INTERNATIONAL STANDARD

ISO 20462-3

First edition
2005-11-01

Photography — Psychophysical experimental methods for estimating image quality —

Part 3: Quality ruler method

Photographie — Méthodes psychophysiques expérimentales pour estimer la qualité d'image —

Partie 3: Méthode «quality ruler»

Reference number
ISO 20462-3:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

ISO 20462-3 was prepared by Technical Committee ISO/TC 42, *Photography*.

ISO 20462 consists of the following parts, under the general title *Photography — Psychophysical experimental methods for estimating image quality*:

— *Part 1: Overview of psychophysical elements*

— *Part 2: Triplet comparison method*

— *Part 3: Quality ruler method*

Introduction

There are many circumstances under which it is desirable to quantify image quality in a standardized fashion that facilitates interpretation of results within a given experiment and/or comparison of results between different experiments. Such information can be of value in assessing the performance of different capture or display devices, image processing algorithms, etc. under various conditions. However, the choice of the best psychometric method for a particular application may be difficult to make, and interpretation of the rating scales produced by the numerical analyses is frequently ambiguous. Furthermore, none of the commonly used rating techniques provides an efficient mechanism for calibration of the results against a standardised numerical scale or associated physical references, which is desirable when results of different experiments are to be compared or integrated.

The three parts of ISO 20462 address the need for documented means of determining image quality in a calibrated fashion. Part 1 provides an overview of practical psychophysics and aids in identifying the better choice between the two alternative approaches described in Part 2 (triplet comparison method[1][2]) and Part 3 (quality ruler method[3]). These two techniques are complementary and together are sufficient to span a wide range of practical applications. Parts 2 and 3 document both specific experimental methods and associated data reduction techniques. It is the intent of these methods to produce results that are not merely directional in nature, but are expressed in terms of relative or fixed scales that are calibrated in terms of just noticeable differences (JNDs), so that the significance of experimentally measured stimulus differences is readily ascertained.

The quality ruler method described in this part of ISO 20462 is particularly suitable for measuring quality differences exceeding one JND. The ratings given by an observer can be converted to JND values in real time, rather than having to wait until the entire experimental data set has been collected and analysed. Furthermore, with suitable reference stimuli, the quality ruler method permits the results to be reported using the standard quality scale (SQS), a fixed numerical scale that:

a) is anchored against physical standards;

b) has one unit corresponding to one JND; and

c) has a zero point corresponding to an image having little identifiable information content.

Reflection prints calibrated against the absolute SQS, which are referred to as standard reference stimuli (SRS), will be available on the I3A website. This part of ISO 20462 also describes how users can conveniently generate their own quality ruler images with correct relative calibrations and, if desired, calibrate them absolutely against the SRS.

The International Organization for Standardization (ISO) draws attention to the fact that it is claimed that compliance with this document may involve the use of US Patent Numbers 6,639,999 and 6,658,139 concerning the quality ruler given in Clauses 4 to 6.

ISO takes no position concerning the evidence, validity and scope of this patent right.

The holder of this patent right has assured ISO that he is willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with ISO. Patent inquiries may be addressed to:

General Council and Senior Vice President
Eastman Kodak Company
345 State Street
Rochester, NY 14650
USA

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified above. ISO shall not be held responsible for identifying any or all such patent rights.

Photography — Psychophysical experimental methods for estimating image quality —

Part 3: Quality ruler method

1 Scope

This part of ISO 20462 specifies:

a) the nature of a quality ruler;

b) hardcopy and softcopy implementations of quality rulers;

c) how quality rulers may be generated or obtained; and

d) the standard quality scale (SQS), a fixed numerical scale that may be measured using quality rulers.

2 Normative references

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

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

3 Terms and definitions

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

3.1
artefactual attribute
attribute of image quality that, when evident in an image, nearly always leads to a loss of overall image quality

EXAMPLE Examples of artefactual attributes include noise and aliasing.

NOTE The commonly used terms *defect* and *impairment* are similar in meaning.

3.2
attribute
aspect, dimension, or component of overall image quality

cf. **artefactual attribute** (3.1) and **preferential attribute** (3.10)

EXAMPLE Examples of image quality attributes include image structure properties such as sharpness and noise; colour and tone reproduction properties such as contrast, colour balance, and relative colourfulness; and digital artefacts such as aliasing, contouring, and compression defects.

3.3
image quality
impression of the overall merit or excellence of an image, as perceived by an observer neither associated with the act of photography nor closely involved with the subject matter depicted

NOTE The purpose of defining image quality in terms of third-party (uninvolved) observers is to eliminate sources of variability that arise from more idiosyncratic aspects of image perception and pertain to attributes outside the control of imaging system designers.

3.4
instructions
set of directions given to the observer for performing the psychophysical evaluation task

3.5
just noticeable difference
JND
stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task

cf. **quality JND** (3.12)

3.6
magnitude estimation method
psychophysical method involving the assignment of a numerical value to each test stimulus that is proportional to image quality; typically, a reference stimulus with an assigned numerical value is present to anchor the rating scale

NOTE The numerical scale resulting from a magnitude estimation experiment is usually assumed to constitute a ratio scale which, ideally, is a scale in which a constant percentage change in value corresponds with one JND. In practice, modest deviations from this behaviour occur, complicating the transformation of the rating scale into units of JNDs without inclusion of unidentified reference stimuli (having known quality) among the test stimuli.

3.7
multivariate
describing a series of test or reference stimuli that vary in multiple attributes of image quality

3.8
observer
individual performing the subjective evaluation task in a psychophysical method

3.9
paired comparison method
psychophysical method involving the choice of which of two simultaneously presented stimuli exhibits greater or lesser image quality or an attribute thereof, in accordance with a set of instructions given to the observer

NOTE Two limitations of the paired comparison method are as follows.

a) If all possible stimulus comparisons are done, as is usually the case, a large number of assessments are required for even modest numbers of experimental stimulus levels [if N levels are to be studied, $N(N-1)/2$ paired comparisons are needed].

b) If a stimulus difference exceeds approximately 1,5 JNDs, the magnitude of the stimulus difference cannot be directly estimated reliably because the response saturates as the proportions approach unanimity.

However, if a series of stimuli having no large gaps are assessed, the differences between more widely separated stimuli may be deduced indirectly by summing smaller, reliably determined (unsaturated) stimulus differences. The standard methods for transformation of paired comparison data to an interval scale (a scale linearly related to JNDs) perform statistically optimized procedures for inferring the stimulus differences, but they may yield unreliable results when saturated responses are included in the analysis.

3.10
preferential attribute
attribute of image quality that is invariably evident in an image, and for which the preferred degree is a matter of opinion, depending upon both the observer and the image content

EXAMPLE Examples of preferential image quality attributes include colour and tone reproduction properties such as contrast and relative colourfulness.

NOTE 1 Because the perceived quality associated with a preferential attribute is dependent upon both the observer and image content, in studies involving variations of preferential attributes, particular care is needed in the selection of representative sets of stimuli and groups of observers.

NOTE 2 The term *noticeable* in *just noticeable difference* is not linguistically strictly correct when applied to a preferential attribute, but is nonetheless retained in this part of ISO 20462 for convenience. For example, the higher contrast stimulus of a pair differing only in contrast might be readily identified by all observers, whereas there might be a lack of consensus regarding which of the two images was higher in overall image quality. Nonetheless, if the responses from the paired comparison for quality were in the proportion of 75:25, the image chosen more frequently would be said to be one JND higher in quality. The JND is best regarded as a measurement unit tied to the predicted or measured outcome of a paired comparison.

3.11
psychophysical method
experimental technique for subjective evaluation of image quality or attributes thereof, from which stimulus differences in units of JNDs may be estimated

cf. **magnitude estimation** (3.6), **paired comparison** (3.9), **quality ruler** (3.13), and **triplet comparison methods** (3.22)

3.12
quality just noticeable difference
quality JND
measure of the significance or importance of quality variations, corresponding to a stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task in which multivariate stimuli pairs are assessed in terms of overall image quality

NOTE See **attribute JND** (3.3) and **quality JND** (3.14) in ISO 20462-1:— for greater detail.

3.13
quality ruler method
psychophysical method that involves quality or attribute assessment of a test stimulus against a series of ordered, univariate reference stimuli that differ by known numbers of JNDs

3.14
reference stimulus
image provided to the observer for the purpose of anchoring or calibrating the perceptual assessments of test stimuli in such a manner that the given ratings may be converted to JND units

NOTE The plural is reference stimuli.

3.15
scene
content or subject matter of an image, or a starting image from which multiple stimuli may be produced through different experimental treatments

NOTE Typically, stimuli depicting the same scene are compared in a psychophysical experiment, because it is the effect of the treatment that is of interest, and differences in image content could cause spurious effects. In cases where scene content is not matched, a number of scenes should be used so that scene effects may be expected to average out.

3.16
standard quality scale
SQS
fixed numerical scale of quality having the following properties:

a) the numerical scale is anchored against physical standards;

b) a one unit increase in scale value corresponds to an improvement of one JND of quality; and

c) a value of zero corresponds to an image having so little information content that the nature of the subject of the image is difficult to identify

3.17
standard reference stimuli
SRS
set of reflection prints used in the hardcopy quality ruler, which vary in sharpness and are calibrated against the standard quality scale (SQS)

NOTE The SRS will be available on the I3A web site.

3.18
stimulus
image presented or provided to the observer either for the purpose of anchoring a perceptual assessment (a reference stimulus) or for the purpose of subjective evaluation (a test stimulus)

NOTE The plural is stimuli.

3.19
suppression
perceptual effect in which one attribute is present in a degree that seriously degrades image quality and thereby reduces the impact that other attributes have on overall quality, compared to the impact they would have had in the absence of the dominant attribute

NOTE To generate reference stimuli that are separated by a specified number of JNDs based on variations in one attribute, it will be necessary to ensure that other attributes do not significantly suppress the impact of the attribute varied.

3.20
test stimulus
image presented to the observer for subjective evaluation

NOTE The plural is test stimuli.

3.21
treatment
controlled or characterized source of the variations between test stimuli (excluding scene content) that are to be investigated in a psychophysical experiment

EXAMPLE Examples of treatments include different image processing algorithms, variations in capture or display device properties, changes in image capture conditions (e.g. camera exposure), etc.

NOTE Different treatments may be achieved through hardware or software changes, or may be numerical simulations of such effects. Typically, a series of treatments is applied to multiple scenes, each generating a series of test stimuli. The effect of the treatment may then be determined by averaging the results over scene and observer to improve signal to noise and reduce the likelihood of systematic bias.

3.22
triplet comparison
psychophysical method that involves the simultaneous scaling of three test stimuli with respect to image quality or an attribute thereof, in accordance with a set of instructions given to the observer

NOTE The triplet comparison method is described in more detail in ISO 20462-2.

3.23
univariate
describing a series of test or reference stimuli that vary only in a single attribute of image quality

4 Quality ruler experiments

4.1 General properties of quality rulers

A quality ruler is a univariate series of reference stimuli depicting the same scene and having known stimulus differences expressed in JNDs of quality. The reference stimuli are presented to the observer in a fashion facilitating:

a) the identification of the reference stimuli closest in quality to the test stimulus; and

b) the comparison of the test stimulus to those reference stimuli under rigorously matched viewing conditions.

Both hardcopy (Clause 5) and softcopy (Clause 6) implementations of quality rulers are described in this standard. Ruler images may be generated by the user (Clause 7). Reflection prints varying in sharpness and calibrated against the standard quality scale (SQS) are referred to as standard reference stimuli (SRS) (Clause 8).

NOTE The SRS will be available on the I3A web site.

The SRS may be used as ruler images or used to calibrate user-generated ruler images on an absolute basis, as distinguished from the relative calibration described in Clause 7.

4.2 Experimental conditions and reported results

Requirements regarding observer selection, test stimulus properties, instructions to the observer, viewing conditions, and reporting of results are set forth in ISO 20462-1.

NOTE 1 Sample instructions to the observer for hardcopy and softcopy quality ruler experiments are provided in informative Annexes A and B, respectively.

The viewing requirements of ISO 3664 shall be met except as modified in 4.4 of ISO 20462-1:—.

Reported values of quality in JNDs or SQS units shall be specifically identified if they are calculated from data 20 % or more of which fall at one of the ends of, or outside, the range of the quality ruler from which they were derived.

NOTE 2 Values based on ratings outside the range of the ruler will be less reliable because of extrapolation effects. In addition, when test samples fall within a JND or two of the high quality end of the ruler, a slight bias may result from observers avoiding use of ratings outside the ruler range. When preferential attributes (e.g. of colour and tone reproduction) are assessed using a quality ruler, it may be desirable to degrade all the test stimuli slightly by blurring (in the case of a ruler varying in sharpness) to allow headroom for test stimuli that are preferred over the reference stimulus.

The pedigree of the rulers used shall be reported, which entails specifying whether they are standard reference stimuli (SRS) or were otherwise generated. If the latter, the attribute varied in the rulers shall be stated. If such rulers vary in sharpness, the method of calibration shall be stated, which shall either be by comparison with SRS or using the average scene relationship (see 7.2).

4.3 Attributes varied in quality rulers

Clause 7 describes the generation of reference stimuli for rulers varying in sharpness, through modification of the modulation transfer function (MTF) of the system generating the images. Quality rulers may alternatively vary in other attributes, although only one attribute shall change within a given ruler. Alternative attributes that are varied in a quality ruler should be artefactual in nature.

NOTE The variation of preferential attributes within quality rulers is discouraged because of the additional variability associated with such attributes. Sharpness has been selected as the reference attribute because of several desirable characteristics:

a) it is easily manipulated through image processing;

b) it is correlated with MTF, which is readily determinable;

c) it has low scene and observer variability; and

d) it exerts a strong influence on quality in practical imaging systems.

Quality rulers varying in attributes other than sharpness shall be calibrated by having their reference stimuli rated against quality rulers varying in sharpness and meeting the criteria stated in this part of ISO 20462. The calibration experiment shall meet the specifications set forth in ISO 20462-1 and this part ISO 20462, with the exception that data from a minimum of 20 observers shall be averaged to determine the calibration.

5 Hardcopy quality ruler implementation

5.1 Physical apparatus

The hardcopy quality ruler apparatus shall consist of the following:

a) a sliding or translating fixture onto or into which a series of reference stimuli may be mounted or inserted (the *ruler*);

b) a test stimulus fixture in close proximity to the ruler;

c) a base surface upon which the ruler and the test stimulus fixture are attached;

d) an illumination system; and

e) a headrest or other device constraining the viewing distance (the distance from the observer's eye to the test and reference stimuli).

The ruler shall be constructed so that the observer may easily slide it to bring any of two reference stimuli into direct comparison with the test stimulus. In this triangular configuration of one test stimulus and two reference stimuli, the illumination level, illumination angle, viewing distance, and viewing angle shall be sensibly matched between the three stimuli. These features are illustrated in Figure 1.

The illumination angle should be 45° and shall fall between 30° and 60°. The viewing distance to any of the three stimuli shall be constrained by the headrest or equivalent mechanism to a range not exceeding 4 % of the value of the arithmetic average viewing distance. The range of the viewing distances of the three stimuli at a given observer head position shall not exceed 2 % of the arithmetic average viewing distance. The viewing angle should be normal to the stimulus surfaces and shall be within 10° of being perpendicular. Specular reflections from the stimuli shall not be visible from the observer's position.

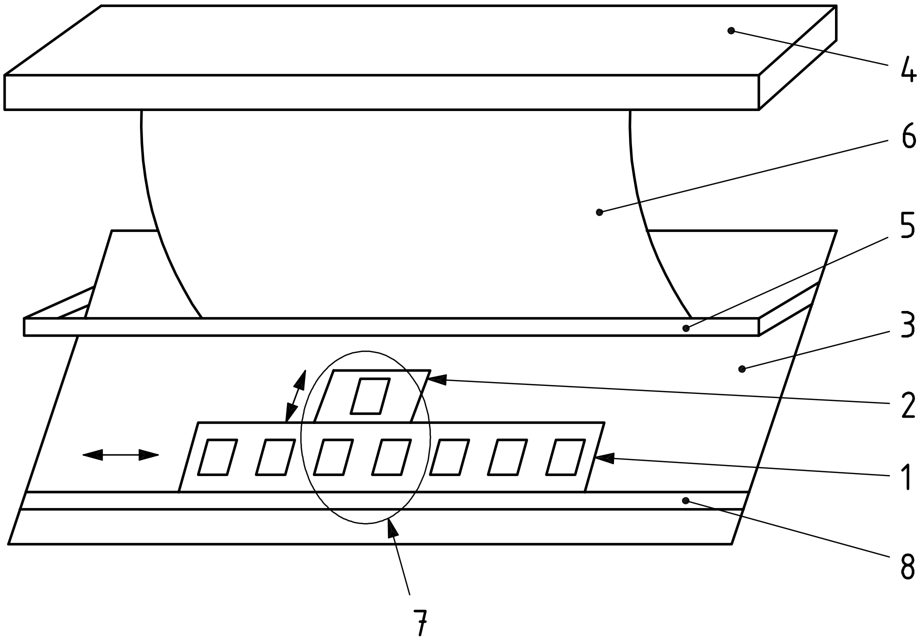
NOTE Achieving the closely matched viewing conditions of the test stimulus and the two reference (ruler) stimuli in the triangular configuration (which facilitates rating interpolation by the observer) is simplified if the physical separation of the three stimuli is minimized. Because some rulers may contain landscape (horizontal) format images and others portrait format (vertical) images, it may be advantageous for the test stimulus fixture to translate vertically. To match viewing angles between the test and reference stimuli, the receiving surface of the test stimulus fixture may have to be tilted.

5.2 Reference stimuli

The reference stimuli shall be ordered from highest to lowest quality from left to right in a horizontally translating ruler or top to bottom in a vertically translating ruler. These stimuli should be spaced by increments of approximately three JNDs. Each stimulus shall be labelled with an integer, and the observer shall provide ratings interpolated to the nearest integer value, which should correspond to approximately one JND scale resolution. The integer labels shall be chosen so that negative ratings are unlikely.

NOTE 1 The use of two interpolating positions between stimuli (for example, stimuli labelled 3 units apart with interpolation to one unit) has been found to yield a uniform and unbiased use of the numerical ratings, whereas when three interpolation positions are available, the numbers corresponding to the reference stimuli and those halfway in between can be used more frequently than those at the one-quarter or three-quarters positions. This result, combined with the difficulty of making evaluations more precise than one JND, leads to the recommendation that the reference stimuli be separated by approximately three JNDs.

NOTE 2 One suggested set of integer labels are 3, 6, 9, … from high to low quality.



Key

1 ruler

2 test stimulus fixture

3 base surface

4 illumination

5 head rest bar

6 black cloth to reduce glare

7 triangular configuration

8 ruler track

Figure 1 — Example of a hardcopy quality ruler apparatus

6 Softcopy quality ruler implementation

6.1 Physical apparatus

The softcopy quality ruler apparatus shall consist of the following:

a) one or more emissive devices such as video monitors with the necessary hardware and/or firmware to display images;

b) a keypad or other means of data entry by the observer;

c) a headrest or other device constraining the viewing distance [the distance from the observer's eye to the monitor faceplate(s)]; and, optionally,

d) a lighting system for controlling the surround illumination to influence the state of adaptation of the observer.

When two identical digital images are displayed simultaneously on the display device(s), their appearance shall be sufficiently similar that in paired comparisons for quality, the more frequently chosen image position (for example, the right monitor) shall not be selected more than 60 % of the time.

To minimize structural artefacts associated with the display, the viewing distance shall exceed 2 500 times the monitor line spacing (or pixel centre separation). The viewing distances (from the observer's eye to the faceplate at the centre of the image) shall be constrained by the headrest or equivalent mechanism to a range not exceeding 4 % of the value of the arithmetic average viewing distance. The range of the viewing distances at a given observer head position shall not exceed 2 % of the arithmetic average viewing distance. The viewing angle shall be within 10° of being perpendicular to the display faceplate at the centres of the images. The angle subtended by the centres of the images from the observer's position should not exceed 30° to avoid requiring the observer to turn their head to change their view from one image to the other.

6.2 Reference stimuli

The reference stimuli should be spaced by increments of approximately one JND.

The maximum precision of a single determination is plus or minus one-half of the reference stimulus spacing.

6.3 Controlling software

The software that controls the display of test and reference stimuli and records the data shall provide the following functions, listed in sequential order:

a) selection of the test stimulus to be evaluated;

b) random selection of the display position of the test stimulus;

c) selection of the initial reference stimulus to be provided;

d) display of the selected stimuli at their selected positions;

e) recording of the stimulus chosen by the observer;

f) selection of a new reference stimulus based upon the observer's response;

g) display of the new reference stimulus, which replaces the previous one;

h) repetition of e) to g) until the test image has been rated differently against two adjacent reference images;

i) recording of the final rating (arithmetic average of the values of the adjacent reference stimuli); and

j) return to a) for a new test stimulus, until all test stimuli have been evaluated.

The selection of the test stimulus a) should be random except that test stimuli may be grouped by scene, in which case the group order should be random, as well as the treatment order. The selection of the initial reference stimulus c) should be random. The choice of the new reference stimulus f) shall be based upon the previous responses of the observer for the present test stimulus. The new reference stimulus shall be higher (lower) in quality than the highest (lowest) quality reference stimulus identified by the observer as being lower (higher) in quality than the test image. Once adjacent reference stimuli (in terms of their order of quality) have received different ratings relative to the test stimulus (the higher quality reference being preferred to the test stimulus, which was chosen over the lower quality reference), the condition of h) is met and the process shall terminate for that test stimulus *i*).

It is recommended that the new reference stimulus f) be chosen so that it falls approximately halfway between the lowest quality reference stimulus preferred over the test stimulus and the highest quality reference stimulus not chosen over the test stimulus, so that an approximately binary search is carried out. Until some reference stimulus has won (lost) a paired comparison with the test stimulus, the highest (lowest) quality reference stimulus may be used as a proxy. An example of pseudocode performing such a binary search is provided in informative Annex C.

7 Generation of quality ruler stimuli

7.1 General requirements

Excluding the effect of the attribute varied within the quality ruler, the reference stimuli shall have high image quality, with pleasing colour (if applicable) and tone reproduction, and an absence of significant degradation from artefacts under the existing viewing conditions.

NOTE These requirements are intended to prevent the suppression by other attributes of the effect on overall image quality of the attribute varied within the ruler.

7.2 Modulation transfer functions (MTFs)

The MTF of the complete imaging system generating a reference stimulus for a quality ruler varying in sharpness shall be characterised by measurement of neutral test targets and/or equivalent calculations based upon linear systems theory. MTFs shall be determined in both horizontal and vertical orientations, at the centre of the image area (on-axis) and at one or more points halfway between the centre of the image and the corners of the image (50 % field position). In computing the overall system MTF, the on-axis position shall have a weight of 3/7 and the off-axis position (or mean of positions) a weight of 4/7, and the field-weighted horizontal and vertical MTFs shall be weighted by 1/3 and 2/3, the higher weight being assigned to the poorer MTF, which shall be defined to be that with lesser mean modulation transfer from 0 to 30 cycles per degree (CPD) at the eye of the observer.

The system MTF so determined shall closely conform to the shape of the monochromatic MTF of an on-axis diffraction-limited lens, $m(\nu)$, which is given by

$$\begin{aligned} m(\nu) &= \frac{2}{\pi}\left(\cos^{-1}(k\nu) - k\nu\sqrt{1-(k\nu)^2}\right) && k\nu \leqslant 1 \\ m(\nu) &= 0 && k\nu > 1 \end{aligned} \tag{1}$$

where

ν is spatial frequency in CPD at the eye of the observer;

k is a constant.

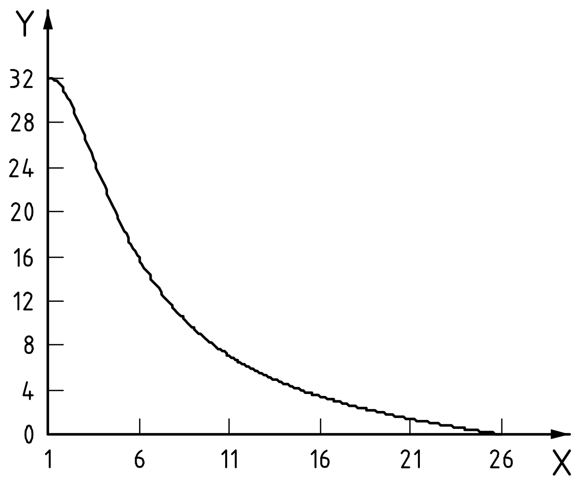
NOTE 1 For a diffraction-limited lens, the constant k would equal the product of the wavelength of light and the lens aperture (f-number). However, in this application, Equation (1) is only being used to represent a possible shape of an entire imaging system MTF, so k is better regarded as being reciprocally related to system bandwidth.

For purposes of verifying whether the shape of a system MTF conforms sufficiently closely to the shape of Equation (1), an equivalent k value shall be determined by finding the value of k such that the area under the MTF of Equation (1) equals that under the system MTF over the frequency range of 0 to 30 CPD. The MTF given by Equation (1) for the value of k so derived shall be referred to as the aim MTF. The system MTF shall be considered to be within conformance and valid for use if the mean fractional modulation transfer of the system and aim MTFs over each of the frequency bands 0 to 5, 5 to 10, …, and 35 to 40 CPD agree to within 0,05.

A relative quality JND value associated with a given value of k for an average scene and typical colour and tone reproduction shall be computed via Equation (2).

$$\text{JNDs} = \frac{17\,249 + 203\,792\,k - 114\,950\,k^2 - 3\,571\,075\,k^3}{578 - 1\,304\,k + 357\,372\,k^2} \quad (1 \leqslant 100k \leqslant 26) \qquad (2)$$

The difference in quality JNDs between two reference stimuli depicting an average scene and having conforming system MTFs shall be computed as the difference between the scale values produced by Equation (2).

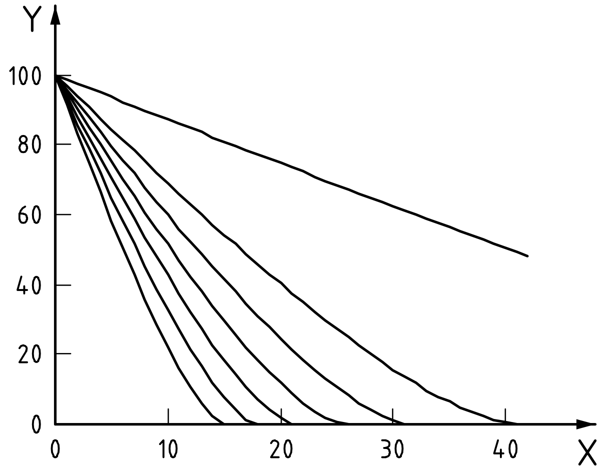


Key

X 100 k

Y relative JNDs

Figure 2 — Plot of Equation (2)



Key

X frequency, cycles per degree

Y modulation transfer, %

Figure 3 — MTFs from Equation (1) spaced by 3 JNDs

NOTE 2 The absolute scale value of Equation (2) has no assigned significance; only differences between scale values of valid system MTFs are ascribed meaning.

NOTE 3 Figure 2 shows the behaviour of Equation (2). For demonstration purposes, a series of values of k were chosen giving 3 JND increments of quality according to Equation (2) (these values were $10^4 \times k$ = 100, 245, 320, 392, 469, 558, and 666). The associated MTF curves from Equation (1) are plotted in Figure 3, with the lower k values corresponding to the higher MTFs.

The deviations of the system MTF shapes within a single ruler series should differ from the aim MTF shapes in as consistent a fashion as possible to minimize errors in the computed differences in JNDs.

If the system MTFs are not within conformance, the reference stimuli shall be calibrated in the same fashion as would stimuli varying in an attribute other than sharpness, as described in 4.3.

7.3 Scene-dependent ruler calibration

To reflect the different dependence of quality on attribute level in different scenes, quality rulers depicting different scenes should be individually calibrated in JNDs by presenting them as test stimuli in a quality ruler experiment against standard reference stimuli (SRS). If a quality ruler is not so calibrated, but rather Equation (2) is used to assign JND values, results obtained from the ruler shall be averaged with results from at least two other rulers, and none of the scenes depicted in these rulers shall be of a type expected to have unusually strong or weak quality dependencies on the attribute varied.

NOTE By averaging the results of several ruler scenes, potential biases caused by using the calibrations for an average scene may be mitigated. Scenes with important high-frequency information, such as some landscapes, are likely to have stronger than average quality dependencies on MTF. Conversely, scenes with particularly limited bandwidth, like some portraits, are likely to have quality change more slowly with MTF than would be the case for an average scene.

8 Standard quality scale (SQS) determinations

8.1 Properties of the SQS

The standard quality scale (SQS) is a fixed numerical scale of image quality that is anchored against physical standards. The scale units are quality JNDs and more positive values indicate higher image quality. An SQS value of zero corresponds to an image having so little information content that the nature of the subject of the image is difficult to identify. The physical standards associated with the SQS scale are referred to as standard reference stimuli (SRS).

NOTE Sets of the SRS reflection prints, which vary in sharpness, will be available on the I3A web site. The calibration of the SQS and SRS are described in informative Annex D.

8.2 Experimental requirements

SQS values shall be determined through a hardcopy quality ruler experiment using standard reference stimuli (SRS). The experiment shall meet the specifications set forth in ISO 20462-1 and this part of ISO 20462, with the following exceptions.

a) The viewing distance shall be 406 mm, with tolerances as given in 5.1.

b) Data from a minimum of 20 observers and 6 scenes shall be averaged to determine a reported SQS value for an experimental treatment.

NOTE The more stringent scene and observer number requirements for reporting SQS values, compared to JND differences, reflects the absolute rather than relative nature of the scale. Individual stimuli may be annotated with an SQS value if they have been assessed by at least 20 observers, as described here, but their values may not be reported in connection with a particular experimental treatment, such as an attribute level, a device, an algorithm, etc.

Annex A
(informative)

Sample instructions for a hardcopy quality ruler experiment

The following is an example of the instructions that might be read by the test administrator to the observer in a hardcopy quality ruler experiment. Text in italics directs the administrator to perform certain actions and so is not read aloud. Examples of the results from this experiment are shown in Annex E.

Display the quality ruler depicting Scene #1 for demonstration purposes.

First, I would like to thank you for participating in this study. Please put on the lab coat and gloves and make yourself comfortable in the chair in front of the viewing table.

In this experiment, you will be evaluating the overall quality of prints made from images that have one or more of the three colour planes shifted out of register. This misregistration may affect the image sharpness and may cause various image artefacts. Let me show you some examples of these images.

Give the subject preview Print #1.

Some images you will see may exhibit only small or even unnoticeable levels of unsharpness and colour fringing around edges. For example, in this image, look at the windows in the building on the right.

Give the subject preview Print #2.

Some images may appear as two or more sharp, coloured images offset from each other as in this scene. Notice the horizontal edges in the lipstick cases in this image.

Do you have any questions about the attribute that you will be judging today?

You will be evaluating these samples using a quality ruler like the one in front of you now. The ruler provides you with a series of prints at different levels of quality, produced by variations in sharpness. The ruler print quality decreases from left to right; however, the ruler numerical values increase from left to right, to reflect quality degradation. The numerical values are defined so that one unit is approximately one just noticeable difference. Here is how you use the ruler:

a) Place the test image flat in the holder above the ruler. Slide the ruler right or left to permit comparison of the test sample with different ruler images, which should be underneath the test sample for a valid assessment.

b) Locate the position of equality on the ruler such that each print farther right is lower in quality than the test image, and each image farther left is higher in quality.

c) Read off the number of the position of equality. If this position falls in between two ruler prints, as it often will, please select an intermediate number from the ruler scale. For example, if the point of equality is in between the ruler prints “12” and “15”, but is closer to “12”, you might assign the test print a value of “13”.

If you feel a test print is higher in quality than the “3” print on the ruler, you may assign it an integer value less than 3. If the test print is lower in quality than the “21” print you may assign it an integer value greater than 21.

When you are finding the point of equality on the ruler, it is important to consider overall quality. It may help to imagine that the image is one that you treasure, and you have a choice of two prints of that image: one is misregistered, whereas the other is unsharp. Compare the test print with different ruler prints and decide which one you would choose to keep. From these comparisons, identify the position on the ruler from which prints to the right are less desirable than the test sample, and those to the left are more desirable.

Note that although some misregistered images may simply appear unsharp, please do not just match the sharpness of the test and ruler images. It is important to concentrate on assessing overall quality, because most of the misregistered images will show artefacts in addition to unsharpness.

Let me show you how I would evaluate one print.

Evaluate preview Print #3 for the observer.

Finally, please disregard any physical damage such as scratches on the test prints during your evaluations.

Do you have any questions at this time?

Answer questions, then begin test.

Annex B
(informative)

Sample instructions for a softcopy quality ruler experiment

Thank you very much for participating in this study. Please wear the black gown provided to reduce stray light.

In this experiment, you will be evaluating the overall quality of a series of images using a psychophysical technique called the softcopy quality ruler. Please remember there are no right or wrong answers because we are asking you which image you perceive as having higher overall quality.

This is how we are asking you to evaluate the test images:

a) A pair of images will be presented to you on the two monitors in front of you. For each pair of test and reference images, we ask you to indicate which image you perceive to be of higher overall quality. You enter your decision by pressing either the LEFT or RIGHT key on the entry keypad.

b) The test images shown on one monitor represent different cases of tone reproduction. These variations can make the images appear “snappy” or “flat,” and they can also determine how colourful the images appear. In addition, some images were rendered lighter or darker in overall tone. During this session you will be evaluating 12 different tone variations in each of 6 different scenes.

c) You will be comparing each test image with a series of reference images appearing on the other monitor. These reference images all have the same tone scale and colour rendering; however, each image has a different sharpness. We ask you to select the image you perceive to be of higher overall quality and would prefer if you were allowed to keep only one image. In some cases the image pairs may appear very similar; however, you should make a choice.

d) When you have completed the series of comparisons to evaluate a given test image, a tone will sound to indicate you are evaluating a new test image. Note that the monitors on which the test image and reference images are displayed will change in a random fashion.

Demonstrate how you would evaluate an image using the first trial image.

There are two other features of the system of which you should be aware:

— If you feel you made an incorrect response and would like to re-evaluate the image you are currently evaluating, press the REDO key. This will restart the evaluation sequence for the current image you are evaluating and record your new answer. Once you use the REDO key, you do not need to take any action to continue with the experiment.

— If you feel you want to re-evaluate a previous image that is not currently on the screen, press the PREV key. Each time you press the PREV key you will index back one test image. When you have reached the image you want to re-evaluate, evaluate it just as you normally would. When your evaluation is complete, the tone will sound. You can then index forward, using the NEXT key, to where you were prior to your detour. The system remembers all your intermediate responses. You can hit the NEXT key until the system stops indexing forward; it will not index past an image you have not evaluated.

Review the Table B.1 with the observer and be sure they are comfortable with the keypad functions.

Table B.1 — Instructions for conducting the evaluation

If the following is true:	then press this key:
You perceive the left image to be higher in overall quality	LEFT
You perceive the right image to be higher in overall quality	RIGHT
You want to restart the evaluation of the *current* test image	REDO
You want to re-evaluate any *previous* test image	PREV
You want to return to your *current* test image after a re-evaluation	NEXT

I will be leaving this set of instructions and the table with you for reference during the experiment. Do you have any questions about how the keypad is used to enter your responses and control the system?

Please evaluate the next few trial images for practice.

Watch the observer evaluate several test images from the trial sequence.

Do you have any questions about the changes in tone reproduction occurring in the test images?

Please feel free to evaluate some more of the images in the trial set. When you feel comfortable using the softcopy ruler, we will end the trial, and you can begin the actual experiment.

Start the experiment when the observer is ready. Stay until the observer has evaluated five images.

Do you have any questions about any part of the experiment? If not, I will be leaving to let you proceed with the experiment. If you have any questions during the experiment you may contact me at extension 12345 using the telephone on the table behind you.

Annex C
(informative)

Sample code of a binary search routine for the softcopy quality ruler

The pseudocode shown below conducts a binary search for the point of equivalence between a test image and a softcopy quality ruler. It is assumed that the test image has already been displayed on a randomly chosen monitor.

The reference images in the softcopy quality ruler are indexed from 1 to *num_ref* from highest to lowest quality. The function *ran_uni* is a uniform pseudorandom number generator over the open range zero to one (i.e. the returned values are greater than zero and less than one). The function *nint* rounds a real number to the nearest integer. The variable string *choice*, having a value of either "test" or "reference", is the image identified as having higher quality by the observer. The variable *qual_test* is the output rated quality in JNDs or SQS values based upon the ruler calibration array *qual_ref* [*num_ref*].

The algorithm randomly selects a starting ruler image index, which provides the initial value of the integer variable *cur_ref*, the current reference level. The integer variable *min_ref* is the index of the lowest quality ruler image that has been chosen as being higher in quality than the test image, whereas *max_ref* is the index of the highest quality ruler image that has been chosen as being lower in quality than the test image. These variables are initialized to values of 0 and *num_ref* +1, respectively, which represent positions just outside the extreme ruler levels. After each paired comparison, *cur_ref* and either *min_ref* or *max_ref* are updated. The process is terminated when *min_ref* and *max_ref* differ by one.

If the test image was chosen in preference to the highest quality ruler image, or the lowest quality ruler image was chosen in preference to the test stimulus, the quality rating is linearly extrapolated. The integer variable *extrap* is set to 0 if the rating is interpolated, 1 if the rating is extrapolated past the high-quality end of the ruler, and 2 if extrapolated past the low-quality end.

```
* initialization;
min_ref = 0;
max_ref = num_ref + 1;
cur_ref = nint( 0,5 + num_ref × ran_uni(seed) );

* perform binary search;
do until ( max_ref – min_ref = 1 );
      display reference_image[cur_ref];
      obtain choice;
      if ( choice = "reference" ) then;
            min_ref = cur_ref;
else if ( choice = "test" ) then;
            max_ref = cur_ref;
      endif;
      cur_ref = nint( ( min_ref + max_ref) / 2 );
enddo;
* compute quality rating;
if ( min_ref = 0 ) then;
      extrap = 1;
      qual_test = ( 3 × qual_ref[1] – qual_ref[2] ) / 2;
else if ( max_ref = num_ref + 1 ) then;
      extrap = 2;
      qual_test = ( 3 × qual_ref[num_ref] – qual_ref[num_ref – 1] ) / 2;
else;
      extrap = 0;
      qual_test = ( qual_ref[max_ref] + qual_ref[min_ref] ) / 2;
endif;
```

Annex D
(informative)

Calibration of the standard quality scale (SQS) and its reference stimuli

This annex describes how the standard quality scale (SQS) was defined and calibrated, and how the standard reference stimuli and their MTF aims were generated. A large number of reflection prints representative of consumer and professional photography were assembled. These stimuli spanned essentially the full range of quality realized in such images, and they varied widely in the levels of many different artefactual and preferential attributes of image quality. These stimuli were assessed for overall quality by expert observers using magnitude estimation, thereby generating a preliminary numerical scale. Small groups of very similarly rated stimuli clustered around each of a number of scale positions, which spanned the full range of the numerical scale, were identified. These groups were individually rank-ordered for overall quality by both trained observers and representative consumers. For each pair of stimuli within a group, the probability of the stimulus with the higher scale value being chosen in a paired comparison was deduced from the rankings given to each stimulus by each observer. The resulting probability was converted to JNDs using Equation (B.2) from ISO 20462-1:—, Annex B, which is repeated here.

$$\text{JNDs} = \frac{12}{\pi}\sin^{-1}\left(\sqrt{p}\right) - 3 \qquad \text{(D.1)}$$

The resulting value for each stimulus pair was plotted against the scale value difference between the stimuli. As the scale value difference increased (indicating a larger quality difference between the stimuli), the probability of the higher scale value stimulus being chosen increased accordingly. A smooth curve was fit through the data, and the stimulus scale value difference predicted to yield 75:25 proportions was read off the plot. This value is called the JND increment of the scale, and it is the rating scale difference that corresponds to one JND of stimulus difference. This procedure was repeated for each of the other groups, and the resulting JND increments were plotted against the average scale value of their associated group of stimuli. These data were fit well by a straight line from the lowest to the highest scale values, providing a continuous function of JND increment versus scale value.

The next step was to convert the preliminary rating scale to a final scale of JNDs of quality. This was accomplished using the following equation[4]

$$Q(s) = Q_\text{r} + \int_{s_\text{r}}^{s} \frac{\text{d}s'}{\Delta s_J\,(s')} \qquad \text{(D.2)}$$

where

s is the scale value,

$Q(s)$ is the quality in JNDs at s,

s_r is a reference value that maps to Q_r,

$\Delta s_J(s)$ is the JND increment, which in this instance was linearly related to s.

The integrand is an infinitesimal scale value change divided by the JND increment at that scale value, which has units of scale values per JND. The integrand thus corresponds to infinitesimal JNDs, which are accumulated by the integral starting at the reference position. The choice of the reference mapping, which is arbitrary, was made as follows. The point on the rating scale at which it became difficult to identify the nature of the principal subject matter in the image was mapped to an SQS value of zero JNDs. On this scale, a typical excellent image might have a value of approximately 30 JNDs. Calibrated primary physical standards were constructed using the multivariate images that were rated in this experiment.

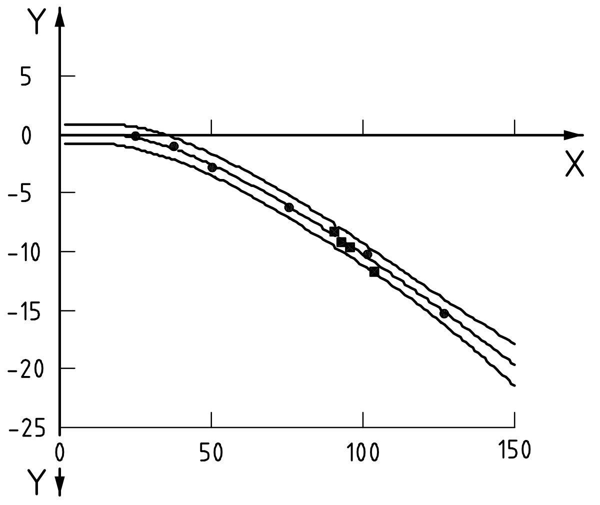
To construct standard reference stimuli, series of univariate images were made by progressively blurring high fidelity scene captures. These MTF series were assessed for quality by trained experts, using the primary multivariate standard. Regressions of quality versus an MTF-based objective correlate of sharpness were obtained, allowing, for this family of MTF variations, interpolation to arbitrary quality positions. From this relationship the MTFs given in 7.2 were derived.

JNDs are proportional to z-scores, which are mean stimulus differences divided by the standard deviation of the perceptual process (often called discriminal dispersion). Thus, JNDs are perceptual signal to noise ratios that depend on both the physical differences between stimuli (signal) and the assessment variability (noise). Consequently, a constant physical stimulus difference measured in multiple psychophysical experiments will yield unequal JND values if the discriminal dispersion varies between experiments. Perceptual variability is a function of the observers, the task being performed, and the nature of variation among the stimuli. For example, assessment of overall quality, multivariate stimuli, and preferential attributes is more disperse than evaluation of individual attributes, univariate stimuli; and artefactual attributes, respectively. As described above, the JNDs and SQS values of this part of ISO 20462 are based upon assessment of overall quality of stimuli varying in many attributes, both artefactual and preferential in nature.

Annex E
(informative)

Example of results from quality ruler experiments

This annex provides examples of the types of results obtained from a quality ruler experiment. The instructions for the first experiment described here were given in Annex A. The artefact under study in this experiment was misregistration of colour records. The treatments consisted of shifts of one or more colour records by integral numbers of pixels. A primary series of treatments involved shifts of the green record only in amounts varying from sub-threshold to seriously degrading. Ancillary treatments included moderate shifts of the blue record, the red record, and both records simultaneously in orthogonal (90°) and in opposite (180°) directions. Each treatment was applied to 12 different pictorial scenes, which were rated against a hardcopy quality ruler by 20 observers. Null stimuli having no misregistration, and so identical to one of the quality ruler positions, were included to test for observer bias. These null stimuli were assigned quality loss values of zero and values for other stimuli were determined by difference. By assigning empirically optimized "weights" to each colour record, an objective metric of visual misregistration, expressed as a subtense at the observer's eye in arc-seconds, was computed for each treatment. Figure E.1 depicts the results, pooled over all observers and scenes. A regression with a 95 % confidence interval is based upon the green record shift data only (circles). The ancillary treatment data (squares) falls within the regression confidence interval, supporting the general correlation of the visual subtense metric with perceived quality loss from misregistration for different geometries of different colour records. Figure E.2 depicts regression curves derived for subsets of observers and scenes to show the impact of their variability. In addition to the regression curve from Figure E.1 for average observers and scenes, in Figure E.2 curves are shown for the 50 % more (less) sensitive observers assessing the 25 % most (least) susceptible scenes, which lie below (above) the average curve. These results may be reported because each datum is based upon (12 scenes × 25 %) × (20 observers × 50 %) = 30 determinations, meeting the minimum criterion of ISO 20462-1.



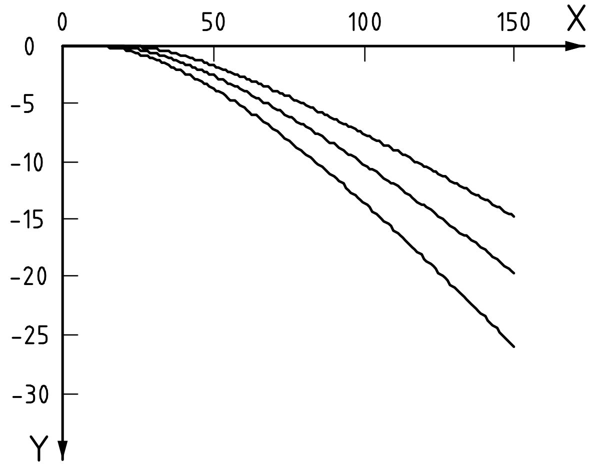
Key

X visual misregistration, arc-seconds

Y JNDs of quality

NOTE Reprinted from the *Handbook of Image Quality: Characterization and Prediction* with permission of the publisher, Marcel Dekker, Inc. See Reference [5].

Figure E.1 — Results pooled over all scenes and observers



Key

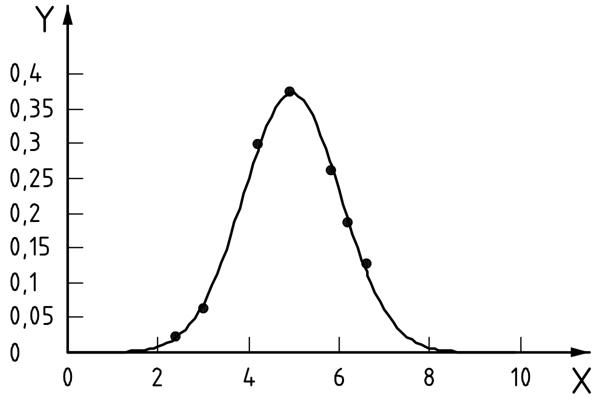
X visual misregistration, arc-seconds

Y JNDs of quality

NOTE Reprinted from the *Handbook of Image Quality: Characterization and Prediction* with permission of the publisher, Marcel Dekker, Inc. See Reference [5].

Figure E.2 — Results for different observer/scene subsets

Data from several studies of preferential attributes of colour and tone reproduction are presented next. Figures E.3 and E.4 show the results of a softcopy quality ruler experiment characterising the impact of overall neutral scale contrast on quality. Each scene was rendered at a series of contrast positions, so it was possible to determine the contrast values preferred for each scene by each observer. These preferred values were accumulated into the probability density distribution shown in Figure E.3. The x-axis metric of Figure E.3 is a linear transform of the logarithm of contrast; in this space, the preference distribution is Gaussian (that is, it is a normal distribution). The quality ruler data also permitted the determination of the average quality loss function, which quantifies how quality falls off as the contrast deviates from the value preferred by an observer. Figure E.4 shows this quality loss function; the x-axis is the deviation from the preferred value (zero corresponds to the preferred position). The quality loss function is fit well by a quadratic function with the quality loss defined to be zero at the preferred contrast, so that the parabola passes through the origin. The combination of the preference distribution and the quality loss function together provide a compact but rather complete characterization of the effect of contrast on quality. Powerful predictions may be made based on these two functions.[6] For example, by convolving the functions, one may compute the average improvement in quality that would result if all images were somehow rendered at the exact contrast that the observer preferred in that individual image. One application of such a value is that it defines an upper bound on the potential benefit of adaptive, customized digital imaging algorithms.



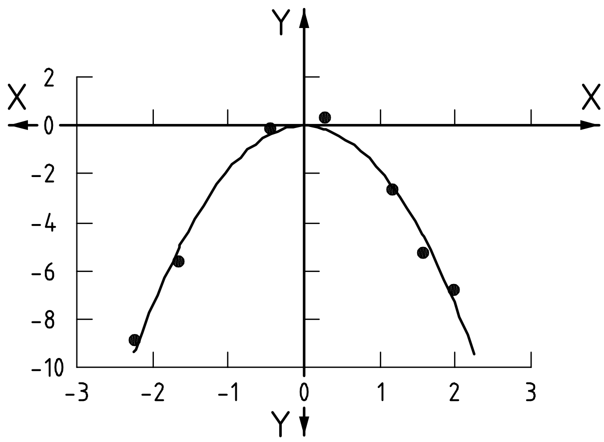
Key

X contrast metric

Y probability density

NOTE Reprinted from the *Handbook of Image Quality: Characterization and Prediction* with permission of the publisher, Marcel Dekker, Inc. See Reference [5].

Figure E.3 — Preference distribution



Key

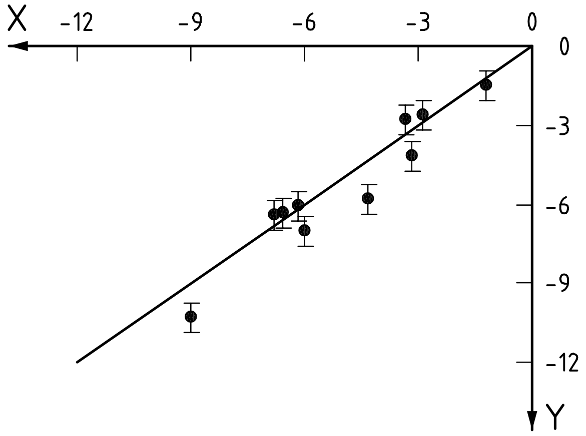
X contrast metric difference from optimum

Y JNDs of quality

NOTE Reprinted from the *Handbook of Image Quality: Characterization and Prediction* with permission of the publisher, Marcel Dekker, Inc. See Reference [5].

Figure E.4 — Quality loss function

A final example of quality ruler results are shown in Figure E.5. Separate quality ruler experiments were conducted on neutral colour balance, contrast, detail in saturated colours, and the reproduction of skin-tones, foliage, and blue sky. In each experiment, attributes other than that being studied were held constant or nearly so. The results of these univariate experiments were combined using the multivariate formalism[7] to predict the overall quality of images in which all of these attributes simultaneously varied and affected quality. Such images were then generated and evaluated in a separate multivariate quality ruler experiment. The measured quality values are plotted with 95 % confidence intervals in Figure E.5; the dashed line shows the predicted values (it is not a regression line; no fitting or adjustments were made to the data). There is good agreement between the measurements and predictions, demonstrating the degree of consistency possible between independent univariate and multivariate quality ruler experiments.



Key

X predicted JNDs of quality

Y JNDs of quality

NOTE Reprinted from the *Handbook of Image Quality: Characterization and Prediction* with permission of the publisher, Marcel Dekker, Inc. See Reference [5].

Figure E.5 — Verification of multivariate consistency

Bibliography

[1] MIYAZAKI, K., KANAFUSA, K., UMEMOTO, H., TAKEMURA, K., URABE, H., HIRAI, K., ISHIKAWA, K. and HATADA, T. A standard portrait image and image quality assessment (II) — Triplet comparison, *Proc. SPIE*, **4300**, 2001, pp. 309–313.

[2] TAKEMURA, K., MIYAZAKI, K., URABE, H., TOYODA, N., ISHIKAWA, K. and HATADA, T. Developing a new psychophysical experimental method to estimate image quality, *Proc. SPIE,* **4421**, 2001, p. 906

[3] KEELAN, B.W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, ISBN 0-8247-0770-2, 2002, Chpt. 7 and 8

[4] KEELAN, B.W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, ISBN 0-8247-0770-2, 2002, Chpt. 5

[5] KEELAN, B.W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, ISBN 0-8247-0770-2, 2002

[6] KEELAN, B. W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, ISBN 0-8247-0770-2, 2002, Chpt. 4

[7] KEELAN, B. W. *Handbook of Image Quality: Characterization and Prediction*, Marcel Dekker, Inc., New York, ISBN 0-8247-0770-2, 2002, Chpt. 11

[8] TORGERSON, W. S. *Theory and Methods of Scaling*, Wiley, New York, 1958

[9] ISO 20462-1:2005, *Photography — Psychophysical experimental methods for estimating image quality — Part 1: Overview of psychophysical elements*

[10] ISO 20462-2:2005, *Photography — Psychophysical experimental methods for estimating image quality — Part 2: Triplet comparison method*

ICS 37.040.01

Price based on 24 pages