INTERNATIONAL STANDARD

ISO 20462-2

First edition
2005-11-01

Photography — Psychophysical experimental methods for estimating image quality —

Part 2: Triplet comparison method

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

Partie 2: Méthode comparative du triplet

Reference number
ISO 20462-2:2005(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 20462-2 was prepared by Technical Committee ISO/TC 42, *Photography*.

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

— *Part 1: Overview of psychophysical elements*

— *Part 2: Triplet comparison method*

— *Part 3: Quality ruler method*

Introduction

This part of ISO 20462 is necessary to provide a basis for visually assessing photographic image quality in a precise, repeatable and efficient manner. This part of ISO 20462 is needed in order to evaluate various test methods or image processing algorithms that may be used in other international and industry standards. For example, it should be used to perform subjective evaluation of exposure series images from digital cameras as part of the work needed for future revisions of ISO 12232.

The opportunities to create and observe images using different types of hard copy media and soft copy displays have increased significantly with advances in computer-based digital imaging technology. As a result, there is a need to develop requirements for obtaining colour-appearance matches between images produced using various media and display technologies under a variety of viewing conditions. To develop the necessary requirements, organizations, including the CIE and the ICC, are developing methods to compensate for the effect of different viewing conditions, and to map colours optimally across disparate media having different colour gamuts.

Such technical activities are often faced with the need to evaluate proposed methods or algorithms by visual assessment based on psychophysical experiments. K.M. Braun *et al.*[1] examined five viewing techniques for cross-media image comparisons in terms of sensitivity of scaling, and mental and physical stress for the observers. CIE TC1-27 “Specification of Colour Appearance for Reflective Media and Self-Luminous Display Comparisons” proposed guidelines for conducting psychophysical experiments for the evaluation of colorimetric and colour-appearance models[6]. Accordingly, for the design and evaluation of digital imaging systems, it is of great importance to develop a methodology for subjective visual assessment, so that reliable and stable results can be derived with minimum observer stress.

When performing a psychophysical experiment, it is highly desirable to obtain results that are precise and reproducible. In order to derive statistically reliable results, large numbers of observers are required and careful attention should be paid to the experimental setup. Multiple (repeated) assessments are also useful. Observer stress during the visual assessment process can adversely affect the results. The order of image presentation, and the types of questions or questionnaires addressed by the observers, can also affect the results.

Table 1 gives a comparison of three visual assessment techniques commonly used for image quality evaluation. The advantages of the category methods include low stress and high stability, since the observer’s task is to rank each image using typically five or seven categories. However, its scalability within a category is less precise. One of the most common techniques for image quality assessment is the paired comparison method. This method is particularly suited to assessing image quality when precise scalability is required. However, a serious problem with the paired comparison method is that the number of samples to be examined is to be relatively limited. As the number of the samples increases, the number of combinations becomes extensive. This causes excessive observer stress, which can affect the accuracy and repeatability of the results. The third method, commonly known as magnitude scaling, is magnitude estimation. This method is extremely difficult when the psychophysical experiments are conducted using ordinary (non-expert) observers to perform the image quality assessment.

Table 1 — Comparison of typical psychophysical experimental methods

Name of method	Scalability	Stability	Stress
Category	Low	High	Low
Magnitude estimation	Medium	Low	Medium
Paired comparison	High	High	High

G. Johnson *et al*.[3] have proposed "A sharpness rule", where the magnitude of sharpness was analyzed in terms of resolution, contrast, noise and degree of sharpness-enhancement. Likewise, preferred skin colour may be considered not only from the viewpoint of chromaticity, but also with respect to the lightness, background and white point of the display media[4]. These examples show that image quality is not always evaluated by a single attribute, but may vary in combination with multiple attributes. In cases where a psychophysical experiment is designed for a new application, the experimenter may need to vary many attributes simultaneously during the course of the experiment. In these situations, the number of the samples to be examined becomes excessively large, making it difficult to employ the paired comparison technique.

Photography — Psychophysical experimental methods for estimating image quality —

Part 2:
Triplet comparison method

1 Scope

This part of ISO 20462 defines a standard psychophysical experimental method for subjective image quality assessment of soft copy and hard copy still picture images.

2 Terms and definitions

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

2.1
just noticeable difference
JND
stimulus difference that would lead to a 75:25 proportion of responses in a paired comparison task

2.2
psychophysical experimental 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** (2.5), **paired comparison** (2.3) and **triplet comparison methods** (2.4)

2.3
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 too many of the stimulus differences are large enough ($>$ 1,5 JNDs) that they produce saturated responses.

2.4
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

2.5
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.

2.6
observer
individual performing the subjective evaluation task in a psychophysical method

3 Two-step psychophysical method

This part of ISO 20462 defines a new psychophysical experimental method, which satisfies the following requirements:

— enables a large number of samples to be examined;

— provides precise scalability;

— provides low observer stress;

— suitable for ordinary (non-expert) observers;

— provides high repeatability of the results.

The method comprises two steps. The first step is a “category step”, and the second step is a “triplet comparison step” which is newly developed for this purpose.

The reason for applying the “category step” is to reduce the number of the samples to an appropriate number which is determined by the purpose of each experiment. Typically this number is less than 27 samples. Category scaling using three categories, such as “favourable”, “acceptable” and “unacceptable” (or “acceptable”, “just acceptable” and “unacceptable”) is used for the first step, and samples are selected according to the number of samples required in the following step. If the number of test samples examined is relatively small, then the first step should be omitted, and the psychophysical experiment should start directly from the second step.

The second step is conducted in order to derive a precise scaling based on an interval scale. The present proposal is to use a newly developed triplet comparison method. In this method three samples are compared at a time, thereby achieving high assessment accuracy while keeping the experimental scale realistic.

NOTE If the normal paired comparison method were used with 21 samples, a total of 210 combinations would need to be examined. This is time-consuming and imposes excessive stress upon the observers. Furthermore, paired comparison methods require a significant number of observers in order that a precise scaling can be derived. This will result in an experiment that is excessively large and unrealizable.

4 Experimental procedure

4.1 Step 1

Proceed as follows.

a) Prepare the test images to be examined.

b) Observe each sample and rank it into 3 categories; “favourable”, “acceptable” and “unacceptable”.

c) Count the number of test images in each category.

d) Select the samples that will be used in Step 2 (4.2) from the upper category. It is recommended that the number of samples, N, be less than 27 in order to avoid observer stress during the experiment. The number of samples should obey the following equations:

$$N = 6K + 1 \text{ or } N = 6K + 3, \tag{1}$$

where

N is the number of samples;

K is an integer number.

NOTE It is possible to use 5 or 7 categories in the case of many samples.

4.2 Step 2

Proceed as follows.

a) Create combinations of samples for use in the triplet comparison step. Each combination shall consist of three samples. If the total number of the samples selected for the triplet comparison step satisfies Equation (1), then it is possible to arrange each combination of samples such that each pair of samples will only ever be viewed together once during the course of the experiment.

b) Observe the samples and rank them into 5 categories;

- 1: favourable,
- 2: acceptable,
- 3: just acceptable,
- 4: unacceptable, and
- 5: poor.

c) Apply Scheffe’s method for statistical analysis to obtain an interval scale.

NOTE See Annex E.

d) Convert interval scale to JNDs.

NOTE See Annex F.

Annex A
(informative)

Comparison between a paired comparison and a triplet comparison technique

The paired comparison method has traditionally been the most popular psychophysical method, capable of providing a high level of reliability and accuracy. However, the reproducibility of Scheffe's method with assessment scales (variations over repeated assessments) and the stress imposed on observers (due to prolonged assessment time caused by the increase in the number of combinations and fluctuation in the assessment scaling for paired comparison, etc.) have not been fully investigated.

The triplet comparison method has the desirable feature of reducing the level of stress on the observer. This is due to shortened assessment times and is expected to improve assessment accuracy and reproducibility. However, no experiments to validate these advantages have been conducted. Furthermore, the triplet comparison method inevitably yields a level of duplication in comparison for certain sample numbers, and the procedure for determining the minimum number of sample combinations has not yet been established. For various reasons, including those cited above, the triplet comparison method has not been commonly used in general subjective assessment experiments.

A series of experiments were conducted in order to assess the two comparison methods from the following aspects:

a) reproducibility (consistency) in terms of order fluctuation over a number of repeated assessments;

b) accuracy evaluated by the correlation between the orders determined by the two methods;

c) degree of difficulty expressed in terms of the degree of fluctuation for each sample, the necessary assessment time and the difficulty reflected in introspective reports;

d) stress on observers reflected in their introspective reports;

e) comparison of expert observers with naïve observers.

A set of experiments to assess favourable skin colour (tones) using the sample set described in References [5] and [6] was conducted for both comparison methods.

The experiments were repeated five times and the results, which are described in detail in Reference [7] of the Bibliography, are summarized as follows.

— In general the overall trends in assessment made by each method are similar.

— The triplet method can accommodate larger scales of assessment and is capable therefore of separating “favourable” samples from “unfavourable” ones more easily than the paired comparison method when assessment deviation is taken into consideration. A method for analysis that is more in agreement with the objectives of the assessment is therefore expected.

— It was generally noted that the assessment result obtained from the first run of the experiment was unreliable. The standard of the assessment scaling and its stability improved with subsequent repetitions of the experiment.

— The time required for assessment with the triplet method was about 1/3 of that required by the paired comparison method.

In conclusion, the two methods are similar with respect to their consistency and accuracy. The level of stress induced by the triplet method on the observer (due to assessment time) was about one third of the stress induced by the paired comparison method. This indicates that the triplet comparison method has the potential of achieving consistent, accurate (reliable) results while simultaneously reducing the level of stress induced on the observer. The primary aim of the triplet comparison technique is therefore fulfilled.

Annex B
(informative)

Number of sample combinations for triplet comparison

The number of sample combinations for paired comparison, N, is expressed by

$$N = {}_nC_2 = n(n - 1)/2$$

where n is the number of samples and n = 2, 3, 4, 5, etc.

For the method of triplet comparison, if the number of samples selected, n' = 7, 9, 13, 15, 19, 21, 25 and 27, then it is possible to select sample combinations that eliminate the duplication of samples across combinations. More generally, the number of samples, n', can be expressed as:

$$n' = 6k + 1,\ 6k + 3\ (k = 1, 2, 3, 4, 5, 6, \text{etc.})$$

For any value n', the number of sample combinations, N', is calculated as:

$$N' = n'\ (n' - 1)/6$$

Let us place n' points on the circumference of a circle to form an n'-sided regular polygon. Each apex of the polygon, is assigned an integer value 1, 2, 3, . . . , n'. We define the notation whereby (p, q, r) represents a triangle comprising the apices p, q and r, and where the triangle apices represent a combination of samples for the triplet comparison method.

Examples of combinations without duplication are shown in Table B.1. In this table function f is defined as follows;

$$f(i) = 1 + \text{modulo}(i - 1, n')$$

where modulo(i – 1, n') represents the remainder for the division of (i – 1) by n'.

For the case of n' = 7, congruent triangles represented by (1, 2, 4), (2, 3, 5), (3, 4, 6), (4, 5, 7), (5, 6, 1), (6, 7, 2) and (7, 1, 3) give combinations without duplication.

For the case where n' = 6k + 1 and k = 1, 2, 3, 4, etc., combinations without duplication are represented by combining k differently shaped triangles chosen from the n' congruent acute angle triangles and the n' congruent obtuse angle ones.

For n' = 9, combinations without duplication can be achieved by triangles (1, 2, 4), (4, 5, 7), (7, 8, 1), (2, 3, 5), (5, 6, 8), (8, 9, 2), (1, 3, 6), (4, 6, 9), (7, 9, 3), (1, 5, 9), (4, 8, 3) and (7, 2, 6).

For n' = 15, the first thirty combinations are specified as follows: two combinations correspond to the apices of triangles (1, 3, 9) and (1, 2, 5). Twenty-eight combinations are formed when the apices of triangles (1, 3, 9) and (1, 2, 5) are moved to their adjacent apices, respectively, to form a further 14 triangles each. The remaining five combinations are specified by the apices of the five regular triangles (with a side length of 5) that are formed by shifting apex (1) of the regular triangle (1, 6, 11) to 2, 3, 4 and 5 respectively. The sample combinations are obtained without duplication.

For the case where n' = 6k + 3 and k = 2, 3, 4, etc., combinations without duplication are represented by combining k differently shaped triangles chosen from the n' congruent acute angle triangles, the n' congruent obtuse angle triangles and using the n'/3 regular triangles.

