International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.912

(03/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Audiovisual quality in multimedia services

Subjective video quality assessment methods for recognition tasks

Recommendation ITU-T P.912

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	**Series**	**P.900**
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.912

Subjective video quality assessment methods for recognition tasks

Summary

Recommendation ITU-T P.912 defines subjective assessment methods for evaluating the quality of one-way video used for target recognition tasks. "Target" refers to something in the video that the viewer needs to identify (e.g., a face, object, or number). Target recognition video (TRV) is video that is used as a tool in order to accomplish a specific goal through the ability to recognize specific targets of interest in a video stream. TRV can be used in various video services such as surveillance, human identification, licence plate identification, telemedicine, robot control, and remote monitoring and decision-making.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.912	2008-08-13	9	11.1002/1000/9514
2.0	ITU-T P.912	2016-03-15	9	11.1002/1000/12774

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.912

Subjective video quality assessment methods for recognition tasks

1 Scope

This Recommendation defines subjective assessment methods for evaluating the quality of one-way video used for target recognition tasks. "Target" refers to something in the video that the viewer needs to identify (e.g., a face, object, or number). Target recognition video (TRV) is video that is used as a tool in order to accomplish a specific goal through the ability to recognize specific targets of interest in a video stream. TRV can be used in various video services such as surveillance, human identification, licence plate identification, telemedicine, robot control, and remote monitoring and decision-making.

This Recommendation considers three categories of target:

1) human identification (including facial recognition);
2) object identification;
3) alphanumeric identification.

Each of these areas requires specific video test material that spans realistic conditions with stimuli that are carefully chosen to allow multiple scenarios to be created repeatedly with different objects of interest, in different lighting conditions or with small changes in scene details.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.913] Recommendation ITU-T P.913 (2016), *Methods for the subjective assessment of video quality, audio quality and audio-visual quality of Internet video and distribution quality television in any environment*.

3 Definitions

This Recommendation defines the following terms:

3.1 discrimination class (DC): One of four levels of visual discrimination at which the target can be analysed.

- Elements of the action – in a very broad and general sense, identification of the series of events that took place.
- Target presence – recognition or detection of the presence or absence of valid targets.
- Target characteristics – recognition of unique characteristics of the target (e.g., markings, scars, tattoos, dents, colour).
- Target positive recognition – recognition of a specific instance of the target (e.g., recognition of a person, a specific object, or an exact alphanumeric sequence).

3.2 pretest: An experiment run on a small set of subjects to determine any problem with the experiment. The experiment is redesigned based on the pretest data and then the pretest data is discarded.

3.3 scenario group (SG): A collection of scenes of the same basic scenario, with very slight differences between the scenes.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

CPD Cycles Per Degree

DC Discrimination Class

PVS Processed Video Sequence

SG Scenario Group

TRV Target Recognition Video

5 Source signal

Test sequences should follow the general principles stated in [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate) and [b-ATIS-0100801.01], which specify that scenes should be consistent with the transmission service under test, and should span the full range of spatial and temporal information. It is critical for the nature of these evaluations that the stimuli used actually reflect the true operational parameters of the conditions under which the video material is collected. If the stimuli used cannot actually cover the entire range of scenarios possible for the application area for identification, the application description has to be explicitly limited. For example, the results should not be generalized. Unlike other subjective assessment methods developed for quality evaluations, this method is directed at the usefulness of the video material to complete a task and not the quality of the video itself.

6 Test methods and experimental design

For video that is used to perform a specific task, it may not be appropriate to rate the quality of the video according to a subjective scale, such as the absolute category rating (ACR) [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate). The goal of test methods for TRV is to assess the ability of a viewer to recognize the appropriate information in the video, regardless of the viewer's perceived quality of the viewing experience. To assess the quality level of TRV, methods that reduce subjective factors and measure the ability of a participant to perform a task are useful in that they avoid ambiguity and personal preference.

The application of TRV is directly related to the ability of the user to recognize targets at increasing levels of detail. These levels are referred to as discrimination classes (DCs). When determining the DC for particular scenarios, it needs to be considered that, for a set distance from the camera to the object of interest, the DC directly correlates to decreasing video resolution of the target, and therefore the object is represented by fewer cycles per degree (CPD) of resolution. Fewer CPD of resolution also mean that the object subtends less of the information content of the video, making identification of the target more difficult.

CPD, the key parameter, is affected by the resolution of the object and (potentially) the distance between the camera and the object [b-Leszczuk, 2011]. Consequently, it relates to the achievable DC.

Examples of the achievable DC are shown in Figure 1. If the distance between the camera and the object is 50 m, "Target Positive Recognition" is possible; for 215 m "Target Characteristics"; but for 430 m only "Target Presence".

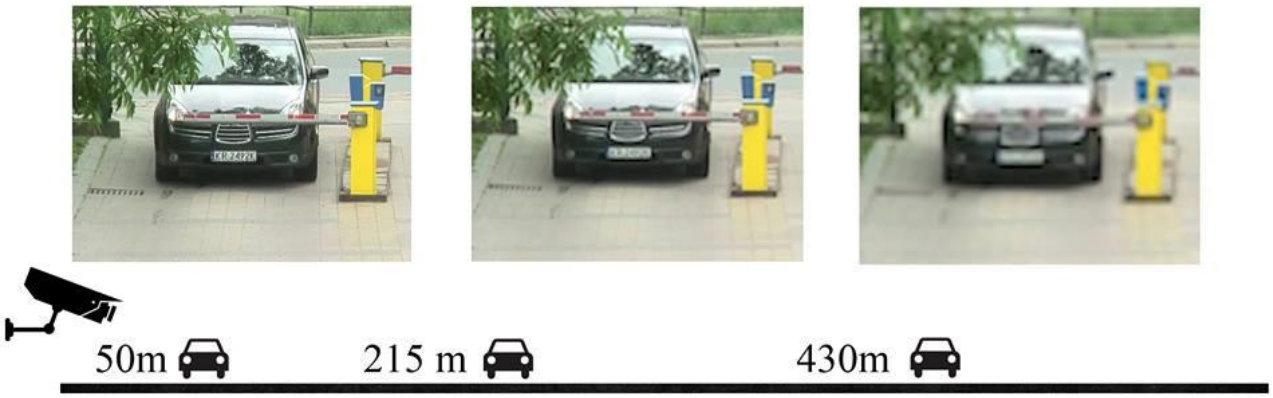

Figure 1 – DC in testing methods for various distances between camera and object

Experimental methods should consist of responding to questions relating to the content in the image or video. The parameter addressed by the question is the target to be recognized.

6.1 Multiple choice method

This method is appropriate for all DC levels and target categories (human, object and alphanumeric). For this method, the video is shown above a list of verbal labels representing the possible answers. After presenting the video, the viewers must choose the label closest to what they recognized in the clip. The use of fixed multiple choices eliminates any possible ambiguity that could arise from open questions, and allows for more accurate measurements.

The number of choices offered to the viewer depends on the number of alternative scenes being presented. The use of "Unsure" as one of the listed choices is discouraged, but allowed. The experimenter should be aware that individual subjects tend to overuse the "Unsure" choice, leading to contamination of results. Consequently, special care must be taken when "Unsure" is one of the listed choices.

An example of the test screen a viewer would see is shown in Figure 2.

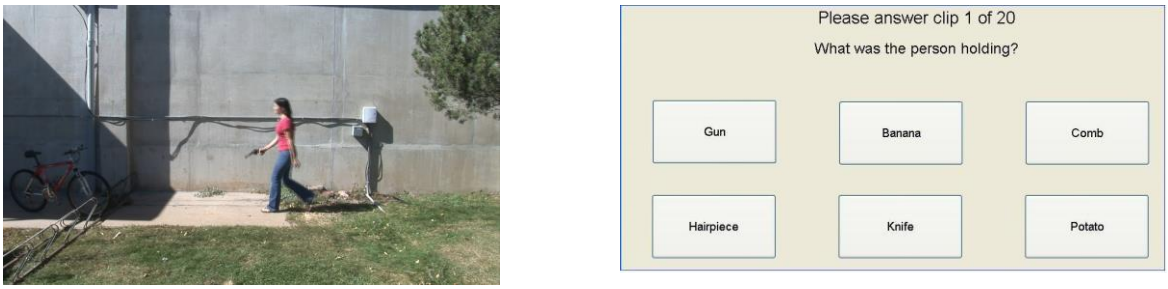

Figure 2 – Multiple choice method for recognition of objects in a video clip: test screen and list of possible visual recognition responses

6.2 Single answer method

If there is an unambiguous answer to an identification question, the single answer method may be used. This method is appropriate for alphanumeric character recognition scenarios. A viewer is asked which letter(s) or number(s) was/were present in a specific area of the video, and the answer can be

evaluated as either correct or incorrect. Alternatively, fuzzy logic may be used (e.g., Hamming distance or Levensthein distance), as seen in [b-Leszczuk, 2014].

Yes or No tests also fall under this method. A viewer may be asked if a certain object was present in the clip, for example. In this method, it is important to ensure that the procedure used to gather viewer responses is easy to understand, so that the test interface does not distract from the cognitive processing required for actual identification of the alphanumeric characters or object. Care must also be taken to avoid terminology that can differ from participant to participant.

The use of "Unsure" as the third possible answer is discouraged, but allowed. The experimenter should be aware that individual subjects tend to overuse the "Unsure" choice , leading to contamination of results. Consequently, special care must be taken when "Unsure" is one of the listed choices.

An example of an alphanumeric single answer viewer screen is shown in Figure 3.

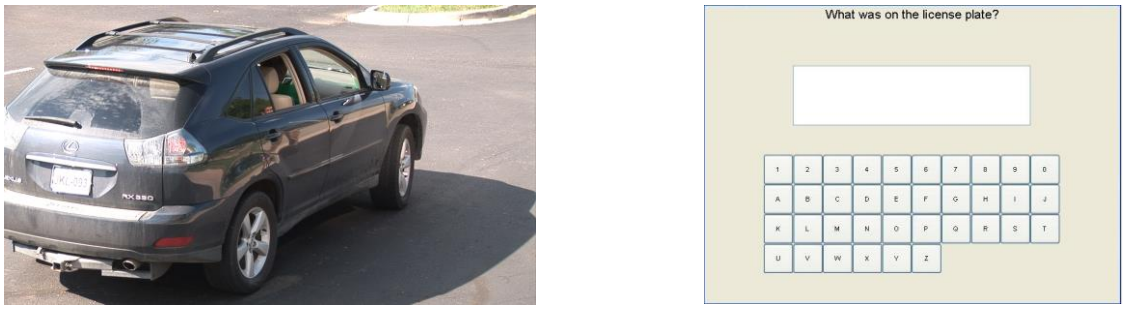

Figure 3 – Single answer method for unambiguous responses to video object recognition: test screen and keyboard for entry of alphanumeric character recognition

6.3 Timed task method

A viewer may be asked to watch for a particular action or object to be recognized in the video clip. When the viewer perceives that the target has occurred, a timer button can be pushed. In the timed task, the experimenter is able to determine whether the time falls within an acceptable time-frame for decision-making. These time-frames are defined by the field in which the video is used, e.g., a person responding to a riot who needs to identify whether crowd members have real weapons versus a person who is chasing a car and needs to read the licence plate.

6.4 Real-time vs. viewer-controlled viewing

Depending on the nature of the task, TRV test methods can be used, either in real time, without the ability to freeze or rewind, or they can be used for non real-time analysis. The experiment should mimic the real world application of the video. If the intended use of the video is analysis, the subject under test should have the capability to control the playing of the test clip.

6.5 Scenes

Since TRV is generally used to perform a recognition task, the scenes should contain targets consistent with the application under study. However, because the measurements are focused on a subject's ability to identify objects and actions, the possibility that a viewer may memorize the scene content and use other visual clues to remember the identity of the target must be addressed. Therefore, an individual scene may be replaced by a set of scenes containing multiple versions, with controlled differences between the versions. This is called a scenario group (SG). For example, the scenario could be that a person walks across the field of view carrying an object. The SG would consist of multiple shots using different objects or different people. The number of scenes in a SG should be large enough so that scene memorization is unlikely. An example of three scenes from one scenario group is shown in Figure 4. The scene content is almost identical except for the single change in the object being held.

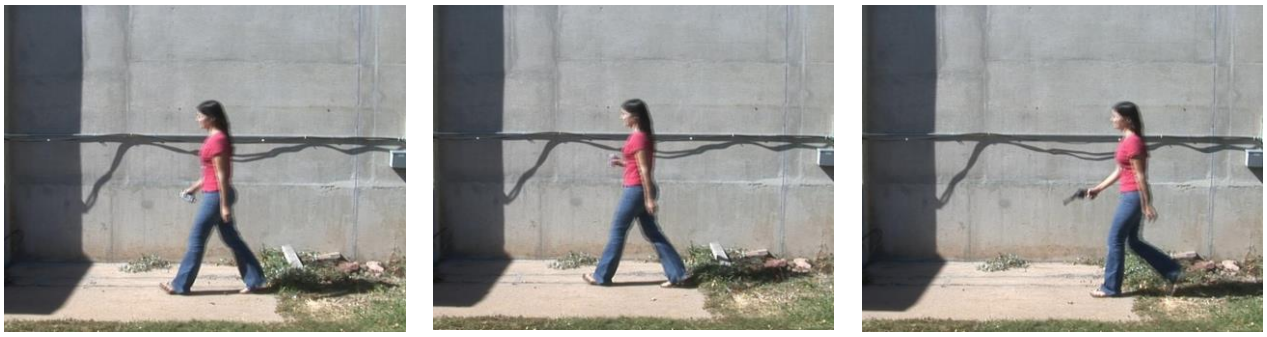

Figure 4 – A scenario group of images: scene 1, scene 2, scene 3

The content of the scenes should be determined by experts in the application for which the video is used. These experts should identify critical *tasks*, critical *scenes* in which these tasks are accomplished and critical *parameters* of the scenes. These parameters are used in the design of the experiment to create the set of multiple choice answers. The scenes should be created in a way that the parameters of interest appear in the video at a resolution that can be realistically expected in practice; i.e., the parameters should occupy a realistic percentage of the field of view.

6.6 Experimental design

The experimenter should follow the guidelines outlined in [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate).

6.7 Reference conditions

The experimenter should follow the guidelines outlined in [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate).

7 Evaluation procedures

In clause 7, a laboratory test is described. Description of a crowdsourcing environment is described in Appendix I.

7.1 Viewing and listening conditions

The experimenter should follow the guidelines outlined in [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate).

7.2 Processing and playback system

The experimenter should follow the guidelines outlined in [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate).

7.3 Subjects

Subjects who are experts in the TRV application field should be used. For certain areas of application testing, where neither specific experience nor expertise is required, use of non-expert subjects is also permissible. Such non-experts must be motivated in other than a professional manner (e.g., they are paid). The validity of this approach is shown in [b-Leszczuk, 2012]. The number of subjects should follow the recommendations of [ITU-T P.913] (or [ITU-T P.910] for those situations where it is more appropriate).

7.4 Instructions to subjects and training session

The subject should be given the context of the task before the video clip is played, and told what they are looking for or trying to accomplish. If questions are to be answered about the content of the video, the questions should be posed before the video is shown, so that the viewer knows what the task is.

It is safe to assume that there are no easy tasks. Even something as easy as recognizing a character must be described in detail. This means the instructions must clearly state what subjects must do if:

1 they cannot recognize a character;

2 they have doubts;

3 they can recognize some, but not all, characters.

The optimal training session must show all specific cases and the correct scoring behaviour (i.e., that desired by the experiment design).

Especially difficult is to define a task for specialists, e.g., medical doctors. In this case, the running of a pretest on a small group before running any larger experiment is strongly recommended. A typical number of subjects for a pretest is approximately 20% of the total. A pretest group can consist of a single person. Specialists often change the task so that it fits a real situation typical for a particular specialist better. This can change the experimental conditions and finally harm the experiment itself. Therefore, it is very important in the pretest to clearly explain the task, the reason for running the test and the reasons why the experiment has been set up in a particular way. Feedback from the pretest is used to improve the experiment before running it with actual subjects.

8 Statistical analysis and reporting of results

The first step of the analysis is subject screening to eliminate those who did not pay attention or who did not understand the task. Further statistical analyses vary slightly depending on the scoring method.

8.1 Subject screening

This technique is optional.

In order to detect abnormal subjects, it is not enough to compare the results obtained by one subject to the average obtained in the experiment, since in a typical experiment different subjects perform different tasks (see clause 6.5). Even with careful design, the tasks performed by one subject can be more difficult than the tasks performed by the average subject. An algorithm for solving the problem of different task difficulty performed by different subjects is proposed in [b-Janowski, 2012].

The algorithm proposed assumes that tasks can be partially ordered. For example, consider an experiment where the goal is to specify detection probability as a function of the compression bitrate. The processed video sequences (PVSs) obtained for the same source and lower bitrate are likely to

have less information and likely the detection is at least not easier than for a higher bitrate. Also if an object covers fewer pixels on a screen, it is not easier to detect.

Based on the above assumption, consider a list of tasks $T_1^{(l)}, \dots, T_{n_l}^{(l)}$, which are ordered by difficulty. The list starts from the easiest, $T_1^{(l)}$, to the most difficult task, $T_{n_l}^{(l)}$. If a subject did not manage task $T_j^{(l)}$ and some tasks $T_{j+k}^{(l)}$ were correctly performed, this indicates that a subject did not pay enough attention or that another problem occurred. Moreover, if a subject did not manage task $T_j^{(l)}$ and another subject correctly solved task $T_{j+1}^{(l)}$, the error is less serious compared with the situation where task $T_{j+10}^{(l)}$ was solved correctly. Therefore, the penalty for a subject from not solving a task that was solved by another subject must be a function of the difference in task difficulty. The final proposed equation is:

$$ssq_{i,j,l} = (1 - r(i,j,l)) \sum_{m=1}^{M} \sum_{k=j+1}^{n_l} (k-j) r(m,k,l) \quad (1)$$

where $r(i,j,l)$ is 0 if the task $T_j^{(l)}$ was performed incorrectly by subject i and 1 otherwise, and M is total number of subjects.

The result of equation (1) is 0 if task j is performed correctly by subject i. It is also 0 if all more difficult tasks were performed incorrectly by all subjects. A higher value indicates that more subjects or more difficult tasks were done correctly.

The final value obtained for user i is:

$$Sq_i = \sum_{l=1}^{L} \sum_{j=1}^{n_l} ssq_{i,j,l} \quad (2)$$

where L is the total number of partially ordered groups.

A large value of Sq_i indicates that a subject is irrelevant or that the task conditions and results must be double checked. For example, Sq_i detects a subject in one experiment who confused the terms "radio" and "mobile phone". The value of Sq_i depends on the experiment size and the length of the partial sorted groups within the experiment, therefore a threshold value cannot be specified. It must be used as an indicator for further investigation or, if no outliers are detected, proving that all subjects behave correctly and the results can be analysed further.

Show Sq_i distribution when the results are reported.

8.2 Further statistical analysis

The statistical analysis for each method varies slightly.

For all conditions, a correlation and understanding of the number of CPD or area subtended of the target are taken into consideration to determine the correlation between success and CPD.

For cases where there are multiple answers, a statistical validity indicator is required.

8.2.1 Multiple choice

For multiple-choice answers, the probability of an incorrect answer needs to be balanced against the ability to answer the questions correctly. The statistical metric in this situation requires an examination of the stability of the answers within and between subject performance metrics. "Unsure" answers should be pooled with those that are incorrect.

8.2.1.1 Recognition probability as a function

For multiple choice, estimate the probability of correct answer as a function of independent variables, like bitrate or camera quality. Probability as a function of independent variables can be calculated by logistic regression [b-Agresti, 2002]. Logistic regression can be found in almost all statistical packages. The main difference between logistic and linear regression is that instead of modelling the

response variable (i.e., 1=correct and 0=incorrect), logistic regression models the probability that the response variable has particular value. The simplest logistic regression is given by:

$$\pi(x) = \frac{\exp(ax+b)}{1+\exp(ax+b)} \quad (3)$$

where $\pi(x) = P(Y = 1|X = x)$ i.e., the probability of a correct answer, in which the explanatory variable is x; a and b are model parameters.

Much more complicated parameters and explanatory variables can be considered, e.g., if there are two explanatory variables x and y, a logit model could be:

$$\pi(x) = \frac{\exp(ax+by+cxy+d)}{1+\exp(ax+by+cxy+d)} \quad (4)$$

Note that for such model both x and y must be normalized, such that the unit change of x is similar (i.e., has a similar influence on probability) to the unit change of y.

If answer "Unsure" is used, it is analysed in the same was as any other answer.

8.2.1.2 Comparing different conditions

A function showing the probability of correct recognition as a function of specific parameters is one way of analysing multiple-choice data. The second important way of analysing multiple-choice data is to compare two or more different conditions. For each condition, the numbers of correct and incorrect recognitions are collected. The goal is to determine whether the observed difference is statistically significant.

The collected data can be represented by a matrix. Assuming that there are k different conditions, the matrix has the form shown in Table 1.

Table 1 – Collected data matrix representation

Condition	Correct	Incorrect	Sum
A_1	n_{11}	n_{10}	$n_{1\cdot}$
A_2	n_{21}	n_{20}	$n_{2\cdot}$
…	…	…	…
A_k	n_{k1}	n_{k0}	$n_{k\cdot}$
Sum	$n_{\cdot 1}$	$n_{\cdot 0}$	N

In order to answer the question whether all conditions are statistically the same, the χ^2-test is needed.[1] The test is performed by comparing the detection probability obtained for each condition with the overall detection probability. The overall detection probability is calculated to be

$$p = \frac{\sum_{i=1}^{k} n_{i1}}{N} \quad (5)$$

where $N = \sum_{i=1}^{k}(n_{i1} + n_{i0}) = \sum_{i=1}^{k} n_{i\cdot}$ and k represents the number of conditions. The χ^2-test is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_{i0} - n_{i\cdot}(1-p)\right)^2}{n_{i\cdot}(1-p)} + \sum_{i=1}^{k} \frac{(n_{i1} - n_{i\cdot}p)^2}{n_{i\cdot}p} \quad (6)$$

Commonly in experiments, the number of answers given for each condition, denoted here by $n_{i\cdot}$, is the same value; denote it n. In such a case, equation (6) simplifies to:

[1] Note that χ^2 test does not answer which condition is different, it only answers if all of them are the same or not. Comparing some conditions is described later in this clause.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_{i0} - n(1-p))^2}{n(1-p)} + \sum_{i=1}^{k} \frac{(n_{i1} - np)^2}{np} \quad (7)$$

The value obtained is compared using the χ^2-distribution with $k - 1$ degrees of freedom. If the value obtained is greater than the value of the χ^2-distribution calculated for a specific significance level (typically 0.05), the hypothesis that all conditions are the same is rejected.

After comparing many conditions, comparison of a subset of these conditions can be of interest. In this case, the significance level has to be correctly set. The reason why this has to be done can be explained in an example. Assume that 100 conditions are considered. The test shows that there is no statistical difference. Nevertheless, if the first condition is compared with all other conditions, statistically five of them could be statistically significantly different. The simplest way to correct the significance level is to use the Bonferroni correction, whose formula is:

$$\alpha_{cor} = \alpha / N \quad (8)$$

where α_{cor} is the corrected significance level, α is the significance level used to compare the group, and N is the total number of comparisons, typically $N = k(k-1)$, where k is the number of conditions considered.

8.2.2 Single answer

For single answer conditions, where answers are either correct or incorrect, a statistical metric to determine whether the subject performs above the random chance of answering correctly should be implemented. "Unsure" answers should be pooled with incorrect answers.

For a single answer, the correctness of the answer can be analysed on a different scale. The simplest scale is 0-1 correct/incorrect. The correctness threshold can be different depending on the specific analysis. Since the final results are of the 0-1 type, the results obtained are similar to those for the multiple-choice case and the same analytical tools must be used.

If the correctness of the answer is analysed, different models can be used. It is difficult to describe all options since the answer can be very different depending on the answer type. Most probably, correctness can be analysed by the generalized linear model described in [b-Agresti, 2002].

8.2.3 Timed task

For timed tasks, the statistical analysis should incorporate two metrics that are finally correlated against each other to understand the impact of correctness versus time taken to perform the task.

The timed factor is a straight average of time to identify the object, that is then weighted against the correctness of the answer. For the correctness factor, the same statistical analysis as for single answer conditions is also applied.

For timed tasks, the statistical analysis must incorporate time as an explanatory variable. Time can be a numerical value "how long it took to finish the task, in seconds" or it could be "number of replays of the movie before a decision was made." The analysis must indicate the influence of time on the result obtained.

Appendix I

Crowdsourcing Environment

(This appendix does not form an integral part of this Recommendation.)

I.1 Introduction

One of the main problems of recognition tasks is the obvious limitation of source sequences reuse as described in clause 6.5. The best way to protect against source sequence remembrance is to prevent the showing of the same source sequence to the same subject more than once. Nevertheless, such a solution has an obvious drawback: it requires a much larger number of subjects. For laboratory tests, it is difficult to achieve a sufficient number of subjects. A natural solution is crowdsourcing, which gives access to thousands of potential subjects at the same time.

The advantage of accessing large number of subjects comes with a price of lack of control over the subjects and environment. This Recommendation describes possible ways to increase the accuracy of results obtained and is based on a white paper of which more details can be found in [b-Hossfeld, 2014].

I.2 Definitions

I.2.1 crowdsourcing: Obtaining the needed service by a large group of people, most probably an on-line community.

I.2.2 test: Subjective assessments in a crowdsourcing environment.

NOTE – I.2.2 follows terminology presented in [b-Hossfeld, 2014].

I.2.3 worker: Person participating in a crowdsourcing test.

NOTE – I.2.3 follows terminology presented in [b-Hossfeld, 2014].

I.2.4 task: Set of actions that a worker needs to perform to complete a subscribed part of the test.

NOTE – I.2.4 follows terminology presented in [b-Hossfeld, 2014].

I.2.5 question: A single event that requires an answer for a worker. A task contains many questions.

I.2.6 campaign:A group of similar tasks. It also contains a more detailed description of the part of the test that is under investigation, like workers' payment, and indicates subjective assessments in a crowdsourcing environment. A test can contain multiple campaigns.

NOTE – I.2.6 follows terminology presented in [b-Hossfeld, 2014].

An example of using definitions I.2.1 to I.2.6 is as follows. A research question calls for a subjective experiment. It was decided to run the experiment as a crowdsourcing test. The test goal is to answer the given question. This test can be divided into multiple campaigns, namely preliminary, main and corrected. Each campaign contains multiple questions that have to be answered. Those questions are grouped into tasks. Each task is assigned to a single worker, but a single question can be asked in many different tasks, also some tasks can be identical, although they are performed by different workers.

I.3 Software

In order to be able to run a crowdsourced test, a worker has to have access to the test environment. Implementation of the test as a web service, which than can be easily accessed by anyone with an Internet connection, is advised. Of course, other solutions, like software code or an application, can also be used, but the number of workers willing to install additional software, compared to those willing to access a specific web page, is much smaller. Even the use of a very specific web browser plugin can restrict the number of participants significantly.

Regardless of the test software, it is important to include a feedback channel, which can help in detecting errors or improving the test.

I.4 Designing a task

The task preparation should take into account all lessons learned from any laboratory study, if such studies were conducted. Any additional questions asked by subjects should be addressed. Note that a worker cannot ask an additional question or at least it is not easy to do. Therefore, all problems should be solved before the task is sent to the workers.

Also the task itself has to be easy. Any question asked should be tested against any misinterpretation. Consultation with non-native speakers to ask their opinion is a good idea, since it is probable that some workers do not speak English well. For the same reason, use simple English in all descriptions, questions and messages presented to workers. If possible, enrich the text with pictures. For example, if the task is to recognize an object, it is recommended that pictures of the object be added, not only written descriptions.

It is important that the task be short, so that a single worker does not take a long time to complete it. The fact that the task is performed in a home environment needs to be taken into account. A long task can be easily interrupted by external events, like a phone call. A shorter task is more likely to be finished in a single session. A short task requires that detailed description be limited. This is also in line with the simplification of the description. It is also common to collect social information, which also should be limited to the most important questions. The answering interface should be prepared in such way that it can be filled in fast.

The task should be short, but training cannot be skipped. It is very important for a worker to know that this question is just a training so he can explore the interface. Since each worker supplies very few answers, it is important to ensure that none of them are lost due to misunderstanding the interface.

Even a correctly design interface can be problematic for some workers, and others can simply cheat or answer randomly. Therefore, unreliable workers need to be detected. Detecting unreliable workers should be included in the experimental design by adding specific questions. In a recognition test, extending the test by adding obvious questions is a good idea. Also, some repeated questions that detect random clickers are needed. An interesting idea is presented in [b-Gardlo, 2014] where an additional monitor test is used. Such a test has two advantages: the first is the detection of random clickers and the second is the detection of workers with very bad screens or lighting conditions.

A task for a worker is similar to a typical subjective experiment for a subject. Nevertheless, the test software has to log much more information than software used in the laboratory study. In general, all information available should be recorded. Among required parameters are: response time; browser type and version; operating system; and screen resolution. This information is crucial for detecting unreliable workers. The most important step after data compilation is finished and before the main conclusions are taken is to clear the data of answers given by unreliable workers. Explain the plan carefully in advance and specify how detection is to be done. The detection method should be based on more data than just the answers to the main research questions. The logs created in the experiment design should be used.

I.5 Distribution of the campaign

After creating the test platform, distribute it among subjects. There are two main ways to advertise a specific campaign.

1 Using social media and mailing lists

   a Advantages:

      i it is possible to get to specific group, e.g., policemen;

      ii quite often it does not include additional costs;

iii workers willing to make a task for free are most of the time honest.

b Disadvantages:

i the mailing list or social media generate(s) a very specific (probably biased) group of workers;

ii since no payment is made for the task, a large number of tests will not be completed, unless the test is extremely short or involves gamification;

iii the speed of collecting the data is, most of the time, very rapid just after announcement, but falls away rapidly, meaning that the web server can be overloaded;

iv it is difficult to predict how many answers will be collected;

v checking whether an individual ran the task once only is difficult.

2 Using specific services (called crowdsource platforms) gathering people willing to make micro tasks

a Advantages:

i the speed of collecting the data can be adjusted;

ii the task is advertised constantly by the service;

iii a large number of data can be collected in a short period of time.

b Disadvantages:

i some workers will use the test just to get money and their answers are random;

ii every answer, even those given by workers answering randomly, costs some money;

iii workers are pooled from a specific group of people willing to make money by doing micro tasks.

I.6 Data analysis

Even with careful subject validation, assume that subjects are different. Since each sequence is validated by a diverse subgroup of subjects, the difference in recognition probability can be characterized only by subgroup of subjects, not by difference in conditions. Nevertheless, results presented in [b-Korshunov, 2012] show high correlation between results obtained in a laboratory environment and those from crowdsourcing. Such a result is possible only after removing unreliable subjects.

Bibliography

[b-ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

[b-ITU-T P.920] Recommendation ITU-T P.920 (2000), *Interactive test methods for audiovisual communications*.

[ITU-R BT.500-13] Recommendation ITU-R BT.500-13 (2012), *Methodology for the subjective assessment of the quality of television pictures*.

[b-ATIS-0100801.01] ATIS-0100801.01.1995 (R2011), *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*

[b-Agresti, 2002] Agresti, A., *Categorical data analysis*, 2nd edition (2002), Hoboken, NJ: Wiley.

[b-Gardlo, 2014] Gardlo, B.; Egger, S.; Seufert, M.; Schatz, R. (2014), Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. In: *IEEE International Conference on Communications (ICC)*, pp.1070-1075, 10-14 June.

[b-Hossfeld, 2014] Hoßfeld, T.; Hirth, M.; Redi, J.; Mazza, F.; Korshunov, P.; Naderi, B.; Seufert, M.; Gardlo, B.; Egger, S.; Keimel, C. (2014), *Best practices and recommendations for crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing*, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003 Qualinet).

[b-Janowski, 2012] Janowski, L. (2012), Task-based subject validation: Reliability metrics. In: *Fourth International Workshop on Quality of Multimedia Experience* (QoMEX), pp. 182-187, 5-7 July.

[b-Leszczuk, 2011] Leszczuk M., Janowski L., Romaniak P., Głowacz A., Mirek R. (2011), Quality assessment for a licence plate recognition task based on a video streamed in limited networking conditions, In: *Proceedings of the Multimedia Communications, Services and Security*, Poland, June, Springer. <http://link.springer.com/chapter/10.1007/978-3-642-21512-4_2>

[b-Leszczuk, 2012] Leszczuk, M., Koń, A., Dumke, J., Janowski, L. (2012), Redefining the ITU-T recommendation P.912 for requirements of the subjects of quality assessments in recognition tasks, In: *Proceedings of the 5th International Conference on Multimedia Communications, Services and Security*, Krakow, Poland, June, Springer CCIS Vol. 287. <http://link.springer.com/chapter/10.1007%2F978-3-642-30721-8_19>

[b-Leszczuk, 2014] Leszczuk, M. (2014), Optimising task-based video quality: A journey from subjective psychophysical experiments to objective quality optimisation, In: *Multimedia Tools and Applications*, United States, January, Springer Vol. 68. <http://link.springer.com/article/10.1007%2Fs11042-012-1161-6>

[b-Moyer] Moyer, S., Hixson, J., Edwards, T., and Krapels, K. (2004), *Identification of Small Handheld Objects for Electro-Optic/FLIR Applications*, Optical Engineering. P. 0632201-1 – 12. <http://adsabs.harvard.edu/abs/2004SPIE.5407..116M>

[b-O'Connor] O'Connor, J., O'Kane, B., Ayscue, K., Bonzo, D., and Nystrom, B. (1998), *Recognition of human activities using handheld thermal systems*, Proceedings of the SPIE Conference on Sensor Technology for Soldier Systems, Orlando, Florida, April, SPIE Vol. 3394.
<http://adsabs.harvard.edu/abs/1998SPIE.3394...51O>

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2016