INTERNATIONAL STANDARD

ISO/IEC 29170-2

First edition
2015-08-15

Information technology — Advanced image coding and evaluation —

Part 2:
Evaluation procedure for nearly lossless coding

Technologies de l'information — Codage d'image avancé et évaluation —

Partie 2: Mode opératoire d'évaluation pour codage presque sans perte

Reference number
ISO/IEC 29170-2:2015(E)

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT), see the following URL: Foreword — Supplementary information.

The committee responsible for this document is ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 29170 consists of the following parts, under the general title *Information technology — Advanced image coding and evaluation*:

— *Part 1: Guidelines for coding system evaluation (forthcoming)*

— *Part 2: Evaluation procedure for visually lossless coding*

Introduction

This International Standard normalizes a procedure to evaluate coding systems by subjective methods. The procedure is particularly useful for evaluating lightly compressed coding systems used, for instance, in display stream compression where a source compresses image data sent to a display. Examples of display streams include but are not limited to a wired link between a set-top box unit and a television or between a mobile host graphics processor and a display panel module in a mobile appliance. Viewers of these displays should be unaware that a coding system is employed in the device or system. A coding system will be considered visually lossless if the test results meet a pre-defined acceptable quality level demonstrated by the performance criteria described in this Specification under the viewing conditions specified and media sets provided.

Appliances that may require visually lossless performance for compressed display streams include: computer monitor displays, televisions, mobile phone and tablet displays. Data compression for these systems allows existing display links to carry more display data than is possible with uncompressed image streams or to reduce system power consumption or both.

The types of coding systems tested by this procedure may have the following properties:

a) The presence of a coding system should be undetectable to a user who is viewing the display.

b) The coding system operates in real-time, with negligible latency, low complexity hardware and minimal memory in both the encoder and the decoder.

This procedure builds on prior standardization and best practices embodied in ISO 3664, ISO 20462-2 and ISO/IEC TR 29170-1.

Information technology — Advanced image coding and evaluation —

Part 2: Evaluation procedure for nearly lossless coding

1 Scope

This Technical Specification normalizes evaluation and grading of a light coding system used for displays and display systems, but is independent of the display technology. This procedure measures whether an observer can distinguish between an uncompressed reference and the reconstructed image to a pre-determined, statistically meaningful level.

The procedure compares individual images with two possible forced choice comparison test methods. This procedure relies on subjective evaluation methods designed to discern image imperfections on electronic colour displays of any technology or size.

Image selections for testing a specific coding system has bearing on the results this procedure will yield, but specific images required for testing are not within scope, excluding an informative annex describing self-test certification. Image categories may vary between end-usage products. For example, content relevant to television manufacturers may or may not be relevant to computer display manufacturers. Due to the nature of this procedure as a visual psychophysical test, observer's age is considered a meaningful parameter of the results.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3664, *Graphic technology and photography — Viewing conditions*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

3.1
advance time
time a reference image or test image is displayed during an interleaved pair comparison test

3.2
algorithm
unique combination of test conditions that contribute to a unique test image, for example, the combination of coding an image with one compression level and one coding method represents coding with one algorithm

Note 1 to entry: Coding an image with a different compression level and the same coding method represents a second unique algorithm.

3.3
blank time
time between trials when the display shows no stimulus

3.4
block of trials
long experiment is logically divided into a series of trials

Note 1 to entry: There is an expectation that the observer iteratively completes each trial presented during a short block of time without stopping for breaks beyond those allocated by the software presentation timing.

3.5
control image
cropped test image coded to ensure defects are easily detectable in a paired comparison test

3.6
evaluator
expert in the field of image or compression artefact analysis or vision science, who prepares images, categorizes images or instructs the observer

3.7
expert observer
observer skilled in vision science or coding technology and image artefacts

3.8
image
still representation of pixels rendered by the display

3.9
just noticeable difference
stimulus difference that would lead to a 0,75 probability of correct responses in a two-alternative forced-choice task

3.10
non-expert observer
observer with no special skills or training in vision science or coding technology

3.11
observer
individual performing the subjective evaluation procedure by evaluating stimulus to make paired comparison task choices

3.12
picture element
pixel
smallest element that is capable of generating the full intended functionality (e.g. colour and gray scale) of the display

Note 1 to entry: In a multicolour display, the smallest addressable element capable of producing the full colour range or the smallest element that is capable of generating the full functionality of the display.

3.13
reference image
cropped original image displayed during a trial

3.14
session
description of the full time interval in which an observer participates in the experiment on a given day

Note 1 to entry: It can encompass multiple blocks of trials. A full experiment can be divided into multiple sessions.

3.15
test image
cropped version of a reconstructed image after coding displayed during a trial

3.16
trial
single unit of the experiment

Note 1 to entry: A set of images will be presented on the screen for a defined viewing time. The trial is complete only when the observer has entered a response.

3.17
viewing time
number of seconds an observer views paired stimuli. The viewing time has a maximum value in all situations

3.18
visually lossless
fully decoded image or multimedia sample presenting a display that is visually indistinguishable from the original uncompressed data over the same spatial area and when viewed under the same conditions

4 Symbols and abbreviated terms

AQL acceptable quality level

PPD pixels per degree

RGB red-green-blue

5 Test methods

5.1 Protocol selection

The evaluator shall select the most suitable test protocol from those listed in Table 1 and contained in the listed annex. The protocol selection depends on the media under test, end display usage and the target AQL.

Table 1 — Visually lossless testing protocol

	Test Protocol	Comment
Annex A	forced choice paradigm with a non-flickering image	sensitive testing, suitable for testing images compressed and reconstructed by light coding systems, but not as sensitive as interleaved testing
Annex B	forced choice paradigm with interleaved images	most sensitive testing, suitable for testing images compressed and reconstructed by light coding systems

5.2 Media selection

Media selection takes images or other media for display rendering relevant to the display type and user applications.

Evaluators shall follow media selection procedures in Annex C after selecting a test protocol.

5.3 Observers

5.3.1 Observer selection

The observers shall be selected from a general population that may include both experts and non-experts. The observers for the experiment shall not include evaluators who participated in the media selection for the experiment being conducted.

The observer population should include variations in gender, ethnicity and age. The experiment is visual in nature and age can strongly correlate to visual acuity, therefore, participant age for this procedure favours the age range for the observer from 18 to 30 years old.

An observer's age shall not exceed 40 years.

This procedure recommends evaluators recruit a suitable number of observers sufficient to include no less than 10 observers who pass visual acuity (see 5.3.2) and test reporting (see D.1.2) requirements of this Technical Specification.

5.3.2 Observer visual screening

The following selection criteria shall apply:

a) Observers may wear corrective lens, either glasses or contacts that shall not have multiple focal lengths, e.g. progressive, bifocal or trifocal corrective lens.

b) Observers shall demonstrate normal visual acuity verified by using a Snellen reading test chart where the observer reads at 20/20 from 50 cm.

c) Participants shall demonstrate normal colour vision verified by testing with Ishihara plates or equivalent.

Evaluators can refer to Annex E for tools that help assess an observer's visual acuity.

5.3.3 Instructions to observers

Evaluators shall provide equivalent instructions to all observers with the criteria listed, here:

a) Explain the use of the software to record image assessments.

b) Explain the forced choice task. Every trial requires a response, even a guess.

c) Explain what to do if a conscious mistake is made, such as a finger slip, or the observer looked away and did not see the stimuli. If a retry is permitted (see 5.6), explain the method to retry that trial.

d) Explain where to sit and how to arrange the chair and use of the chin and headrest for the proper viewing position and comfort.

5.3.4 Training session

The evaluator should use a training session consisting of one block of trials for observers new to software used in subjective image testing. A training session will:

a) Use the control images from the experiment as test images.

b) Do not report data from the training session in the data results.

c) Use the same viewing time limit as the experiment.

d) Prompt the observer when a correct or incorrect response is entered. If incorrect, continue by repeating the test image until a correct response is entered.

Observers who have participated in the experiment usually do not need to repeat a training session.

5.4 Viewing conditions

5.4.1 Lighting and display calibration

Viewing conditions shall be consistent with ISO 3664 viewing condition for images displayed on colour monitors. Exceptions include:

a) The luminance of the white displayed on the monitor shall be between 100 cd/m^2 and 140 cd/m^2.

b) Displays that do not contain calibration tables should be avoided, but if used should have gamma setting verified close to 2,2 and deviations noted in the test report. Record maximum luminance of the display used in the procedure.

c) The colour of surrounding walls and ceilings do not require a specific colour but shall not cause distracting reflections that may affect the vision of the observer. An appropriate viewing booth is optionally desirable.

d) The display monitor shall render at least the number of bits per component present in the tested images.

5.4.2 Viewing distance

Viewing distance shall be controlled and ensured using a chin and forehead rest for observers. Figure 1 shows an example chin and forehead rest. The rest may be height adjustable to ensure small differences between observer's head size can be accommodated, the chair for the observer should have a large adjustment range to ensure the observer is as comfortable as possible during the procedure.

Figure 1 — Example chin and forehead rest

The viewing distance is a function of the pixels per degree (PPD) to be subtended at the set viewing distance.

The viewing distance shall be determined according to Table 2 based on horizontal pixels, the display pixel density and pixels subtended in one degree of arc.

Table 2 — Viewing distance versus display size and resolution

Condition	PPD[a]	D[b] cm
Viewing distance	30	D equals the larger of the value in the following equation or 12cm[c] $D = \frac{W}{H_{RES} \times \tan(\frac{1}{PPD})}$

[a] The experiment requires a consistent display orientation to be maintained and mobile display may have a different width and pixel resolution in landscape versus portrait orientation. PPD is calculated for each orientation. Detailed work on computer displays and mobile devices tends to be closer than for general entertainment, e.g., television, and requires evaluation with a more aggressive PPD than would be the case for Snellen acuity (30 cycles/degree or PPD = 60)

[b] W is the screen width (cm) and H_{RES} is the number of pixels across the display horizontally as viewed by the observer.

[c] The minimum focusing distance for normal vision is predetermined as 12 cm by this Technical Specification.

5.4.3 Viewing position

An observer shall be seated in a comfortable position oriented with respect to the display as shown in Figure 2 and Figure 3:

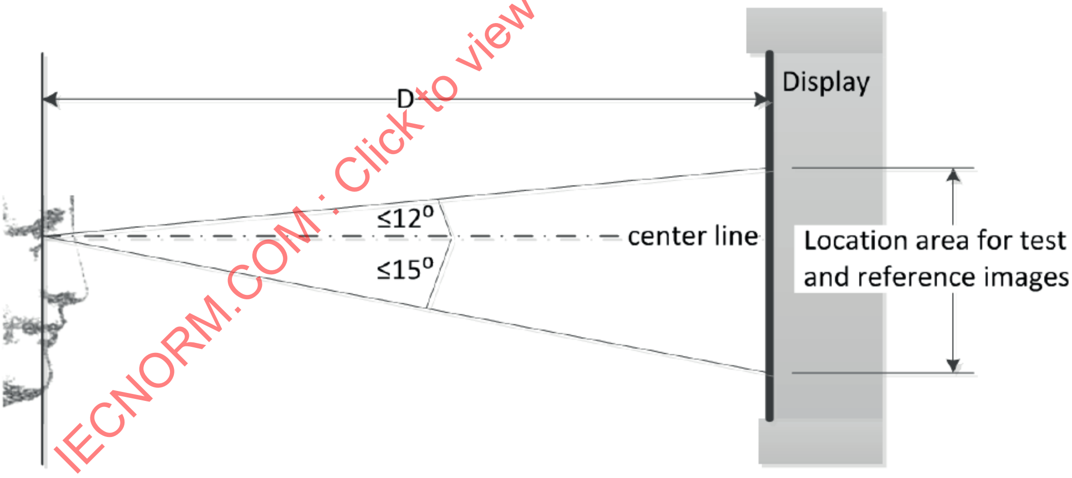


Figure 2 — Observer position with respect to the test display (side view)

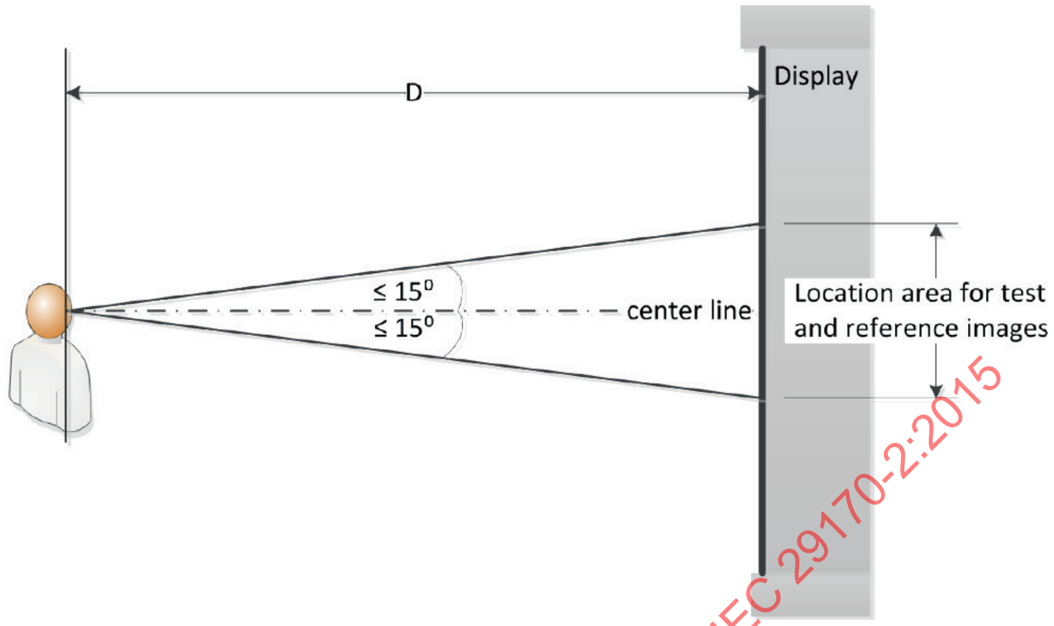


Figure 3 — Observer position with respect to the test display (top view)

5.5 Viewing time

An observer shall view paired stimuli on the display for no more than 4 s.

If no choice is made within the viewing time, the display shall blank out the stimuli and prompt the observer for a response. The blanking of the screen should use a dark, neutral colour with low luminance in order to not hinder viewing the next stimulus pair. User prompts from the screen should be viewable with mesopic vision, preferable using red lettering.

If the experiment permits retries (see 5.6) of a paired stimulus, the viewing time is reset for each viewing attempt.

5.6 Trial retry

An experiment may permit an observer to retry a trial. The evaluator provides a mechanism to retry paired stimuli when the observer makes a conscious mistake or does not see the trial stimuli. The retry data replaces the prior recorded response as if it did not occur. Evaluators should track the number of retries by each observer over the course of all sessions.

5.7 Test reporting

Evaluators shall report results of the data collected by the procedure as prescribed in Annex D.

5.8 Verification of procedure

Evaluators may self-certify a specific implementation of this procedure by correlating a trial to the results shown in Annex F.

Annex A
(normative)

Forced choice paradigm with non-flickering images test protocol

A.1 Image pair generation

A.1.1 On-screen presentation of images

Each test image shall be presented side-by-side with the reference image. These images shall be separated by a 1,0° ± 0,1° gap separating them based on the PPD. The side on which the test and reference images are presented shall be randomly assigned on each trial.

Centered and above these images a second copy of the reference image shall be presented. Images may be arranged in either portrait or landscape orientation.

All images shall be shown with 1:1 mapping to the display pixels without anti-aliasing.

In all trials, a test image is to be compared against its corresponding reference image.

Figure A.1 shows an example display using a cropped portion of *Boat* as a reference image on top and as a reference image as one of the choices on the bottom next to a test image. The observer prompt is below the images.

Figure A.1 — Example display showing image arrangements for the non-flickering test protocol

A.1.2 Image trial ordering

Two consecutive trials shall present a different test image, even if the reference image is the same. For example, the test image differs in either algorithm or compression level.

During a session, the presentation order of the set of test images shall be randomized.

A.1.3 Control Image

As part of each block of trials, at least 5 % of the test images should contain defects that a typical observer would be able to detect.

A.2 Test duration and timing

The time for an observer to complete one trial is the viewing time, t_{VIEW}, which depends on how fast the observer responds after seeing a stimulus pair plus a software blanking time, t_{BLANK}. These times are:

a) At least a four-second viewing time, t_{VIEW}, should be budgeted for each trial of the experiment.

b) The software shall ensure that there is a 0,25 s screen blank between subsequent trials.

If a block of trials contains a number of test images, N, coded with a number of algorithms, A, at a number of compression levels, C, and each combination is shown to the observer once, and the budgeted time to complete a block of trials is:

$$t_{BLOCK} = N \times A \times C \times (t_{VIEW} + t_{BLANK})$$

A block of trials should not require more than 10 min to complete, in order to minimize observer stress. This allows the observer to take regular breaks from the evaluation. If the block of trials would not exceed four to five minutes, it is acceptable to present each test image two or more times during a single block of trials.

Each session shall not exceed two hours in a one-day period.

The block of trials shall be repeated such that each observer in the experiment has evaluated each test image 30 times.

A.3 Subjective task

On each trial, the observer shall have a fixed duration of time to view the images appearing on the screen.

The observer compares the reference image (above) with the two choices (below).

The observer shall be instructed to select the image that most closely matches the reference image and shall make a computer response as soon as they have made a decision.

If the viewing time elapses, the screen shall be blanked and the experiment shall pause until the observer responds.

Annex B
(normative)

Forced choice paradigm with interleaved images test protocol

B.1 Image pair generation

B.1.1 On-screen presentation of images

Each compressed and reconstructed image (test image) shall be presented side-by-side with the uncompressed image (reference image). The test image shall be temporally interleaved with the reference image. These images shall be separated by a 1,0° ± 0,1° gap separating them based on the PPD. The side on which the test and reference images are presented is randomly assigned on each trial.

All images shall be shown with a 1:1 mapping to the display pixels to without anti-aliasing.

Figure B.1 shows an example display using a cropped portion of *Boat* as a reference image and a test image and the observer prompt below the images.

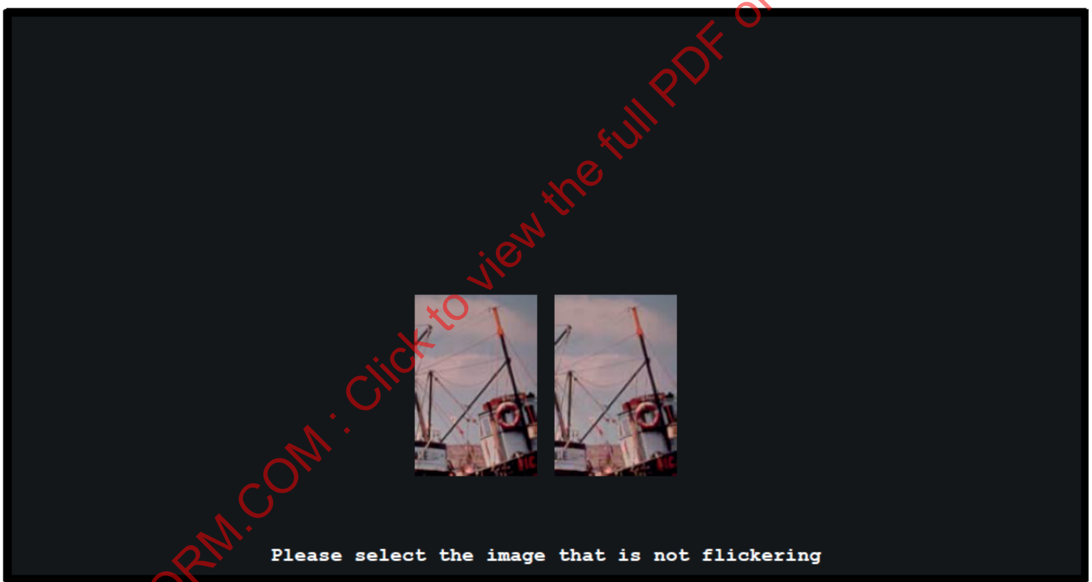


Figure B.1 — Example display with an interleaved test protocol image arrangement

B.1.2 Temporal interleaving

The test image and reference image shall be interleaved at the same rate for all experimental trials. Each image shall be displayed at an advance time specified in Table B.1.

Table B.1 — Advance time and frame rate

Advance time s	Frame rate Hz	Consecutive frames showing one image	Repetition rate Hz – informative only –
0,1	50	5	5,0
0,1	60 (recommended)	6	5,0

B.1.3 Image ordering

Two consecutive trials shall present a different test image, even if the reference image is the same. For example, the test image differs in either algorithm or compression level.

During a session, the presentation order of the set of test images shall be randomized.

B.1.4 Control image

As part of each block of trials, at least 5 % of the test images should contain defects that a typical observer would be able to detect.

B.2 Test duration

The time for an observer to complete one trial is the viewing time, t_{VIEW}, which depends on how fast the observer responds after seeing a stimulus pair plus a software blanking time, t_{BLANK}. These times are:

a) At least a 4 s viewing time, t_{VIEW}, should be budgeted for each trial of the experiment.

b) The software shall ensure that there is a 0,25 s screen blank between subsequent trials.

If a block of trials contains *N* test images coded with *A* number of algorithms, *C* number of compression levels and each combination shown to the observer once, the budgeted time to complete a block of trials is:

$$t_{BLOCK} = N \times A \times C \times (t_{VIEW} + t_{BLANK})$$

A block of trials should not require more than 10 min to complete, in order to minimize observer stress. This allows the observer to take regular breaks from the evaluation. If the block of trials would not exceed four to five minutes, it is acceptable to present each test image two or more times during a single block of trials.

Each session shall not exceed two hours in a one-day period.

The block of trials shall be repeated such that each observer in the experiment has evaluated each test image 30 times.

B.3 Subjective task

On each trial, the observer shall have a fixed duration of time to view the images appearing on the screen.

The observer examines both choices for flicker.

The observer shall be instructed to select the image that does not appear to flicker and make a computer response as soon as they have made a decision.

If the viewing time elapses, the screen shall be blanked and the experiment will pause until the observer responds.

Annex C
(normative)

Media selection and preparation procedure

C.1 Image search and cropping procedure

C.1.1 Image processing

The process of determining the images to be used in the experiment is performed by the evaluator. The process involves the steps from Figure C.1. The evaluator compresses the full image, and inspects the results for artefacts. The process of searching for artefacts can include:

a) Reference/Test image flicker

b) Close-up viewing/ on-screen magnification examination

c) Side-by-side image viewing

d) Full-reference image difference map utilities (e.g. s-CIELAB)

The goal of this search is to identify regions within images which exhibit a bit-wise difference in the compressed image and is thus potentially visually degraded. As part of the image selection, the evaluator may review many images and select only a small fraction of the worst examples of artefacts for inclusion in the experiment.

Once an image with an artefact is found, the evaluator crops a region centred about the artefact for inclusion in the test set. The cropping of the region of interest will greatly reduce the time needed for observers to view and image before making a response.

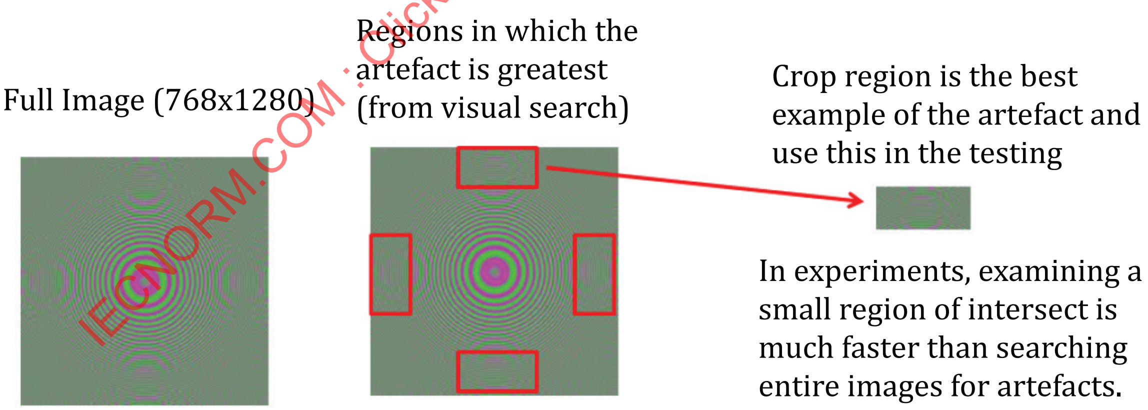


Figure C.1 — Test image processing pipeline

C.1.2 Image orientation

Table C.1 lists orientation options based on the test protocol selection.

Table C.1 — Reference and test image orientation

Test protocol	Landscape orientation	Portrait orientation
Annex A (side-by-side)	Allowed	Allowed
Annex B (interleaving)	Not allowed	Allowed

C.2 Diverse image categories

Image selection should have coverage of at least four critical categories, for example:

a) Natural images of people and animals

b) Natural images of scenery, landscapes and still life

c) Graphics and text that is common in computer and mobile device desktop screens, including an assortment of application icons mixed with text and different text fonts including Arabic, Asian, Cyrillic, Latin and other types of lettering

d) Challenge images of computer graphics test specific image attributes that will be difficult to compress. It is usually possible to create a test pattern that can cause any lossy coding system to fail. This category helps establishes a limit of the quality a candidate algorithm can achieve, even though some defects may be permissible.

The example set of images from these categories used in the reference experiment are shown in Figure C.2.

Figure C.2 — Image category examples, from left to right: Artificial/Challenge, Landscape/Still, Graphics/Text, and People/Animals

C.3 Controlling experiment duration

C.3.1 Limited trial time duration

Subjective evaluation is time consuming. It is important that the experimental sessions be limited in duration to no more than two hours to avoid participant fatigue.

C.3.2 Images to exclude

During the image selection process, some images that exhibit particularly strong artefacts are omitted. In these cases in which there is an obvious failure, it is acceptable to omit the test image from subjective testing as the result is pre-determined, and nothing will be learned from including it in the evaluation. These images should be degraded to the point at which any reasonable observer could be expected to detect the defect with high reliability.

Images that show defects without the presence of the reference image and without spatial dependence, that is, the defects are noticeable at greater than twice the tested PPD shall be excluded from testing in this procedure.

Another class of images to exclude is images that are bit-wise perfect. During image screening, the evaluator should only select crops in which there is an objective difference between the compressed and reference image.

C.3.3 Image to include

In addition to the set of images of interest, each block of trials contains control images. These images should be of the same content as that used in some of the test images, but be compressed to be unambiguously degraded. These control images serve two purposes:

a) Provide a metric in which the observer attentiveness can be measured post-hoc.

b) The observer is aware that there are at least some images that are degraded. Thus, the inclusion of control images reduces feelings of frustration or boredom.

Annex D
(normative)

Test report procedure

D.1 Data analysis calculation

D.1.1 Per-Observer data accumulation

For each observer, the data from all sessions and blocks of trials is collected and separated by test image.

For each test image, divide the number of trials in which the observer correctly selected the reference image by the total number of presentations of that test image.

Segregate the test images that were part of the control image set. Calculate the fraction correct across all the control images.

Note in the report if an observer is allowed to repeat a trial in case of a conscious mistake or did not see the stimuli.

D.1.2 Observer qualification for inclusion in the data set

Performance on the control set shall be greater than 95 % for the observer to qualify for inclusion in remaining analysis. This filter removes participants who may have been inattentive or misunderstood the instructions. It is also an approach to correct mistakes introduced during the observer screening process.

D.1.3 Aggregate statistics

Repeat analysis for each observer who participated in the study.

Upon completion of the analysis, there are a total of $B \times N$ response fractions corresponding to a total of N test images that were used in the testing and B observers. For each of the N test images, calculate the mean and standard deviation (1σ) of the response fractions for all qualifying observers.

Record the minimum and maximum response fraction for the group of observers for each of the N test images.

D.2 Data figures

The results are usually content dependent. It is most informative to compare algorithm and compression ratio performance directly for a single image at a time. One such example comparison is shown in Figure D-1. In this example, four algorithms were used in the testing with two or three different compression levels each. It is convenient to group the data from each algorithm together, for example as groups B, C, D and E along the abscissa of Figure D.1. The ordinate represents the fraction of the trials in which observers correctly select the reference image.

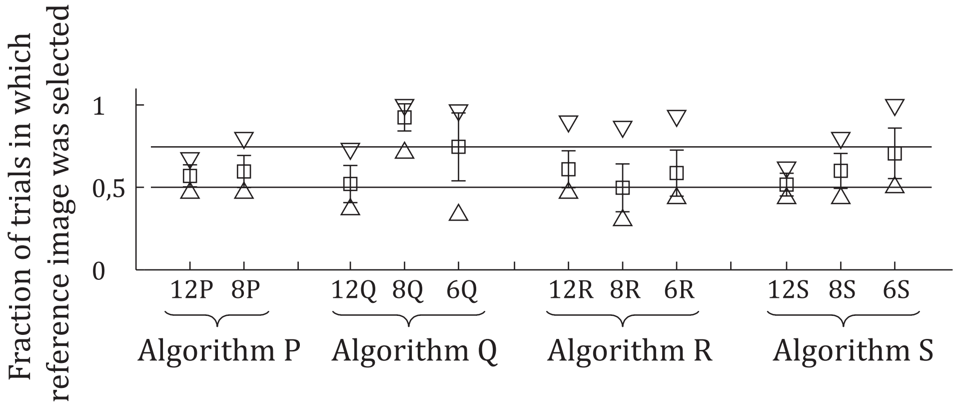
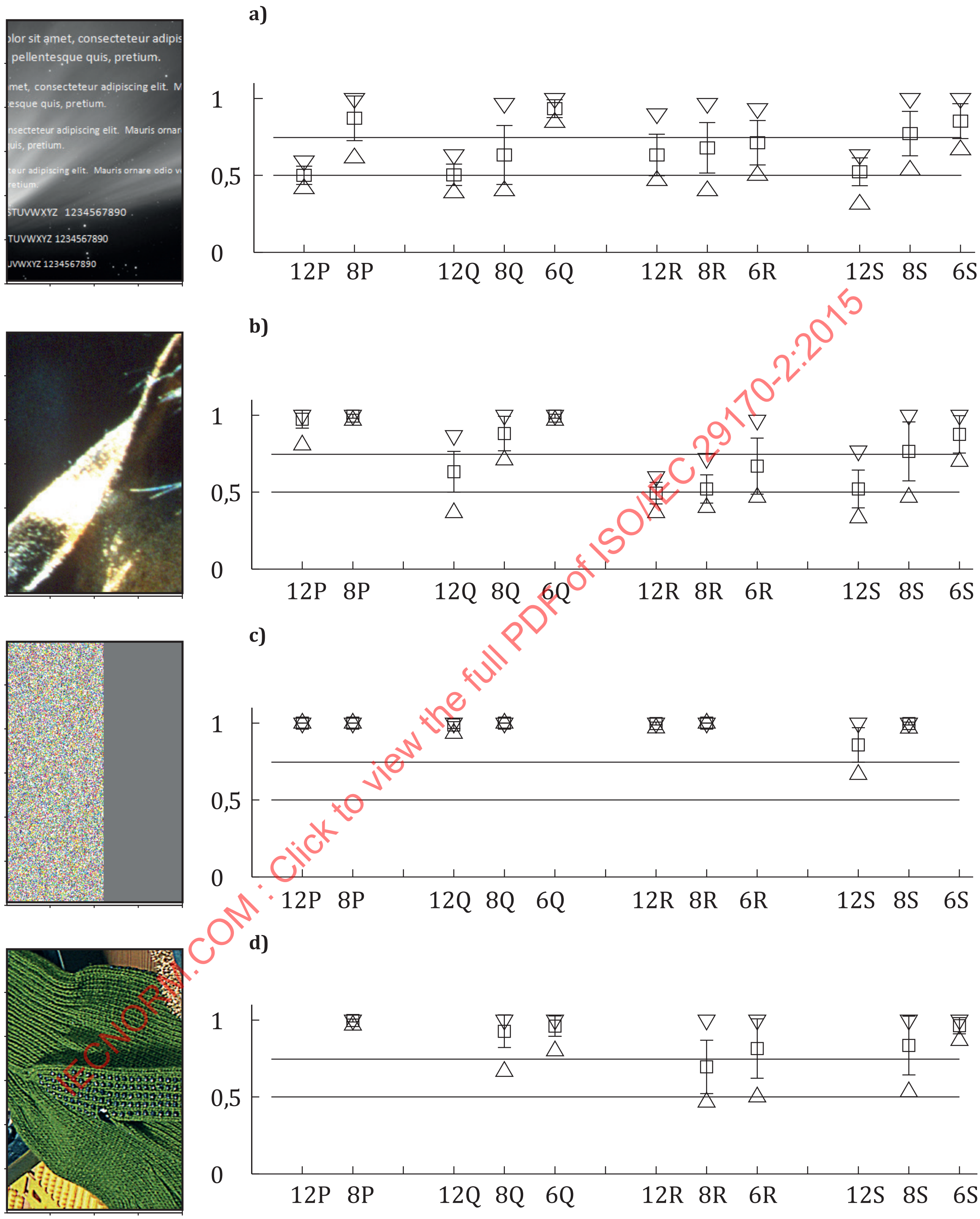


Figure D.1 — Example data presentation for one image

The two horizontal lines at heights of 0,5 and 0,75 represent the chance line and 1 JND, respectively. At 0,5, observers are responding at chance, there is no visible difference. At a response fraction of 0,75, there is 1 just noticeable difference of image degradation. This Specification suggests 1 JND as the threshold for classifying an image as visually lossless.

In Figure D.1 and D.2, the mean response fraction for all observers is plotted as a square, and the standard deviation of the response fraction is shown as the error bar. The highest reported response fraction and the lowest response fraction are indicated with triangles. Adding a thumbnail of the test image next to the data set adds reader convenience to the report. When selecting example images from each category for presenting the results, it is preferable to present the results for the image that had the largest failures.



Key

a) text and graphics — “cleartype”

b) people/animals — “deer”

c) challenge image — “noise boundary”

d) landscape/still life — “woven glove”

Figure D.2 — Example results

D.3 Data interpretation

D.3.1 Identifying visually lossy images

This procedure evaluates the assertion that an algorithm is visually lossless. An algorithm satisfies this criterion when all the observers fail to correctly identify the reference image more than 75 % of the trials. All data points that fall below 75 % in the results represent a visually lossless algorithm. However, see D.3.2 for assessment guidelines of difficult images, usually called engineered or challenge images.

A pre-specified AQL may stipulate that a visually lossless level differs from the above baseline assertion that all observers results fall below 1 JND. An AQL different from 1 JND shall be stipulated prior to testing.

In the text and graphics category of Figure D.2, the “Clear type” image is considered visually lossless only by algorithms 12P, 12Q and 12S. With these compression routines, no observers were able to discriminate the reference from the compressed image. Algorithm 12R does not completely satisfy the visually lossless assertion because at least one of the observers was able to reliably identify the reference and test images.

While it is possible that a visually lossless algorithm could fail the 75 % criteria through guessing, the likelihood of such a false failure (in which there is no visible difference) would occur is closely related to the number of repeated viewings. Table D.1 shows the relationship of the false positive rate to the number of trial repetitions. The table reports the probability that each response fraction could be caused through guessing. (Note that not all probabilities decrease monotonically as some of the response fractions are not feasible with the number of repetitions.) The procedure, with 30 trial repetitions per observer achieves 0,26 % false positive rate at 1 JND.

Table D.1 — Probability of a method failure by chance (false positive detection of defects)

Repetitions	**Response Fraction**								
	0,6	0,65	0,7	**0,75**	0,8	0,85	0,9	0,95	1
5	$5{,}0\times10^{-1}$	$1{,}9\times10^{-1}$	$1{,}9\times10^{-1}$	$\mathbf{1{,}9}\times10^{-1}$	$1{,}9\times10^{-1}$	$3{,}1\times10^{-2}$	$3{,}1\times10^{-2}$	$3{,}1\times10^{-2}$	$3{,}1\times10^{-2}$
10	$3{,}8\times10^{-1}$	$1{,}7\times10^{-1}$	$1{,}7\times10^{-1}$	$\mathbf{5{,}5}\times10^{-2}$	$5{,}5\times10^{-2}$	$1{,}1\times10^{-2}$	$1{,}1\times10^{-2}$	$9{,}8\times10^{-4}$	$9{,}8\times10^{-4}$
15	$3{,}0\times10^{-1}$	$1{,}5\times10^{-1}$	$5{,}9\times10^{-2}$	$\mathbf{1{,}8}\times10^{-2}$	$1{,}8\times10^{-2}$	$3{,}7\times10^{-3}$	$4{,}9\times10^{-4}$	$3{,}1\times10^{-5}$	$3{,}1\times10^{-5}$
20	$2{,}5\times10^{-1}$	$1{,}3\times10^{-1}$	$5{,}8\times10^{-2}$	$\mathbf{2{,}1}\times10^{-2}$	$5{,}9\times10^{-3}$	$1{,}3\times10^{-3}$	$2{,}0\times10^{-4}$	$2{,}0\times10^{-5}$	$9{,}5\times10^{-7}$
25	$2{,}1\times10^{-1}$	$5{,}4\times10^{-2}$	$2{,}2\times10^{-2}$	$\mathbf{7{,}3}\times10^{-3}$	$2{,}0\times10^{-3}$	$7{,}8\times10^{-5}$	$9{,}7\times10^{-6}$	$7{,}7\times10^{-7}$	$3{,}0\times10^{-8}$
30	$\mathbf{1{,}8}\times10^{-1}$	$\mathbf{4{,}9}\times10^{-2}$	$\mathbf{2{,}1}\times10^{-2}$	$\mathbf{2{,}6}\times10^{-3}$	$\mathbf{7{,}2}\times10^{-4}$	$\mathbf{3{,}0}\times10^{-5}$	$\mathbf{4{,}2}\times10^{-6}$	$\mathbf{2{,}9}\times10^{-8}$	$\mathbf{9{,}3}\times10^{-10}$
35	$1{,}6\times10^{-1}$	$4{,}5\times10^{-2}$	$8{,}3\times10^{-3}$	$\mathbf{9{,}4}\times10^{-4}$	$2{,}5\times10^{-4}$	$1{,}1\times10^{-5}$	$2{,}1\times10^{-7}$	$1{,}0\times10^{-9}$	$2{,}9\times10^{-11}$
40	$1{,}3\times10^{-1}$	$4{,}0\times10^{-2}$	$8{,}3\times10^{-3}$	$\mathbf{1{,}1}\times10^{-3}$	$9{,}1\times10^{-5}$	$4{,}2\times10^{-6}$	$9{,}3\times10^{-8}$	$7{,}5\times10^{-10}$	$9{,}1\times10^{-13}$
45	$1{,}2\times10^{-1}$	$1{,}8\times10^{-2}$	$3{,}3\times10^{-3}$	$\mathbf{4{,}1}\times10^{-4}$	$3{,}3\times10^{-5}$	$2{,}7\times10^{-7}$	$4{,}7\times10^{-9}$	$2{,}9\times10^{-11}$	$2{,}8\times10^{-14}$
50	$1{,}0\times10^{-1}$	$1{,}6\times10^{-2}$	$3{,}3\times10^{-3}$	$\mathbf{1{,}5}\times10^{-4}$	$1{,}2\times10^{-5}$	$1{,}0\times10^{-7}$	$2{,}1\times10^{-9}$	$1{,}1\times10^{-12}$	$8{,}9\times10^{-16}$

D.3.2 Assessment of challenge images

It is usually possible to engineer an image that is challenging to code. Subjective analysis can reveal a broad overall failure for these images for many or all of the algorithms. One example of such an image is the third row of Figure D.3 which depicts a “Noisy Boundary“. All algorithms failed by much greater than 1 JND. In this situation, the experimenter would need to review the test images for obvious failures, some of which are shown in Figure D.3. In the case of such engineered test images, there should be rules to determine which classes of failures may be acceptable. For example, edge bleed over and hue shift should always be unacceptable while fine line artefacts could be acceptable at typical viewing distances with a greater PPD than used by this procedure.

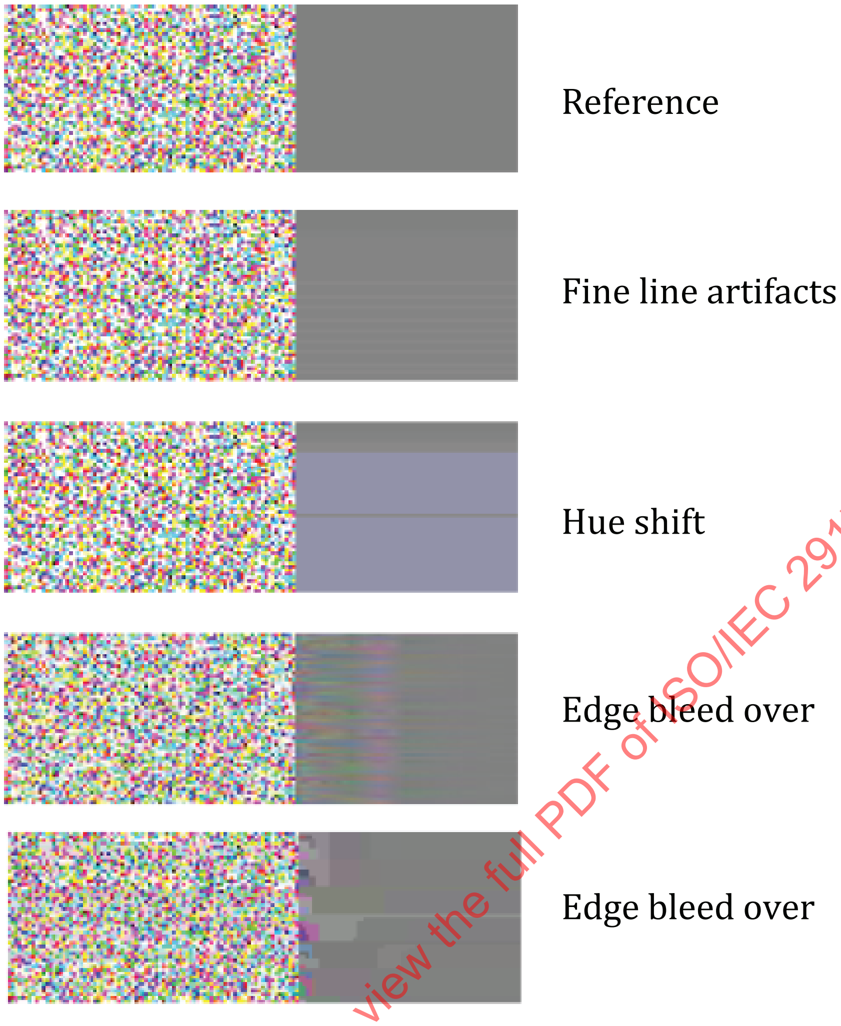


Figure D.3 — Type of obvious failures

D.4 Demographic statistics

The evaluator shall report the demographic breakdown of the observer population by:

a) session dates and times;

b) age, which may be reported as a number or reported by using the following brackets as appropriate to local cultural norms where the experiment is conducted:

  1) 21 or less;

  2) 22 to 26;

  3) 27 to 31;

  4) 32 to 36;

  5) 37 or greater;

c) Gender (optional recorded item);

d) use of corrective lenses or not;