TITLE: Review of the State of the Art on Subjective Image Quality Assessment

SOURCE: WG1

EDITORS: Michela Testolina (EPFL) (michela.testolina@epfl.ch)

PROJECT: ISO/IEC 29170 (JPEG AIC)

STATUS: Approved

DISTRIBUTION: Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch
# Table of Contents

1. INTRODUCTION ........................................................................................................................................2

2. SUBJECTIVE ASSESSMENT METHODS REVIEW .......................................................................................3
   2.1 JPEG AIC Part-1 ..................................................................................................................................3
   2.1.1 Review of ITU-R BT.500 methods ...............................................................................................3
   2.1.2 Review of ITU-R BT.1082 methods .............................................................................................6
   2.2 JPEG AIC Part-2 ..................................................................................................................................6

3. BEST PRACTICES REVIEW .........................................................................................................................7
   3.1 CONTROLLED LAB ENVIRONMENT ...............................................................................................7
   3.2 CROWDSOURCING ENVIRONMENT .................................................................................................9

4. UNCOVERED QUALITY RANGES ...................................................................................................................11
   4.1 HIGH QUALITY TO NEAR-VISUALLY LOSSLESS QUALITY ASSESSMENT .......................................12
1. INTRODUCTION

Image compression plays a fundamental role, making possible efficient storage and delivery of a large number of rich media such as images and video. While ensuring a lower storage consumption, lossy compression introduces distortions that might become visible to the human eye, and in variable amount depending on the content, compression ratio and context such as the environment and the display type. Each compression method produces different artifacts in images, including blocking, blurring or ringing artifacts, color shift, and others. In order to standardize new and efficient lossy compression methods, it is fundamental to conduct perceptual visual quality assessment experiments to assess the impact of the introduced visual artifacts.

The visual quality can be assessed objectively through a number of objective quality metrics, or subjectively by collecting individual opinions on the presented distorted content from a large number of people, often known as subjects or observers. While objective quality assessment is fast and inexpensive, subjective quality assessment is slow and costly, but more reliable as based on the opinion of human observers. Recommendation documents on subjective visual quality assessment methodologies have been proposed by both ITU-T and ITU-R, and reviewed by the Joint Photographic Experts Group (JPEG) in the context of the AIC activity.

In recent years, and with the increasing number of high-quality capture and display devices accessible to the general public, the interest in visually lossless image compression has rapidly grown. In this context, JPEG released standardized methods for assessing the quality of visually lossless approaches in AIC Part-2.

Most methods are focused on controlled laboratory environments and traditional image modalities. In recent years, motivated by the confinements caused by the COVID-19 pandemic, many subjective quality experiments have been conducted using an uncontrolled crowdsourcing approach, in which subjects are hired remotely, conducting the subjective experiment in their own environments.

The aim of this document is to review a wide range of subjective visual quality assessment methods and standards, focusing on image compression applications. Similar review works can be found in the state of the art: [1] compared three different subjective quality assessment methods in a controlled laboratory environment, and [2] compared six different subjective video quality assessment methods. This study aims at reviewing the existing standards and best practices for visual quality assessment paying special attention to the JPEG AIC framework and its previous standardization efforts.
2. SUBJECTIVE ASSESSMENT METHODS

2.1 JPEG AIC Part-1


Subjective methodologies can generally be divided into two main categories: single stimulus (SS) and double stimulus (DS) methods. The main difference is in the number of stimuli that are presented to subjects, where in the SS the subjects rate the quality of one single image and therefore without using any explicit reference, while in the DS methods the subjects rate the visual quality given two stimuli at the time. A third less-known category, not covered in this document, are triple stimulus methods which aim at rating the visual quality given three stimuli at the time.

2.1.1 Review of ITU-R BT.500 methods

2.1.1.1 Single stimulus methods

The single stimulus (SS) methods consist of presenting to the observers a sequence of images, one at a time, asking to rate their visual quality without any reference. This subjective method is a common choice due to its simplicity and the low number of comparisons. As an example, single stimulus subjective quality assessment was used in several experiments such as [6] or [7].

**Absolute Category Rating (ACR):** the test stimuli are presented one at a time and the subjects are asked to rate the visual quality of the images on a discrete scale from 1 to 5, namely (1) bad, (2) poor, (3) fair, (4) good, (5) excellent.

The advantages of this method are the simplicity of its design and the computation of the subjective scores; however, it usually requires long training sessions to acquaint the subjects with the grading scale. Moreover, the subjective scores are occasionally influenced by the subjects’ opinions on the content of the stimulus. To mitigate the influence of the content on the subjective scores, the ACR-HR method may be used.

**Absolute Category Rating with Hidden Reference (ACR-HR):** is a variation of the ACR method where the original image is "hidden" among the distorted stimuli, without informing the subjects of such occurrence. This experimental method allows to remove the variance due to the subjects' personal opinion on the content, and allows to compute the differential mean opinion score (DMOS) rather than the mean opinion score (MOS), thus obtaining a more precise evaluation of the quality of the stimuli. Due to the trade-off between
its simplicity and accuracy, this method is widely used in the state of the art. As an example, this method was utilized for the large-scale and publicly available subjective quality assessment dataset collected by the Laboratory for Image and Video Engineering (LIVE) \[6\], and in \[7\] to evaluate the perceptual quality of learning-based image compression.

**Single Stimulus Continuous Quality Evaluation (SSCQE):** similar to the ACR method, but uses a continuous evaluation scale rather than a discrete one. The advantage of the continuous quality scale is its similarity with the continuous grading scale of objective quality assessment, which leads to a more accurate comparison with objective quality methods.

2.1.1.2 Double stimulus methods

Double stimulus (DS) methods consist in showing to the subjects pairs of stimuli and asking to evaluate the quality of both or other features like the impairment between the two. The grading scale differs from experiment to experiment. Double stimulus methods are in general more time-consuming when compared to single stimulus methods, but are more accurate on specific types of artifacts, such as in the case of shifts in the colors. For this reason, the method has been used recently in multiple subjective quality experiments, e.g. \[8\] and \[9\].

**Double Stimulus Impairment Scale (DSIS):** also known as Degradation Category Rating (DCR), consists of showing to the subjects two stimuli, and asking them to rate the amount of impairment between the two. The position of the original stimulus is always disclosed to the subjects. The grading scale is (1) very annoying, (2) annoying, (3) slightly annoying, (4) perceptible but not annoying, and (5) imperceptible.

The main disadvantage of this method is that, for the same amount of time, the DSIS produces fewer subjective scores compared to the ACR, as the subjects are requested to observe two stimuli at the time rather than one. However, it has the advantage that the subjective scores are in general not influenced by the subjects' opinion on the content and an easier detectability of the color impairment between two images.

The DSIS is widely used in the field of subjective quality assessment of compressed images, for example in \[8\], where the authors compared different learning-based compression methods.

**Double-Stimulus Continuous Quality-Scale (DSCQS):** the subjects are asked to rate the overall quality of both displayed stimuli using a continuous quality rating scale. In this method, the reference stimulus is displayed in a random position, unknown to the subject. As the subjects are asked to rate the quality of two stimuli at every step, this method is the slowest among those presented above. This method is particularly useful for evaluating learning-based compression methods where, at the highest bitrates, the compressed image might present a higher visual quality than the original thanks to pro-processing algorithms. As an example, the DSCQS method was utilized in the subjective visual quality assessment experiment conducted
to evaluate the submissions to the JPEG AI Call for Evidence, co-organized in conjunction with the IEEE MMSP’2020 Challenge on Learning-Based Image Coding, and presented in [9].

**Double Stimulus Comparison Scale (DSCS):** also known as pair comparison (PC), the subjects are asked to evaluate at each step the visual quality of the first stimulus compared to the second taken as reference. The grading scale is discrete, and the grades are (-3) much worse, (-2) worse, (-1) slightly worse, (0) the same, (1) slightly better, (2) better, (3) much better.

This experiment has the largest number of steps, and therefore is the most time-consuming. While this method is the most accurate in evaluating the performance of different compression methods in terms of quality, it has the disadvantage that the bitrates of the compared stimuli are required to be as close as possible, in order to guarantee a fair comparison. Other variants of scaling in this method could include three (Better, The same, Worse) or even only two (Better, Worse). Partial comparison among stimuli is also possible in order to reduce the duration of the tests.

A general summary of the methods introduced in this section is available in Table 1, presenting the main advantages and disadvantages of each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Scale Type</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR</td>
<td>SS</td>
<td>Discrete</td>
<td>Fast and simple</td>
<td>Scores influenced by the subjects' opinion on the content</td>
</tr>
<tr>
<td>ACR-HR</td>
<td>SS</td>
<td>Discrete</td>
<td>Allows to remove the variance due to the subjects' opinion on the content</td>
<td>Requires a long training procedure to acquaint the subjects with the artifacts</td>
</tr>
<tr>
<td>SSCQE</td>
<td>SS</td>
<td>Continuous</td>
<td>Comparable to the continuous grading scale of objective quality metric</td>
<td>Requires a long training procedure to acquaint the subjects with the artifacts</td>
</tr>
<tr>
<td>DSIS</td>
<td>DS</td>
<td>Discrete</td>
<td>Not influenced by the subjects' opinion on the content, reliable in evaluating color impairment</td>
<td>Slower than ACR</td>
</tr>
<tr>
<td>DSCQS</td>
<td>DS</td>
<td>Continuous</td>
<td>Both the original and impaired stimuli are graded</td>
<td>Slower than DSIS</td>
</tr>
<tr>
<td>DSCS</td>
<td>DS</td>
<td>Discrete</td>
<td>Compares all the different stimuli among themselves</td>
<td>Biggest number of comparisons, bitrate matching is critical</td>
</tr>
</tbody>
</table>

Table 1: Summary of the main methods for subjective quality assessment presented in ITU-R BT.500.
2.1.2 Review of ITU-R BT.1082 methods

Additional subjective methods for visual quality assessment of compressed images were proposed in ITU-R Recommendation BT.1082 [5]. Please note that some of the methods proposed in this document are still under review, and are not reported in this document.

**Ratio Scaling:** this method consists in presenting a sequence of images to the observers, asking to numerically grade each image based on the previous one. The numerical scale can be freely decided by the subjects and with no limit to the range. Each stimulus is presented twice. The advantage of this method is that it allows not only to determine the ranking of the compressed stimuli but also "how much better" is one stimulus compared to a different one.

**Graphing Scaling:** uses a grading scale based on adjectives and adverbs, that can be translated into the native language of the subjects (for example, "not quite passable" may be used instead of "poor"). This method has been reported simple and easy to use. In the reported example, the images were printed, no time limit was imposed, and the subjects were free to see the previous and next images.

**Numerical Category Scale:** subjects are asked to grade the images on a linear grading scale, usually with 5 to 10 points. The main advantage is its ease of use, as the subjects might (depending on the country of origin) be familiar with the grading scale from the school system.

**Multi-dimensional scaling:** aims at creating a three-dimensional perceptual space from the results, obtained from stimulus-comparison judgments.

2.2 JPEG AIC Part-2

The mentioned subjective visual quality assessment methods and standards have been mainly designed for web-quality applications, i.e. applications with a limited or variable bitrate requirement, where visual artifacts are usually easily perceivable by the human eye. In recent years, the number of applications that target high visual qualities is increasing [10]. In the context of storage applications, in fact, limited memory consumption is no longer the main requirement, thanks to cheap and large portable storage devices or cloud storage services. Therefore, users demand compression methods that maximize the visual quality of the images, rather than minimizing their bitrate consumption. In this context, the standard single and double stimulus experiments in Section 2.1 are not accurate enough, as it is difficult to assess subtle artifacts with these types of methods. As an example, it is usually difficult to detect slight shifts in the colors when images are presented alone or even side by side.

To address this issue, the JPEG committee has released the standard ISO/IEC 29170-2 (AIC Part-2) [11,12], which includes standardized methods for subjective visual quality assessment of nearly lossless to lossless visual qualities. In particular, two different methods have been proposed in Annex A and B of the standard.
**AIC Part-2 Annex A**: two distorted test images, along with the original image, are presented to the subjects, which are asked to select the closest distorted test image to the original one. The subjects have 4 seconds to select their preference.

**AIC Part-2 Annex B, or “flicker” test**: one distorted test image along with the original image are presented to subjects, side-by-side. The test image is interleaved at a certain frequency with the original image. In the case in which the test image presents some perceivable degradations, the test stimulus will appear "flickering". The position of the original and "flickering" stimuli is random and unknown to the evaluating subjects. The subjects are asked to choose which of the two stimuli presents flickering. If the image presents visible distortions, the flickering will be visible by the human eye and the subjects can detect the flickering image correctly. If the distortions in the test images are not perceivable by the human eye, the subjects are not able to correctly detect the flickering image and will answer randomly.

![Visual representation of the implementation of the method presented in AIC Part-2 Annex B, or “flicker” test. Image from [12].](image)

The method presented in AIC Part-2 Annex B was used for the subjective visual quality assessment of JPEG XS standard. The results can be found in [13].

### 3. BEST PRACTICES REVIEW

#### 3.1 CONTROLLED LAB ENVIRONMENT

The quality of compressed images is usually assessed in controlled environments, where the experiment is conducted in a test laboratory with controlled conditions. It is important to create ideal conditions to avoid noise and bias that can cause fluctuations in the subjective scores. The International Telecommunication Union (ITU) presented multiple public recommendations documents, namely Recommendation ITU-R P.910 [14], Recommendation ITU-R P.913 [15],
and Recommendation ITU-R BT.500 [4], which provide guidelines for the set-up of the laboratory environment for visual assessment of image quality.

The best practices for controlled environment subjective quality assessment were also reviewed by JPEG in JPEG AIC Part-1 [3]:

1. **Observers or subjects selection**: the experiment might be conducted with *experts*, i.e. people who are already familiar with the type of image artifacts assessed in the experiment, or by *non-experts/naïve* subjects, i.e. people that have no previous experience in the type of artifacts or image compression in general.

2. **Visual acuity**: subjects should have a normal or corrected-to-normal visual acuity, which may be measured with the Snellen or Ishihara test.

3. **Number of observers or subjects**: the recommended number of viewers is greater or equal to 15. It anyway depends upon the sensitivity and reliability of the adopted test procedure. Particular care should be devoted to the selection of the subjects, as studies on the consistency between results collected from different testing laboratories have found that systematic differences can occur. In addition, as the experiment results are usually correlated to the visual acuity, the standard recommends favoring subjects in the range of 18 to 30 years old.

4. **Training session**: before the session, all the subjects should be introduced to the experiment and familiarized with the objective of the experiment, grading scale, and timing.

In document ITU-R BT.500 [4], more guidelines are given on the selection of the monitor and room conditions, in particular:

5. **Room illumination**: should be low, for example around 10 lux.

6. **Monitor choice and setup**: the choice of the monitor is crucial as using monitors with different characteristics will yield different subjective scores.
   a. When using consumer displays for subject image assessments, all image processing options must be disabled.
   b. The peak monitor luminance should be 70-250 cd/m^2.
   c. The monitor contrast ratio should be ≤ 0.02.
   d. The maximum and minimum resolution of the monitor (center and corners of the screen) at the used luminance value should always be reported.
   e. Ratio of background luminance behind picture monitor to peak luminance of picture should be 0.15.

7. **Observation angle**: the maximum observation angle relative to the normal should be constrained so that deviations in reproduced color on the screen should not be visible to an observer.

8. **Selection of material**: the test images should be chosen carefully and should represent a large variety of use cases. The selected material should be presented to the subjects in random order and without showing the same content two times in a raw.

For high-quality consumer-grade image coding, the selection of the proper monitor parameters, like the monitor luminance, will play a crucial role: a slight increase in monitor luminance can in fact make the otherwise invisible artifacts visible. This makes the situation
much more challenging in uncontrolled test environments and thus requires special attention, together with viewing distance. This effect has been observed in previous studies, e.g., in JPEG XS-related subjective experiments while using AIC Part-2 flicker test.

A crucial element in all experiments involving human subjects is data protection and privacy respect. Although, the exact details often depend on the legislation and best practices of the country or organization where the experiments are carried out, anonymization of subjects and conditional access to the collected data, including various restrictions on the type of personal data that can be collected, are common elements among many and should be strictly adhered to when running experiments.

### 3.2 CROWDSOURCING ENVIRONMENT

Subjective image quality assessment can alternatively be conducted crowdsourcing by hiring subjects remotely to conduct the assessment directly in their environment. While the experiment is performed in an uncontrolled environment, it is more likely to be similar to the day-to-day viewing conditions of digital media in a realistic setup.

This approach has been adopted since the first decade of 2000, e.g. in the experiment presented in 2009 in [16]. However, in recent years, due to the worldwide confinements caused by the COVID-19 pandemic, this approach has shown increasing interest and popularity. The European Network on Quality of Experience in Multimedia Systems and Services (Qualinet)¹, has worked towards the definition of a number of guidelines for crowdsourcing subjective image quality assessment. Specifically, in 2014, the Qualinet Task Force on Crowdsourcing produced a whitepaper on the best practices and recommendations for Crowdsourced QoE [17]. More recently, ITU-T SG12 works toward a set of recommendations for subjective quality assessment based on crowdsourcing.

In the following, the main recommendations proposed in [17] are summarized:

- **Adopt user-friendly** software, without requiring admin installations. As an example, web-based applications are commonly used, where the subjects only need to connect to a web server without installing any software on their machine.
- **As crowdsourcing experiments** are more diverse in terms of spoken languages and cultural backgrounds for their nature, it is essential to use **simple and direct questions** to minimize misunderstandings among the subjects attending the experiment.
- **It is important to choose a proper duration** for the experiment to avoid fatigue in the subjects, and especially it is recommended to have shorter sessions than those in controlled environments.
- **The reward** should be proportional to the experiment time, to encourage more participants to take part in the evaluation session. Participation of subjects with a direct relationship with the organizer of the experiment should be avoided.
- **As it is not possible to get immediate feedback from the subjects**, it is important to provide a suitable and exhaustive **training session** to avoid poor quality of the

¹ [http://www.qualinet.eu](http://www.qualinet.eu)
collected scores caused by misunderstandings. It is also useful to address well-known issues that have been experienced in previous controlled or uncontrolled environment experiments.

- It is advisable to collect feedback from the subjects, to improve the experiment or to correct common issues identified among participants.
- Collecting event logging, and therefore information about what occurs during the experiment is critical to evaluate the quality of the submitted subjective scores. For example, the clicking behavior, the window resizing operations, the page reloading, and the tab switching, are all important factors to understand the behavior of the subject during the experiment.
- Including honeypot questions is also advisable to evaluate the amount of attention that the subjects are paying to the experiment. As an example, the semantic content of the previous image or simple general questions (e.g. "Five plus 2 = ?") can be asked during the experiment. In this way, it is possible to detect the subjects who have not focused adequately on the experiment, or are answering randomly. The reliability questions can be asked during or after the experiment.

Most of the crowdsourcing approaches are web-based frameworks that can be run through a widely used browser, and that therefore do not require the installation of any software. [18] presented different crowdsourcing frameworks for subjective quality assessment, summarizing the advantages and disadvantages of each. Among the most common ones are Euphoria [16], Crowd MOS [19], QualityCrowd 2 [20], WESP [21], BeagleJS [22], in-momento [23] and Crowdee [24].

Multiple subjective quality assessment experiments based on crowdsourcing have been conducted and reported over the years. Among the most recent experiments, [25] presented the results of a crowdsourcing-based subjective experiment conducted to assess the performance of the submissions to the JPEG AI Call for Evidence, co-organized with the IEEE MMSP’2020 Challenge on Learning-Based Image Coding. The experiment was conducted on Amazon Mechanical Turk by means of the QualityCrowd2 framework [20], using the DSCQS protocol. Figure 2 shows the interface that was displayed to the subjects during the experiments.
Table 2 summarizes the main characteristics of the controlled environment and crowdsourcing subjective quality assessment.

<table>
<thead>
<tr>
<th>Controlled environment</th>
<th>Crowdsourcing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expensive and time consuming</td>
<td>Fast and cheaper</td>
</tr>
<tr>
<td>It is possible to test one variable at the time</td>
<td>It is conducted in an uncontrolled but realistic environment</td>
</tr>
<tr>
<td>Low diversity in the participants</td>
<td>High diversity in the participants</td>
</tr>
<tr>
<td>Relevant for professionals who work in a controlled environment</td>
<td>Relevant for media broadcasting applications</td>
</tr>
</tbody>
</table>

Table 2: Summary of the main characteristics of the controlled environment subjective quality assessment method vs the crowdsourcing subjective quality assessment method

4. UNCOVERED QUALITY RANGES

While a large number of subjective quality assessment methods, summarized in Section 2, have been proposed and standardized, a few quality ranges have been identified for which a suitable subjective quality assessment method have not been identified yet.
4.1 High quality to near-visually lossless quality assessment

The subjective methods proposed in BT.500 have been observed to be more suitable for evaluating the visual appeal of images (how obvious and/or annoying the artifacts are) than for evaluating their visual fidelity (how true the images are to the original), and have been observed to be inaccurate when used for assessing modern compression methods (e.g. JPEG XL and JPEG XS). On the other hand, the methods proposed in JPEG AIC-2, are based on a very sensitive flicker test that will catch even the slightest visual distortion. Therefore, no suitable subjective image quality assessment method has been observed for qualities that a range from high quality to near-visually lossless quality: this is a range of visual qualities where artifacts are not noticeable by an average non-expert viewer without presenting an original reference image, but are detectable by a flicker test.
REFERENCES


