between various laboratories,

recommends

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

2 that the full descriptions of test configurations, test materials, observers and methods should be provided in all test reports;

3 that, in order to facilitate the exchange of information between different laboratories, the collected data should be processed in accordance with the statistical techniques detailed in Annex 2.

NOTE 1 – The development of a library of video material appropriate for the subjective assessment of video quality in multimedia applications needs to be further pursued by Radiocommunication Study Group 6.

Annex 1

Description of assessment methods

1 Introduction

Many countries have begun deploying digital broadcasting systems that will permit the delivery of multimedia and data broadcasting applications comprising video, audio, still-picture, text and graphics.

Standardized subjective assessment methods are needed to specify performance requirements and to verify the suitability of technical solutions considered for each application. Subjective methodologies are necessary because they provide measurements that allow industry to more directly anticipate the reactions of end users.

The broadcasting system needed to deliver multimedia applications is markedly different from the one currently in use: information is accessed through fixed and/or mobile receivers; the frame rate can be fixed or variable; the possible image size has a large range (i.e. SQCIF to HDTV); the video is typically associated with embedded audio, text and/or sound; the video may be processed with advanced video codecs; and the preferred viewing distance is highly dependent on the application.

The subjective assessment methods specified in Recommendation ITU-R BT.500 should be applied in this new context. In addition, investigations of multimedia systems might be carried out with new methodologies to meet the user requirements of the characteristics of the multimedia domain.

This Recommendation describes non-interactive subjective assessment methods for evaluating the video quality of multimedia applications. These methods can be applied for different purposes including, but not limited to: selection of algorithms, ranking of audiovisual system performance and evaluation of the video quality level during an audiovisual connection.

Terms and definitions relating to this Recommendation can be found in Appendix 3 to Annex 1.

2 Common features

2.1 Viewing conditions

Recommended viewing conditions are listed in Table 1. The size and the type of display used should be appropriate for the application under investigation. Since several display technologies are to be used in multimedia applications, all relevant information concerning the display (e.g. manufacturer, model and specifications), used in the assessment should be reported.

When PC-based systems are used to present the sequences, the characteristics of the systems (e.g. video display card) should also be reported.

Table 2 shows an example of the data record for the configuration of multimedia system under test.

If the test images are obtained using a specific decoder-player combination, the images must be separated from the proprietary skin to get an anonymous display. This is necessary to ensure that the quality assessment is not influenced by the knowledge of the originating environment.

When the systems assessed in a test use reduced picture format, such as CIF, SIF or QCIF, etc., the sequences should be displayed on a window of the display screen. The colour of the background on the screen should be 50% grey.

TABLE 1

Recommended viewing conditions as used in multimedia quality assessment

Parameter	Setting
Viewing distance[(1)]	Constrained: 1-8 H Unconstrained: based on viewer's preference
Peak luminance of the screen	70-250 cd/m^2
Ratio of luminance of inactive screen to peak luminance	≤ 0.05
Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white	≤ 0.1
Ratio of luminance of background behind picture monitor to peak luminance of picture[(2)]	≤ 0.2
Chromaticity of background[(3)]	D_{65}
Background room illumination[(2)]	≤ 20 lux

[(1)] Viewing distance in general depends on the application.

[(2)] This value indicates a setting allowing maximum detectability of distortions, for some applications higher values are allowed or they are determined by the application.

[(3)] For PC monitors, the chromaticity of background should approximate as much as possible the chromaticity of "white point" of the display.

TABLE 2

Configuration of the multimedia system under test

Parameter	Specification
Type of display	
Display size	
Video display card	
Manufacturer	
Model	
Image information	

2.2 Source signals

The source signal provides the reference picture directly and the input for the system under test. The quality of the source sequences should be as high as possible. As a guideline, the video signal should be recorded in multimedia files using YUV (4:2:2, 4:4:4 formats) or RGB (24 or 32 bits). 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.

2.3 Selection of test materials

The number and type of test scenes are critical for the interpretation of the results of the subjective assessment. Some processes may give rise to a similar magnitude of impairment for most sequences. In such cases, results obtained with a small number of sequences (e.g. two) should provide a meaningful evaluation. However, new systems frequently have an impact that depends heavily on the scene or sequence content. In such cases, the number and type of test scenes should be selected so as to provide a reasonable generalization to normal programming. Furthermore, the material should be chosen to be "critical but not unduly so" for the system under test. The phrase "not unduly so" implies that the scene could still conceivably form part of normal television programming content. A useful indication of the complexity of a scene might be provided by its spatial and temporal perceptual characteristics. Measurements of spatial and temporal perceptual characteristics are presented in more detail in Appendix 1 to Annex 1.

2.4 Range of conditions and anchoring

Because most of the assessment methods are sensitive to variations in the range and distribution of conditions seen, judgment sessions should include the full ranges of the factors varied. However, this may be approximated with a more restricted range, by also presenting some conditions that would fall at the extremes of the scales. These may be represented as examples and identified as most extreme (direct anchoring) or distributed throughout the session and not identified as most extreme (indirect anchoring). If possible, a large quality range should be used.

2.5 Observers

The number of observers after screening should be a least 15. They should be non-expert, in the sense that they are not directly concerned with picture quality as part of their normal work and are not experienced assessors. Prior to a session, the observers should be screened for (corrected to) normal visual acuity on the Snellen or Landolt chart and for normal colour vision using specially selected charts (e.g. Ishihara).

The number of assessors needed depends upon the sensitivity and reliability of the test procedure adopted and upon the anticipated size of the effect sought.

Experimenters should include as many details as possible on the characteristics of their assessment panels to facilitate further investigation of this factor. Suggested data to be provided could include: occupation category (e.g. broadcast organization employee, university student, office worker), gender and age range.

2.6 Instructions for the assessment

Assessors should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, timing, etc. Training sequences demonstrating the range and the type of the impairments to be assessed should be used with scenes other than those used in the test, but of comparable sensitivity.

2.7 Experimental design

It is left to the experimenter to select the experimental design in order to meet specific cost and accuracy objectives. It is preferable to include at least two replications (i.e. repetitions of identical conditions) in the experiment. Replications make it possible to calculate individual reliability and, if necessary, to discard unreliable results from some subjects. In addition, replications ensure that learning effects within a test are to some extent balanced out. A further improvement in the handling of learning effects is obtained by including a few "dummy presentations" at the beginning of each test session. These conditions should be representative of the presentations to be shown later during the session. The preliminary presentations are not to be taken into account in the statistical analysis of the test results.

A session, that is a series of presentations, should not last more than half an hour.

When multiple scenes or algorithms are tested, the order of presentation of the scenes or algorithms should be randomized. The random order might be amended to ensure that the same scenes or same algorithms are not presented in close temporal proximity (i.e. consecutively).

3 Assessment methods

The video performance of multimedia systems can be examined using Recommendation ITU-R BT.500 methodologies. A list of selected methods is provided in § 3.1.

In § 3.2 is a description of an additional methodology, called SAMVIQ, that takes advantage of the characteristics of multimedia domain and can be used for the assessment of the performance of multimedia systems.

3.1 Recommendation ITU-R BT.500 methodologies

The following Recommendation ITU-R BT.500 methodologies should be used for the assessment of video quality in multimedia systems:

– Double stimulus impairment scale (DSIS) method as described in Recommendation ITU-R BT.500, Annex 1, § 4.

– Double stimulus continuous quality scale (DSCQS) method as described in Recommendation ITU-R BT.500, Annex 1, § 5.

– Single-stimulus (SS) methods as described in Recommendation ITU-R BT.500, Annex 1, § 6.1.

– Stimulus-comparison (SC) methods as described in Recommendation ITU-R BT.500, Annex 1, § 6.2.

– Single stimulus continuous quality evaluation (SSCQE) method as described in Recommendation ITU-R BT.500, Annex 1, § 6.3.

3.2 Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ)

In this method, the viewer is given access to several versions of a sequence. When all versions have been rated by the viewer, the following sequence content can be then accessed.

The different versions are selectable randomly by the viewer through a computer graphic interface. The viewer can stop, review and modify the score of each version of a sequence as desired. This method includes an explicit reference (i.e. unprocessed) sequence as well as several versions of the same sequence that include both processed and unprocessed (i.e. a hidden reference) sequences. Each version of a sequence is displayed singly and rated using a continuous quality scale similar to the one used in the DSCQS method. Thus, the method is functionally very much akin to a single stimulus method with random access, but an observer can view the explicit reference whenever observer wants, making this method similar to one that uses a reference.

The SAMVIQ quality evaluation method uses a continuous quality scale to provide a measurement of the intrinsic quality of video sequences. Each observer moves a slider on a continuous scale graded from 0 to 100 annotated by 5 quality items linearly arranged (excellent, good, fair, poor, bad).

Quality evaluation is carried out *scene by scene* (see Fig. 1) including an *explicit reference, a hidden reference and various algorithms*.

To get a better understanding of the method, the following specific words are defined below:

Scene: audio-visual content

Sequence: scene with combined processing or without processing

Algorithm: one or several image processing techniques.

3.2.1 Explicit, hidden reference and algorithms

An evaluation method commonly includes quality anchors to stabilize the results. Two high quality anchors are considered in the SAMVIQ method for the following reasons. Several tests have been carried out that indicate minimized standard deviations of scores by using an *explicit reference* rather than a hidden or no reference. Particularly to evaluate codec performance, it is better to use an explicit reference to get the maximum reliability of results. A *hidden reference* is also added to evaluate intrinsic quality of the reference, instead of the explicit reference, because the presentation is anonymous as well as processed sequences. The explicit name "reference" has an influence on about 30% of observers. These observers give the highest possible score (100) to the explicit reference and this score is totally different from the corresponding score of the hidden reference. Notably, when there is no available reference the test remains possible but the standard deviation is dramatically increased.

The SAMVIQ method is appropriate for multimedia context since it is possible to combine different features of image processing such as codec type, image format, bit-rate, temporal updating, zooming, etc. One of these features or a combination of them is summarized by the name *algorithm*.

3.2.2 Test conditions

Variation of criticality during a scene is limited because homogeneous contents are chosen following the same rules implicitly used by other methodologies providing a global score (e.g. single stimulus methods). A maximum sequence viewing duration of 10 or 15 s is then sufficient to get a stabilized and reliable quality score. The proprietary decoder-players, or a screen copy of their output, should be used to maintain the appropriate display performance.

3.2.3 Test organization

a) The test is carried out scene by scene as it is described in Fig. 1.

b) For the current scene, it is possible to play and score any sequence in any order. Each sequence can be played and scored several times.

c) From one scene to another, the sequence access is randomized and prevents the observers from attempting to vote in an identical way according to an established order. In fact, inside a test the algorithm order remains the same to simplify analysis and presentation of results. Only, the corresponding access from an identical button is randomized.

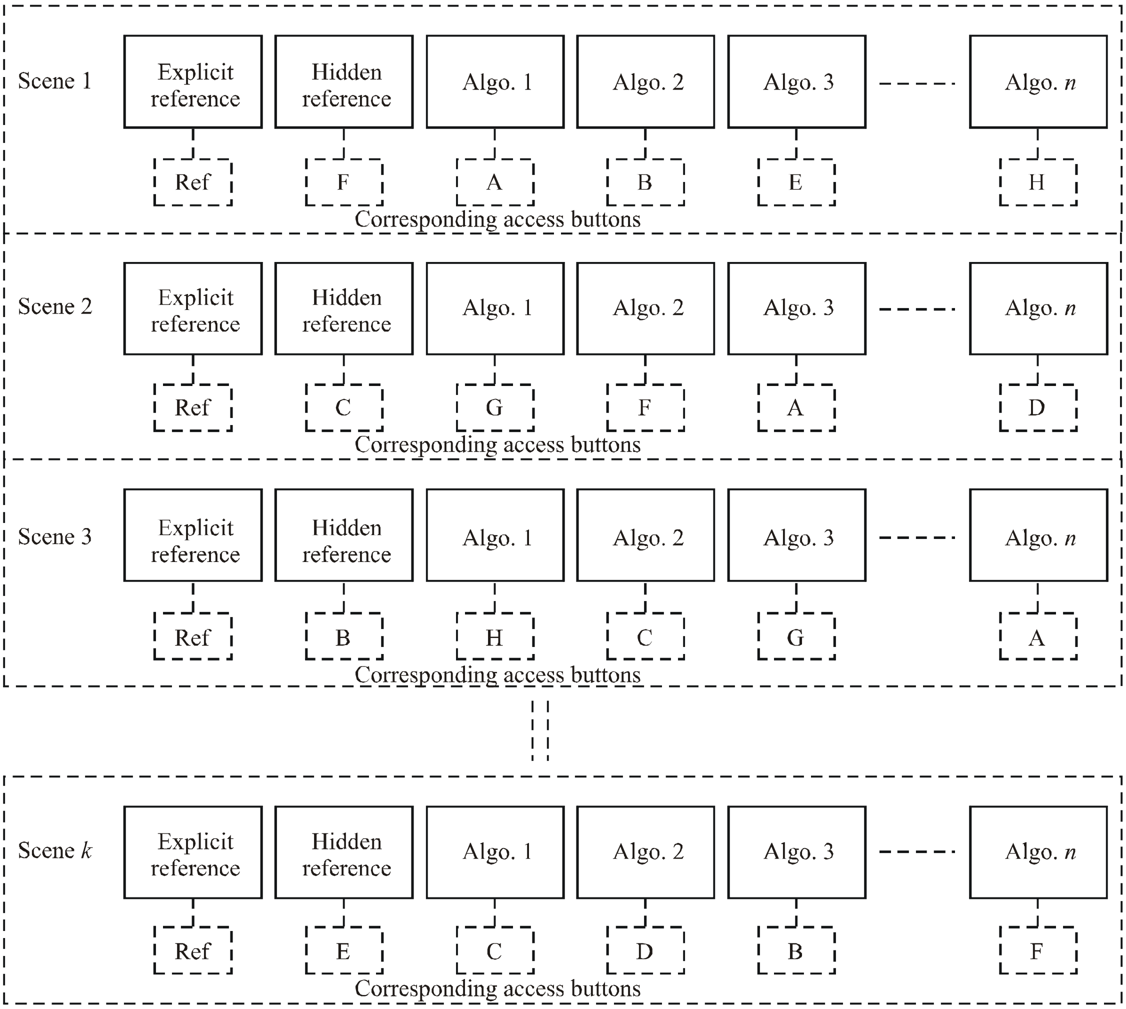
d) For a first viewing, the current sequence must be totally played before being scored; otherwise it would be possible to score and stop immediately.

e) To test the next scene all sequences of the current scene must be scored.

f) To finish the test all the sequences of all the scenes must be scored.

FIGURE 1

Test organization example for the SAMVIQ method



1788-01

The SAMVIQ method is implemented via software. In addition to the access buttons shown in Fig. 1, “play”, “stop”, “next scene” and “previous scene” buttons are necessary to allow the viewer to manage the presentation of the different scenes (see Appendix 2 to Annex 1, for example). When a score has been given by the viewer, it should be shown under the access button corresponding to that scene. When all different versions of a sequence have been graded, the viewer is still allowed to compare scores and modify, if necessary, score values. It is not necessary to review the whole current sequence because large differences have been already highlighted during the first pass viewing.

Annex 2

Presentation and analysis of data

1 Summary information

Accurate information about the test environment is necessary to replicate a test or to compare results across different tests. Therefore, it is suggested to report information about the test environment as described in Table 3.

TABLE 3

Test summary information

Name of the method	
Display technology	
Reference name of the display	
Peak luminance level (cd/m²)	
Black luminance level (cd/m²)	
Black level setup: PLUGE (black to supra black level distance perceived threshold = 8). Otherwise indicates the threshold value	
Background luminance level (cd/m²)	
Illumination (lux)	
Viewing distance: – Not constrained: front of display – Constrained: nH	
Display size (diagonal in inches)	
Width/height display ratio	
Display format (number of columns and lines)	
Image input format (number of columnsand lines)	
Image output format[(1)] (number of columns and lines)	
White colour temperature: D65 otherwise White colour coordinates (x, y)	
Number of effective observers	

[(1)] This information is required when the input image is processed, e.g. rescaled, upon display.

Display characteristics may have an influence on the test results. Additional information such as luminance response (gamma fidelity) and colour primaries should be required for flat panel displays.

The characteristics of the video sequences are important to design a test or explain its results. It is suggested to report spatio-temporal characteristics as described in Appendix 1 of Annex 1. This information should be considered in the collection of test sequences in the library of video material appropriate for the subjective assessment of video quality in multimedia applications.

2 Methods of analysis

The methods of analysis are those described in Recommendation ITU-R BT.500, Annex 2, § 2.

3 Screening of the observers

For the methods listed in Annex 1, § 3.1, screening procedures are described in Recommendation ITU-R BT.500, Annex 2, § 2.3.

The screening for SAMVIQ is described in the next section. However, this procedure could be used for the SS, DSIS and DSCQS methods. The procedure is simpler to implement than the corresponding one used in the Recommendation ITU-R BT.500 for those methods.

3.1 SAMVIQ screening procedure

Each observer must have stable and coherent method to vote fairly degradation of quality for each scene and algorithm. The rejection criteria verifies the level of consistency of the scores of one observer according to mean score over all observers for a given test session. In the SAMVIQ method, as in the DSQCS method, all algorithms (hidden or implicit reference, low anchor, encoded sequences) can be considered. The decision criterion is based on a correlation of individual scores against corresponding mean scores from all the observers of the test.

3.2 Pearson correlation

The relationship between the quality scale and score range of observers is supposed to be linear to apply the Pearson correlation.

The major aim is to verify by a simple method if the scores of one observer are consistent to mean scores of all observers on the whole of the session test. The hidden reference is considered as a high quality anchor. If the low and high anchors are included they increase the correlation score, conversely the correlation offsets between the observers are decreased.

$$r(x,y)=\frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}$$

where:

x_i: mean score of all observers for the triplet (algo, bit rate, scene)

y_i: individual score of one observer for the same triplet

n: (number of algo) × (number of scenes)

i: {codec number, bit rate number, scene number}.

3.3 Spearman rank correlation

The Spearman rank correlation can be applied even if the relationship between the quality scale and score range of observers is not supposed to be linear[1]:

$$r(x,y) = \left[1 - \frac{6 \times \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2}{n^3 - n} \right]$$

where:

x_i: mean score of all observers for the triplet (algo, bit rate, scene)

y_i: individual score of one observer for the same triplet

n: (number of algo) × (number of scenes)

$R(x_i$ or $y_i)$: ranking order

i: {codec number, bit-rate number, scene number}.

3.4 Final rejection criteria for discarding an observer of a test

The Spearman rank and Pearson correlations are carried out to discard observer(s) according to the following conditions:

IF [mean(r) – sdt(r)] > Max Correlation Threshold (MCT).
Rejection threshold = Max Correlation Threshold (MCT).
ELSE Rejection threshold = [mean(r) − sdt(r)].

IF [r (Observer $_i$)] > Rejection threshold.
THEN observer "i" of the test is not discarded.
ELSE observer "i" of the test is discarded.

where:

r = min (Pearson correlation, Spearman rank correlation)

mean(r): average of the correlations of all the observers of a test

sdt(r): standard deviation of all observers' correlations of a test

Max Correlation Threshold (MCT) = 0.85.

The 0.85 MCT value is valid for SAMVIQ and DSCQS methods, otherwise 0.7 MCT value has to be considered for SS and DSIS methods.

[1] Generally Pearson correlation results are very close to Spearman ones.

Appendix 1
to Annex 1

The spatial and temporal information measures

The spatial and temporal information measures given below are single-valued for each frame over a complete test sequence. This results in a time series of values which will generally vary to some degree. The perceptual information measures given below remove this variability with a maximum function (maximum value for the sequence). The variability itself may be usefully studied, for example with plots of spatial-temporal information on a frame-by-frame basis. The use of information distributions over a test sequence also permits better assessment of scenes with scene cuts.

Spatial perceptual Information (SI): A measure that generally indicates the amount of spatial detail of a picture. It is usually higher for more spatially complex scenes. It is not meant to be a measure of entropy nor associated with information as defined in communication theory. The spatial perceptual information, *SI*, is based on the Sobel filter. Each video frame (luminance plane) at time n (F_n) is first filtered with the Sobel filter [Sobel (F_n)]. The standard deviation over the pixels (std_{space}) in each Sobel-filtered frame is then computed. This operation is repeated for each frame in the video sequence and results in a time series of spatial information of the scene. The maximum value in the time series ($\max_{time}$) is chosen to represent the spatial information content of the scene. This process can be represented in equation form as:

$$SI = \max_{time} \{std_{space} [\text{Sobel}(F_n)]\}$$

Temporal perceptual Information (TI): A measure that generally indicates the amount of temporal changes of a video sequence. It is usually higher for high motion sequences. It is not meant to be a measure of entropy, nor associated with information as defined in communication theory.

The measure of temporal information, *TI*, is computed as the maximum over time ($\max_{time}$) of the standard deviation over space (std_{space}) of $M_n(i, j)$ over all i and j.

$$TI = \max_{time} \{std_{space} [M_n(i, j)]\}$$

where $M_n(i, j)$ is the difference between pixels at the same position in the frame, but belonging to two subsequent frames, that is:

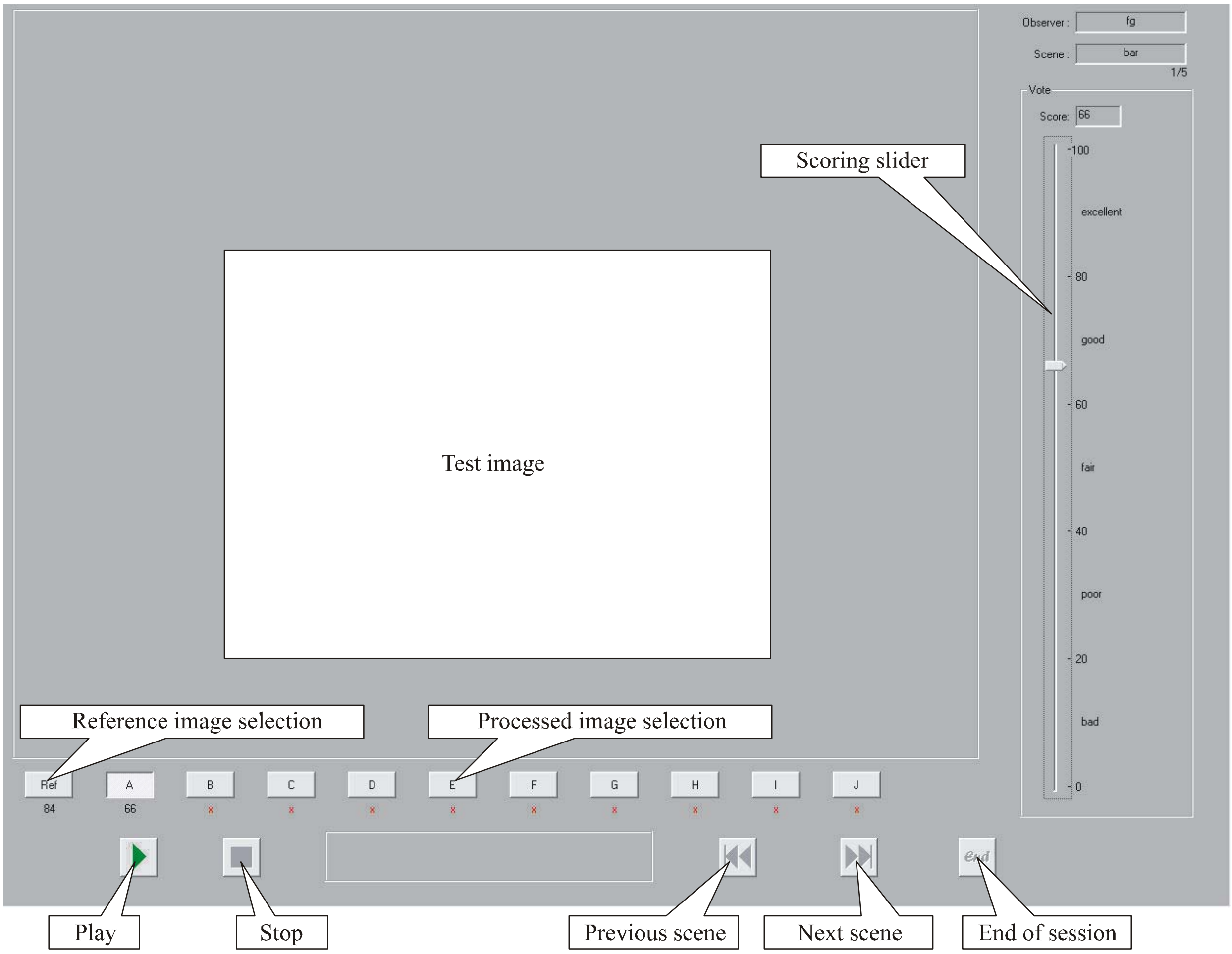
$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$

where $F_n(i, j)$ is the pixel at the i-th row and j-th column of n-th frame in time.

NOTE 1 – For scenes that contain scene cuts, two values may be given: one where the scene cut is included in the temporal information measure and one where it is excluded from the measurement.

Appendix 2
to Annex 1

Example of interface for SAMVIQ



Observer: fg

Scene: bar

Score: 66

1788-02

Appendix 3
to Annex 1

Terms and definitions

Algorithm	One or several image processing operations
AVI	Audio video interleaved
CCD	Charge coupled device
CI	Confidence interval

CIF	Common intermediate format (picture format defined in Recommendation H.261 for video phone: 352 lines × 288 pixels)
CRT	Cathode ray tube
DSCQS	Double stimulus using a continuous quality scale method
DSIS	Double stimulus using an impairment scale method
LCD	Liquid crystal display
MOS	Mean opinion score
SC	Stimulus comparison method
PDP	Plasma display panel
PS	Programme segment
QCIF	Quarter CIF (picture format defined in Recommendation H.261 for video phone: 176 lines × 144 pixels)
SAMVIQ	Subjective assessment of multimedia video quality
Sequence	Scene with combined processing or without processing
Scene	Audiovisual content
S/*N*	Signal-to-noise ratio
SI	Spatial information
SIF	Standard intermediate format [picture formats defined in ISO 11172 (MPEG-1): 352 lines × 288 pixels × 25 frames/s and 352 lines × 240 pixels × 30 frames/s]
SP	Simultaneous presentation
SQCIF	Sub-QCIF
SS	Single stimulus method
SSCQE	Single stimulus using a continuous quality evaluation method
std	Standard deviation
TI	Temporal information
TP	Test presentation
TS	Test session
VTR	Video tape recorder