INTERNATIONAL STANDARD

ISO 20462-1

First edition
2005-11-01

Photography — Psychophysical experimental methods for estimating image quality —

Part 1: Overview of psychophysical elements

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

Partie 1: Aperçu général des éléments psychophysiques

Reference number
ISO 20462-1:2005(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.

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

ISO 20462-1 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. There are a number of psychometric methods described in the literature, such as paired comparison, rank ordering, categorical sort, and magnitude estimation, which might be considered as candidates for experimentally measuring image quality. Several textbooks[1] [3] [4] [5] [9] [12] have reviewed these and other methods and have discussed associated data reduction techniques, which usually are based upon the approach of Thurstone[11] or analogous reasoning. However, the choice of the best 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 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 value of new calibrated psychometric methods in developing comprehensive models of imaging system quality has been demonstrated in a recent work[6] that contains more detailed discussions of many of the informative topics superficially considered herein.

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; specific experimental methods and associated data reduction techniques are described in Part 2 (triplet comparison method[8] [10]) and Part 3 (quality ruler method[6]). Informative Annex A aids in identifying the better choice between the two alternative methods of Parts 2 to 3, which are complementary and together are sufficient to span a wide range of applications. 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 just noticeable differences (JNDs), so that the significance of experimentally measured stimulus differences is readily ascertained.

Photography — Psychophysical experimental methods for estimating image quality —

Part 1: Overview of psychophysical elements

1 Scope

This part of ISO 20462 is part of a multiple-part standard pertaining to the subjective evaluation of pictorial still image quality. This part of ISO 20462

a) defines the units by which image quality is quantified (just noticeable differences, or JNDs);

b) describes the influence of stimulus properties, observer characteristics, and task instructions on results obtained from rating experiments;

c) provides a flow chart for choosing the preferred psychophysical method for determining image quality from among those defined in subsequent parts of ISO 20462.

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 purpose 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** (3.12)

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
attribute just noticeable difference
attribute JND
measure of the detectability of appearance variations, corresponding to a stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task in which univariate stimuli pairs are assessed in terms of a single attribute identified in the instructions

cf. **quality JND** (3.14)

NOTE 1 As an example, a paired comparison identifying the sharper of two stimuli that differ only in their generating system modulation transfer function (MTF), would yield results in terms of sharpness attribute JNDs. If the MTF curves differed monotonically and did not cross, the outcome of the paired comparison would depend primarily upon the observers' ability to detect changes in the appearance of the stimuli as a function of MTF variations, with little or no value judgement required of the observers. The relationship between paired comparison proportions and stimulus differences is discussed in greater detail in Annex B.

NOTE 2 If observers are instead asked to choose which of a pair of stimuli is higher in overall image quality, and if the stimuli in aggregate are multivariate, such that the observer should make value judgements of the importance of a number of attributes, rather than focussing on one aspect of image appearance, it is observed experimentally that larger objective stimulus differences (for example, MTF changes) are required to obtain a 75:25 proportion of responses, which in this case corresponds to a quality JND.

NOTE 3 A JND is a statistical quantity, derived from a number of observations. An observer assessing a single pair of images differing by one attribute JND is unlikely to be confident that he or she has detected the sample difference. A stimulus difference of approximately three JNDs is usually needed for an observer of average sensitivity to feel reasonably certain of his or her assessment.

3.4
categorical sort method
psychophysical method involving the classification of a stimulus into one of several ordered categories, at least some of which are identified by adjectives or phrases that describe different levels of image quality or attributes thereof

NOTE The application of adjectival descriptors is strongly affected by the range of stimuli presented, so that it is difficult to compare the results of one categorical sort experiment to another. Range effects and the coarse quantization of categorical sort experiments also hinder conversion of the responses to JND units. Given these limitations, it is not possible to unambiguously map adjectival descriptors to JND units, but it is worth noting that in some experiments where a broad range of stimuli have been presented, the categories *excellent*, *very good*, *good*, *fair*, *poor*, and *not worth keeping* have been found to provide very roughly comparable intervals that average about six quality JNDs in width.

3.5
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.

3.6
instructions
set of directions given to the observer for performing the psychophysical evaluation task

3.7
just noticeable difference
JND
stimulus difference that leads to a 75:25 proportion of responses in a paired comparison task

cf. **attribute JND** (3.3) and **quality JND** (3.14)

3.8
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.

3.9
multivariate
describing a series of test or reference stimuli that vary in multiple attributes of image quality

3.10
observer
individual performing the subjective evaluation task in a psychophysical method

3.11
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.

3.12
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. 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 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.

3.13
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. **categorical sort** (3.4), **magnitude estimation** (3.8), **paired comparison** (3.11), **quality ruler** (3.15), **rank ordering** (3.16) and **triplet comparison methods** (3.24)

3.14
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 1 See Notes for **attribute JND** (3.3).

NOTE 2 The attribute JND is a measure of detectability of appearance changes, whereas the quality JND is a measure of significance or importance of stimulus differences in terms of their impact on quality. An attribute JND is a useful unit for predicting how observers would react to an advertisement showing images carefully matched in all respects but one, and drawing the attention of the observer to the attribute varying. In contrast, a quality JND is useful for predicting how observers would perceive overall quality as a function of one or more stimulus variations, and so is a more useful quantity in optimizing imaging system design, where different attributes should be balanced against one another. The overall quality of an image may be predicted from a knowledge of the impact of each attribute in isolation, expressed in terms of quality JNDs, whereas the same is not true of attribute JNDs. Therefore, it is often highly desirable to obtain results expressed in quality JNDs, even if the stimuli being assessed are univariate in nature. This can be accomplished if test stimuli are rated against a series of appropriately calibrated reference stimuli, as in the quality ruler method.

3.15
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

NOTE The quality ruler method is described in more detail in ISO 20462-3.

3.16
rank ordering method

psychophysical method involving the arrangement by an observer of a series of stimuli in order of increasing or decreasing image quality or an attribute thereof, in accordance with the set of instructions provided

3.17
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.

3.18
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.

3.19
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

NOTE The standard quality scale is described in more detail in ISO 20462-3.

3.20
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.

3.21
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.

3.22
test stimulus
image presented to the observer for subjective evaluation

NOTE The plural is test stimuli.

3.23
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.

3.24
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.

3.25
univariate
describing a series of test or reference stimuli that vary only in a single attribute of image quality

4 Specification of the experimental conditions and results

4.1 Observer characteristics

Observers shall be free of any personal involvement with the design of the psychophysical experiment or the generation of, or subject matter depicted by, the test stimuli.