ISO/IEC JTC 1/SC 29/WG1 **N100118**
94th meeting – Online – January 17-21, 2022

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE: Use Cases and Requirements for Image Quality Assessment 1.0

SOURCE: WG1

EDITORS: Jon Sneyers (jon@cloudinary.com)
Michela Testolina (michela.testolina@epfl.ch)
Evgeniy Upenik (evgeniy.upenik@huawei.com)
Thomas Richter (thomas.richter@iis.fraunhofer.de)

PROJECT: JPEG AIC

STATUS: Approved

REQUESTED ACTION: Distribution

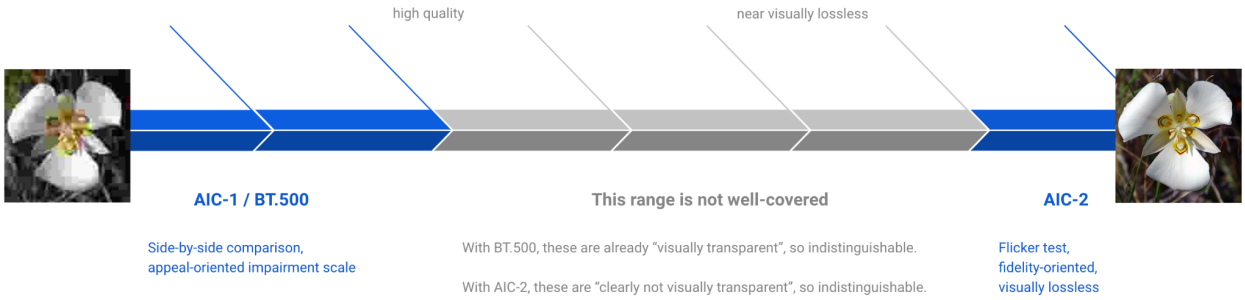
DISTRIBUTION: Public

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

Use Cases and Requirements for Image Quality Assessment 1.0

1. Introduction

The JPEG Committee has launched a new activity on Assessment of Image Coding, also referred to as JPEG AIC. This activity is a continuation of the previous standardization efforts (AIC-1 and AIC-2) and aims to develop a new standard, known as AIC-3. The new standard will be focusing on the methodologies for assessment of images with the quality levels in between the range where ITU-Rec. BT.500 [1] is suitable and the range where AIC-2 [2] is suitable.



Thanks to the increasing storage capabilities and internet speed, modern image codecs have made very high quality images possible and desirable. However, the existing standardized assessment methodologies, e.g. BT.500 [1], that are described in AIC-1 [3], are 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).

This problem has been exacerbated by appeal-oriented improvements in modern image codecs—i.e., codecs have become good at "hiding artifacts", causing BT.500 evaluation methodologies to saturate at a lower bitrate but also at a lower fidelity level. In a side-by-side comparison, images can become indistinguishable from the original at a bitrate as low as 0.5 bpp. However, the visual fidelity of those images can be relatively low at that bitrate. Thus, it is not necessarily the case that the tested image is also visually indistinguishable from the original image in a flicker test.

For example, a subtle tone shift of the background color can be impossible to see in a side-by-side comparison, but in a web use case it can lead to a visible discontinuity at the border where the image is supposed to match the background color of the surrounding page or of an adjacent image. Another example is loss of detail in subtle textures, which can be hard to spot in a side-

by-side comparison but which is still problematic for an e-commerce website selling clothes online where it is important that prospective customers can see the details of the fabric with a fidelity as close as possible to the experience of a physical store.

The methodology described in AIC-2, on the other hand, approaches the problem from the other end: it is based on a very sensitive flicker test that will catch even the slightest visual distortion. While this leads to extremely high fidelity, the bitrates needed to 'pass' such a test tend to be as high as 5 bpp or more. For somewhat lower fidelity targets and more economical bitrates, the AIC-2 test cannot be used.

2. Scope

The scope of AIC activities is to specify standards or best practices w.r.t. subjective and objective image quality assessment methodologies that target a range from high quality to near-visually lossless quality.

High quality, here, is defined as the lowest visual quality level where artifacts are not noticeable by an average non-expert viewer in a side-by-side comparison as in BT.500 [1].
Near-visually lossless quality, here, is defined as the smallest amount of artifacts where a flicker test [2] consistently detects flicker.

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.

The standard shall be aimed at evaluating the detectability of specifically compression artifacts, not necessarily other kinds of distortions like capture/sensor artifacts.

3. Use Cases

In this section, the use cases addressed by the JPEG AIC-3 standard are presented.

3.1 Assessment of Existing and Emerging Image Coding Technologies

Developing new and improving existing image coding technologies require perceptual visual quality assessment at many steps of the pipeline.
Given a set of different coding technologies that claim to perform well at the particular quality targets, proper tools are needed in order to assess and validate their performance for relevant applications.

3.1.1 Assessment of High-Quality Consumer-Grade Image Coding

For consumer-grade photography, mathematically lossless or near-lossless encoding is typically not required nor desirable, but the visual fidelity target for lossy encoding is higher than for web delivery.

3.1.2 Assessment of End-User Web Delivery Image Coding

On the web, visual fidelity targets are relatively low, in order to maximally save bandwidth. Both high-appeal encoding (low bitrate, but “the image looks good” in a no-reference way) and high-fidelity encoding (higher bitrate, accurate color and detail preservation is important) are relevant. Various image codecs are now (becoming) available in browsers, including JPEG-1, PNG, JPEG 2000, WebP, AVIF and JPEG XL. There is still limited knowledge about the compression performance of the different codecs and their encoders (e.g. mozjpeg, various AVIF encoders like libaom, rav1e, svt-av1 and aurora), as well as the type of image content they are most suitable for. Subjective evaluation is needed to get more insights. Objective metrics are valuable to make automated per-image encoder configuration decisions.

3.2 Encoder Perceptual Optimization

In order to improve the trade-offs between perceptual quality, compression density and encoding speed, various approaches can be proposed, e.g., to make better use of coding tools or to prune the search space. Such proposals require validation, in the form of subjective and/or objective evaluation. Encoders can also internally implement objective metrics to optimize for. In this use case, objective metrics are generally preferred in order to allow for faster iterations of development and finetuning of algorithm parameters. The computational complexity of the metric itself is important if the metric is used internally by an encoder.

3.3 Quality Assessment under Various Viewing Conditions

Subjective image quality depends on the viewing conditions, including display device technology, brightness, ambient light, pixel density, viewing distance, etc. Besides the traditional desktop or laptop computer monitor, images are also displayed using a variety of different display devices: including smartphones, smart watches, HDR televisions, beamers, etc. In many practical use cases, what matters is the way images are realistically and commonly viewed on consumer-grade display devices. This could be rather different from the way they are viewed on professional-grade calibrated display devices in a lab-controlled environment as specified in currently adopted standards.

4. Requirements

This project consists of different components that each have their own requirements. Essential “core requirements” are identified with a ‘shall’ and “complementary requirements” are identified with a ‘should’.

4.1 Subjective Quality Assessment and Score Screening

The proposed methodology shall satisfy the following core requirements:

4.1.1. **Quality range:** The standard shall provide discriminative scores in the quality range where both ITU Rec. BT-500 and AIC-2 (flicker test) are not discriminative between two candidate coding technologies;

4.1.2. **Suitable to evaluate compression artifacts**: The standard shall be suitable for evaluating quality degradations introduced by compression;

4.1.3. **Score screening**: The standard shall specify effective score screening methods in order to identify outliers;

4.1.4. **Reliability:** The standard shall ensure that after screening, scores are reliable and consistent (non-self-contradictory);

4.1.5. **Reproducibility:** The standard shall ensure that the scores obtained from one experiment shall be reproducible and confirmable by another experiment that follows the same methodology but performed independently from the first one;

4.1.6. **Scalability**: The standard shall provide a possibility to be efficiently usable at a large scale (hundreds or thousands of test subjects);

4.1.7. **Controllability**: The standard shall provide a possibility to be usable in scenarios with limited control over the experimental setup (e.g., crowdsourcing experiments);

4.1.8. **Variety of image content**: The standard shall be reliable when applied to a variety of image content, i.e. not only photographic images but also synthetic images (graphics, screenshots, etc.);

4.1.9. **Variety of consumption environments**: The standard shall be applicable in different consumption environments, i.e. display devices (TV, PC, Handheld) and viewing conditions;

4.1.10. **Portability**: The standard shall offer multiple viewing setups covering a set of different relevant viewing conditions (TVs, desktops, laptops, smartphones etc.);

4.1.11. **Flexibility**: The standard shall be able to obtain multiple trade-offs between competing requirements;

4.1.12. **Machine-readable scores**: The standard shall specify a format that provides a machine-readable representation of subjective scores in order to facilitate the validation, development, and training of objective quality metrics, and the aggregation and consolidation across multiple experiments.

Furthermore, the following requirements are complementary:

4.1.13. **Suitable for wide-gamut and high-dynamic-range images**: The standard should also be applicable to wide-gamut and high-dynamic-range images;

4.1.14. **Suitable for images with an alpha-transparency component**: The standard should also be applicable to images that contain an alpha-transparency component;

4.1.15. **Suitable to evaluate artifacts other than compression**: The standard should also be able to detect other kinds of distortions, e.g. capture/sensor artifacts;

4.1.16. **Economic feasibility**: The standard should make it possible to obtain results within a reasonable financial budget;

4.1.17. **Timeliness**: The standard should make it possible to conduct an experiment within a reasonable timeframe;

4.1.18. **Aggregation**: The standard should under specific conditions allow data from multiple experiments to be combined in order to widen the scope, improve the accuracy, and track evolutions (e.g. comparing new encoder versions to older encoder versions).

4.2 Objective Quality Assessment

The proposed objective metrics shall satisfy the following core requirements:

4.2.1. **Correlation with subjective scores**: The metrics shall correlate well with subjective scores;

4.2.2. **Reference-based metrics**: The metrics shall assess the perceptual visual quality of an image under test by comparing it to an original undistorted image;

4.2.3. **Variety of image content**: The standard shall be reliable when applied to a variety of image content, i.e. not only photographic images but also synthetic images (graphics, screenshots, etc.);

4.2.4. **Computational resources**: The processing time, memory, and hardware requirements of the metric shall be reasonable.

Complementary requirements:

4.2.5. **No-Reference metrics:** The metrics should assess the perceptual visual quality of an image under test without a need to compare it to an original undistorted image;

4.2.6. **Reduced-Reference metrics**: The metrics should assess the perceptual visual quality of an image under test without a need to explicitly compare it to an original undistorted image, but by using reduced side-information about the original undistorted image;

4.2.7. **Suitable for wide-gamut and high-dynamic-range images**: The standard should also be applicable to wide-gamut and high-dynamic-range images;

4.2.8. **Suitable for images with an alpha-transparency component**: The standard should also be applicable to images that contain an alpha-transparency component;

4.2.9. **Computational resources**: The processing time, memory, and hardware requirements of the metric should be suitable for an image coding pipeline.

4.3 Interchange Format for Objective and Subjective Scores

Core requirements:

4.3.1 **File Format**: The standard shall specify a format for objective and subjective scores allowing interchange for independent analysis by different parties;

4.3.2 **Machine readability**: The format shall use a syntax that is easy to parse by existing tools.

5. Free and Open Source Encouragement

Contributors are welcome and encouraged to develop and provide free and open source implementations of objective metrics as well as tools (frontend, backend, test set generation, etc.) to conduct evaluation experiments and data analysis specified in the standard.

6. Royalty-free Goal

The royalty-free patent licensing commitments made by contributors to previous standards, e.g. JPEG 2000 Part 1 and JPEG XL, have arguably been instrumental to their success. JPEG expects that similar commitments would be helpful for the adoption of new standards.

Bibliography

[1] Recommendation ITU-T BT.500-14, “Methodologies for the subjective assessment of the quality of television images,” International Telecommunication Union (2019).

[2] ISO/IEC 29170-2:2015 Information technology — Advanced image coding and evaluation — Part 2: Evaluation procedure for nearly lossless coding.

[3] ISO/IEC TR 29170-1:2017 Information technology — Advanced image coding and evaluation — Part 1: Guidelines for image coding system evaluation.