 doi:10.20944/preprints202101.0435.v1

Article

New Recommendation for Subjective Video Quality Assessment Methods for Recognition Tasks

Mikołaj Leszczuk [1]*, Lucjan Janowski [1]

[1] AGH University of Science and Technology, Krakow MP 30059, Poland; vq@kt.agh.edu.pl
* Correspondence: leszczuk@agh.edu.pl; Tel.: +48-12-617-3599

Abstract: It was once thought that high QoS (Quality of Service) performance solves recurrent problems of low-quality multimedia services. Since then, solutions have been proposed to ensure a high level of QoE (Quality of Experience). In this document, the authors attempt to outline their understanding of an accurate meaning of quality of multimedia services. Starting from QoS and passing through generalised QoE, the authors focus on aspects of subjective and objective quality modelling and optimisation of visual performance for TRV (Target Recognition Video) applications (such as video surveillance), outlining the path of ITU-T standardisation in this area. The authors revised the ITU-T Recommendation P.912 to reflect improved subjective test techniques developed since this Recommendation was approved. The authors also attempt to predict at least some existing errors of reasoning, which are likely to become evident for the industry in the next decade.

Keywords: QoS; QoE; ITU-T; TRV; P.912; CCTV

1. Introduction

Decades ago, the telecommunications industry believed that high-performance Quality of Service (QoS) techniques resolve any recurrent problems of low-quality multimedia services. However, within some years, it became clear that optimisation of QoS parameters such as throughput, packet loss, delay, or jitter is not the best way of improving the quality experienced by users. More efficient codecs can compensate for the problem of low bandwidth. The impact of packet loss is strongly dependent on their distribution and redundancy coding and transmission. For many applications, buffering multimedia data streams can alleviate significant delays and jitter.

Since discovering that QoS is not a useful metric of network quality, most proposals suggest that quality should be measured on the user level. This process was named Quality of Experience (QoE) [1,2]. Such measurement calls for unique structures (frameworks) of integrated assessment of the quality of video sequences [3]. These structures are increasingly being filled with solutions that attempt to model the overall quality, operating at the intersection of QoS and QoE [4] or only in QoE. However, it has become apparent that such a general approach simply does not work for many visual applications such as target recognition (utility) applications (video surveillance, telemedicine/remote diagnostics, fire safety, backup cameras, games) [5,6].

QoE – the way of perceived quality of multimedia services – depends on several objective and subjective contextual parameters [7]. Only a full understanding, usually only possible with definite limitations of the QoE modelling application area, makes it possible to obtain results consistent with the expectations of service users and, consequently, to optimise quality [8]. Unfortunately, high numbers of contextual parameters mean this research question is still open.

2. Target Recognition Video

In many visual applications, the motion picture's quality is not as important as the ability of the visual system to perform specific tasks for which it is created, given the processed video sequences. Such sequences are called Target Recognition Video (TRV). Regardless of how the concept of TRV quality is understood, its verification is necessary to perform dedicated quality testing. These tests'

 doi:10.20944/preprints202101.0435.v1

basic premise is to find TRV quality limits for which the task can be performed with the desired probability or accuracy.

Such tests are usually subjective tests (psychophysical experiments) with a group of subjects. Unfortunately, due to the complexity of the issue and our relatively low degree of understanding of human cognitive mechanisms, satisfactory computer modelling results of TRV quality have not yet been achieved beyond minimal areas of application.

Given the use of TRV, qualitative tests do not focus on the subject's satisfaction with the video sequence's quality, but instead, they measure how the subject uses TRV to accomplish certain tasks. Purposes of this may include:

- Video surveillance – recognition of vehicle licence plate numbers
- Telemedicine/remote diagnostics – correct diagnosis
- Fire safety – fire detection
- Backup cameras – parking the car
- Games – spotting and correctly reacting to a virtual enemy

The human factor is a significant influence. Therefore it is necessary to ask questions on the procedures to be complied with to make a subjective assessment of TRV quality. In particular, questions arise on:

- Method of selecting the TRV source from which the test TRV (with degraded quality) arises
- Subjective testing methods and the general manner of conducting the psychophysical experiment
- Method of selecting a group of subjects in the psychophysical experiment, especially identification of any prior knowledge of the task
- Training subjects before the start of the experiment
- Conditions in which the test will be carried out
- Methods of statistical analysis and presentation of results

3. Methods for Subjective Evaluation of TRV

The International Telecommunication Union (ITU; French: *Union Internationale des Télécommunications* or UIT), is a specialised agency of the United Nations responsible for all matters related to information and communication technologies. The ITU comprises three sectors, each managing a different aspect of the matters handled by the Union and ITU Telecom. The ITU Telecommunication Standardisation Sector (ITU-T) coordinates standards for telecommunications and Information Communication Technology. ITU-T Recommendations are the names given to telecommunications and computer protocol specification documents published by the ITU-T. Series "P" of ITU-T Recommendations describe telephone transmission quality, telephone installations, and local line networks. ITU-T P.900-P.999 series describe audiovisual quality in multimedia services.

Recommendation ITU-T P.912 [9] "Subjective Video Quality Assessment Methods for Recognition Tasks" addresses questions formulated in the previous section. 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). Recommendation P.912 organises terminology related to subjective TRV testing, introducing appropriate definitions for testing methods (psychophysical experiments).

The initial version (there was just one version so far) of Recommendation P.912 – "P.912 (08/08)" – has been published in August 2008.

Unfortunately, P.912 (08/08) was only the first step in standardising subjective TRV testing methods. In the opinion of the authors, based on research results (their own and independent) and observations conducted during numerous experiments with TRV, many claims of P.912 (08/08) were formulated at too high a level of generality. What is more, selected statements are not supported by research results and are significantly disputable. The P.912 (08/08) has not been adopted widely for use across industry or academia.

 doi:10.20944/preprints202101.0435.v1

In this situation, the authors have taken steps to introduce significant modifications (amendments) to P.912 (08/08). For this purpose, in order to formalise the procedures, the authors have established a collaboration with the Polish Ministry of Administration and Digitisation. The authors have already received a formal nomination as a delegate of the Polish government. The procedure for submitting amendments has been already finalised.

The areas in which the authors have made suggestions relating to the source signal, testing methods and experimental design, evaluation procedures, statistical analysis and reporting of results as well as crowd-sourcing environment.

The detailed scope of the amendments to P.912 (08/08) is discussed in [10] and the following sections.

4. Source Signal

In Clause 5, P.912 (08/08) stated:

Test sequences should follow the general principles stated in [11]. Moreover, [12], which specify that scenes should be consistent with the transmission service under test and span the full range of spatial and temporal information. It is critical for these evaluations that the stimuli used to reflect the actual operational parameters of the conditions under which the video material is collected, and cover the entire range of scenarios possible for the application area that one is identifying. 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 video's quality.

Unfortunately, in some instances, data availability is very limited in practice. Let us consider the impact of studying the quality of still images on the accuracy of X-ray bone fractures' diagnosis. It is clear that due to the low frequency of certain types of fractures, the availability of a database of corresponding images is deficient.

Another example concerns research on the impact of CCTV recordings on the accuracy of licence plate recognition [8]. For this study, a unique video database was created [13]. The recordings have been created using fixed CCTV cameras, recording cars entering the car park at the AGH University of Science and Technology in Kraków, Małopolska, Poland (Fig. 1). Again, it is clear that due to the conditions mentioned above of acquisition, recordings represent a particular CCTV camera, its specific location and direction, a specific distance from the object, and specific lighting conditions. What is more, since the recordings were made in Kraków, most of the licence plates have the letter "K" (distinguishing the province of Małopolska) in the first position on the plate and "R" (distinguishing the county) in the second position.

As shown, contrary to P.912 (08/08), it was tough to ensure complete coverage of the potential applications of the recordings. Any expansion of the record database was laborious, time-consuming, or even impossible. It did not mean that the cited studies were useless; however, their applicability must have been explicitly limited to the scope of the recordings' database. Unfortunately, literature frequently included attempts to extrapolate the applicability of test results (particularly among less experienced researchers), which the authors believe may have been because issues in P.912 (08/08), which frequently included instructions to carry out tests, are not addressed explicitly.

The authors revised Clause 5 of P.912 (08/08) with the following amendments:

Test sequences should follow the general principles stated in [11] and [12], 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 these evaluations that the stimuli used to reflect the actual operational parameters of the conditions under which the video material is collected. ***If the stimuli used cannot cover*** *the entire range of scenarios possible for the application area that one is identifying,* ***the application description needs to be explicitly limited. For example, the results should not be generalised****. 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 video's quality.*

 doi:10.20944/preprints202101.0435.v1

Figure 1. Source signal

5. Testing Methods and Experimental Design

For videos used to perform a specific task, it may not be appropriate to rate video quality according to a subjective scale such as an absolute category rating (ACR) [11]. The goal of test methods for TRV is to assess a viewer's ability to recognise 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 a participant's ability to perform a task are useful in that they avoid ambiguity and personal preference.

In Clause 6, P.912 (08/08) stated:

The application of TRV is directly related to the user's ability to recognise targets at increasing levels of detail. These levels are referred to as Discrimination Classes (DC). 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. Therefore the object is represented by fewer Cycles Per Degree (CPD) of resolution. Fewer CPD of the resolution also means that the object subtends less of the video's information content, making identification of the target more difficult.

CPD, the critical parameter, is affected by the resolution of the object and (potentially) the distance between the camera and the object [14]. Consequently, it relates to achievable DC.

Consequently, the authors revised P.912 (08/08) with the following contribution:

CPD, the critical parameter, is affected by the resolution of the object and (potentially) the distance between the camera and the object [14]. Consequently, it relates to the achievable DC.

Examples of the achievable DC are shown in Fig. 2. 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".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 January 2021 doi:10.20944/preprints202101.0435.v1

Figure 2. *DC in testing methods for various distances between camera and object*

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

5.1. Multiple Choice Method

In Clause 6.1, P.912 (08/08) stated:

The number of choices offered to the viewer will depend on the number of alternative scenes being presented. "Unsure" may be one of the listed choices.

It should be noted that subjects tend to abuse the "Unsure" response. This problem has been observed when applying a Comparison Category Rating (CCR, Table 1), as defined in Recommendation ITU-T P.800 [15], in which subjects tend to abuse the response "0" ("About the Same"). A similar trend was observed independently in TRV studies conducted by the author.

Table 1. Comparison Category Rating (CCR)

3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

Unfortunately, P.912 (08/08) was missing a clear warning against the prudent use of the "Unsure" response (P.912 (08/08) even encouraged its use).

The authors revised that entry in P.912 (08/08) amending it as follows:

The number of choices offered to the viewer will depend 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****.*

5.2. Single Answer Method

In Clause 6.2, P.912 (08/08) stated:

If there is a non-ambiguous 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 what letter(s) or number(s) was present in a specific area of the video, and the answer can be evaluated as either correct or incorrect.

It should be noted that, contrary to P.912 (08/08), it is also possible to apply fuzzy logic [8]. For scenarios where the result of recognition is an alphanumeric string, assistance may come from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 January 2021 doi:10.20944/preprints202101.0435.v1

measuring differences between two strings using the Hamming distance (only for strings of the same length) [16], or Hamming distance's generalisation – the Levenshtein distance [17]. Using the experiment shown in Fig. 3 as an example results containing no more than one error may be regarded as correct [8]. This is because even in the event of a plate being recognised incorrectly, by correlating it with a vehicle database containing the vehicle's make and colour, we substantially reduce the risk of the vehicle being identified incorrectly.

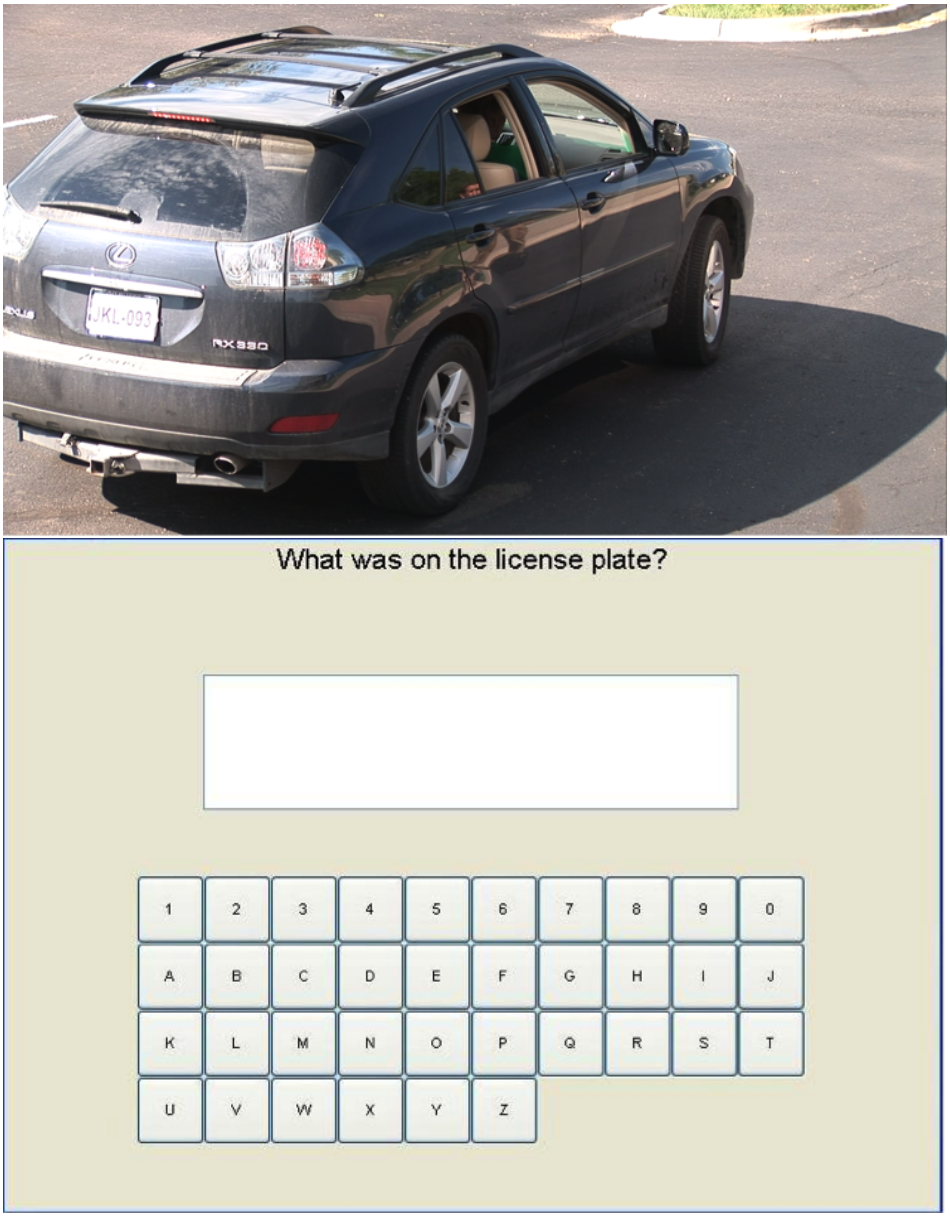

Figure 3. Single answer method.

The authors revised the description of the single choice method, expanding it as follows:

If there is a non-ambiguous 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 what letter(s) or number(s) was 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 Levenshtein distance), as seen in [8].***

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 January 2021 doi:10.20944/preprints202101.0435.v1

6. Evaluation Procedures

In Clause 7, a laboratory test is described. Description of a crowd-sourcing environment is described in Appendix I (see Section 8).

6.1. Subjects

In Clause 7.3, P.912 (08/08) stated:

Subjects who are experts in the target video recognition application field should be used. The number of subjects should follow the recommendations of [11].

To verify this finding, experiments were carried out (co-organised by the authors) testing subjects' ability to recognise particular objects (mobile phone, flashlight, gun, mug, radio, aluminium soda can, electric "Taser" stun gun) shown in video sequences. In the first experiment, the subjects were experts – law enforcement officers [18,19]. When the experiment was repeated with non-experts, very similar results were obtained, as long as the non-experts were compensated for their time [20].

The authors introduced an entry which allows the use of non-expert subjects providing they are motivated in an appropriate manner (such as being paid for their time). Naturally, this is only possible for some testing regions, since non-experts subjects cannot be used in tests associated with (for example) medical diagnostics.

Subjects who are experts in the application field of the TRV should be used. ***For certain areas of application testing, where neither specific experience nor expertise is required, non-expert subjects may also be used. Such non-experts must be motivated in an appropriate manner (such as being paid for their time). In [20] shows the validity of this approach.*** *The number of subjects should follow the recommendations of [11].*

6.2. Instructions to Subjects and Training Session

In Clause 7.4, P.912 (08/08) stated:

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 video's content, the questions should be posted before the video is shown so that the viewer knows what the task is.

However, the topic was not exhausted. Therefore, the authors extended this clause, adding:

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

1 *they cannot recognise a character;*
2 *they have doubts;*
3 *they can recognise 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 tricky is to define a task for specialists, e.g., medical doctors. In this case, the running of pretest on a small group before running any more massive 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. It can change the experimental conditions and finally harm the experiment itself. Therefore, it is essential in the pretest to clearly explain the task, the reason for running the test, and 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.

7. 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.

 doi:10.20944/preprints202101.0435.v1

7.1. Subject Screening

P.912 (08/08) did not contain information about subject screening at all. Therefore, the authors added a new subsection with the following content:

This technique is optional.

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 complicated 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 [21].

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 bit-rate. The Processed Video Sequences (PVS) obtained for the same source and lower bit-rate are likely to have less information, and the detection is likely not easier than for a higher bit-rate. Also, if an object covers fewer pixels on a screen, it is not easier to detect.

7.2. 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 considered to determine the correlation between success and CPD.

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

It is worth noting that the description of changes in this part P.912 (08/08) presented in this paper is not complete (it only contains the essential information) due to the authors' extensive contributions.

7.2.1. Multiple Choice

P.912 (08/08) stated only:

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 examining the stability of the answers within and between-subject performance metrics. "Unsure" answers should be pooled with those that are incorrect.

However, the topic was not exhausted. Therefore, the authors extended this clause, adding information about recognition probability as a function and comparing different conditions.

7.2.2. Single Answer

P.912 (08/08) stated only:

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.

However, the topic was not exhausted. Therefore, the authors extended this clause, adding:

For a single answer, the correctness of the answer can be analysed on a different scale. The most straightforward 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 not easy to describe all the options since the answer can differ depending on the answer type. Most probably, correctness can be analysed by the generalised linear model described in [22].

7.2.3. Timed Task

P.912 (08/08) stated:

 doi:10.20944/preprints202101.0435.v1

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.

However, the topic was not exhausted. Therefore, the authors extended this clause, adding:

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 "several replays of the movie before a decision was made." The analysis must indicate the influence of time on the result obtained.

8. Crowd-Sourcing Environment

P.912 (08/08) did not contain information about the crowd-sourcing environment. Therefore, the authors added a new appendix to P.912 (08/08). This appendix does not form an integral part of this Recommendation.

It is worth noting that the description of changes in this part P.912 (08/08) presented in this paper is not complete (it only contains the essential information) due to the authors' extensive contributions.

8.1. Introduction

One of the main problems of recognition tasks is the apparent limitation of source sequences reuse as described in Clause 6.5. The best way to protect against source sequence remembrance is to prevent showing 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 crowd-sourcing, which gives access to thousands of potential subjects at the same time.

8.2. Definitions

NOTE – 8.2 follows terminology presented in [23].

8.2.1. crowd-sourcing :

Obtaining the needed service by a large group of people, most probably an online community.

8.2.2. test:

Subjective assessments in a crowd-sourcing environment.

8.2.3. worker:

Person participating in a crowd-sourcing test.

8.2.4. task:

Set of actions that a worker needs to perform to complete a subscribed part of the test.

8.2.5. question:

A single event that requires an answer for a worker. A task contains many questions.

8.2.6. campaign:

A group of similar tasks. It also contains a more detailed description of the part of the test under investigation, like workers' payment, and indicates subjective assessments in a crowd-sourcing environment. A test can contain multiple campaigns.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 January 2021 doi:10.20944/preprints202101.0435.v1

8.3. Software

To run a crowd-sourced test, a worker has to have access to the test environment. Implementation of the test as a web service, which then 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 particular web browser plugin can restrict the number of participants significantly.

8.4. 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 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 recognise an object, it is recommended that pictures of the object be added, not only written descriptions.

8.5. 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 crowd-source 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.*

8.6. Data Analysis

Even with careful subject validation, assume that subjects are different. Since a diverse subgroup of subjects validates each sequence, the difference in recognition probability can be characterised only by a subgroup of subjects, not by a difference in conditions. Nevertheless, results show a high correlation between results obtained in a laboratory environment and those from crowd-sourcing. Such a result is possible only after removing unreliable subjects.

9. Conclusions

The discussion of statements contained in ITU-T Recommendation P.912 (08/08) showed that some of the findings and observations required the verification of specific provisions of the Recommendation. The authors revised Recommendation P.912 (08/08) to reflect improved subjective test techniques developed since this Recommendation was approved. Sufficient justification existed to support a new ITU-T work item, and contributions to this topic have been encouraged by ITU-T.

Ultimately, the amended Recommendation P.912 (03/16) has a broader scope; expands target testing methods; provides better instruction and training of subjects; improves conditions for testing, statistical analysis and reporting; and extends the applicability of techniques in the field of crowd-sourcing for the subjective assessment of the quality of TRV.

Author Contributions: Introduction, ML; Target Recognition Video, ML; Methods for Subjective Evaluation of TRV, ML; Source Signal, ML; Testing Methods and Experimental Design, ML; Evaluation Procedures, LJ; Statistical Analysis and Reporting of Results, LJ; Crowd-Sourcing Environment, LJ; Conclusions, ML

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cerqueira, E.; Zeadally, S.; Leszczuk, M.; Curado, M.; Mauthe, A. Recent advances in multimedia networking. *Multimedia Tools and Applications* **2011**, *54*, 635–647. doi:10.1007/s11042-010-0578-z.
2. Grega, M.; Janowski, L.; Leszczuk, M.; Romaniak, P.; Papir, Z. Quality of experience evaluation for multimedia services, Szacowanie postrzeganej jakosci uslug (QoE) komunikacji multimedialnej. *Przeglad Telekomunikacyjny* **2008**, *81*, 142–153.
3. Mu, M.; Romaniak, P.; Mauthe, A.; Leszczuk, M.; Janowski, L.; Cerqueira, E. Framework for the integrated video quality assessment. *Multimedia Tools and Applications* **2012**, *61*, 787–817.
4. Leszczuk, M.; Janowski, L.; Romaniak, P.; Papir, Z. Assessing quality of experience for high definition video streaming under diverse packet loss patterns. *Signal Processing: Image Communication* **2013**, *28*, 903–916.
5. Leszczuk, M.; Stange, I.; Ford, C. Determining image quality requirements for recognition tasks in generalized public safety video applications: Definitions, testing, standardization, and current trends. Broadband Multimedia Systems and Broadcasting (BMSB), 2011 IEEE International Symposium on, 2011, pp. 1–5. doi:10.1109/BMSB.2011.5954938.
6. Möller, S.; Raake, A., Eds. *Quality of Experience: Advanced Concepts, Applications and Methods*; Springer: Cham, 2014. doi:10.1007/978-3-319-02681-7.
7. Brunnström, K.; Beker, S.A.; De Moor, K.; Dooms, A.; Egger, S.; Garcia, M.N.; Hossfeld, T.; Jumisko-Pyykkö, S.; Keimel, C.; Larabi, M.C.; others. Qualinet white paper on definitions of quality of experience, 2013.
8. Leszczuk, M. Optimising task-based video quality. *Multimedia Tools and Applications* **2014**, *68*, 41–58. doi:10.1007/s11042-012-1161-6.
9. ITU-T. ITU-T P.912, Subjective video quality assessment methods for recognition tasks.
10. Leszczuk, M. Revising and Improving the ITU-T Recommendation P. 912. *Journal of Telecommunications and Information Technology* **2015**.
11. ITU-T. ITU-T P.910, Subjective video quality assessment methods for multimedia applications, 1999.
12. ITU-T. ANSI T1.801.01, Digital Transport of Video Teleconferencing/Video Telephony Signals –– Video Test Scenes for Subjective and Objective Performance Assessment, 1995.
13. Leszczuk, M.; Janowski, L. Database for video quality assessment in license plate recognition. Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2013, 2013, pp. 51–55.
14. Leszczuk, M.; Janowski, L.; Romaniak, P.; Głowacz, A.; Mirek, R. Quality assessment for a licence plate recognition task based on a video streamed in limited networking conditions. International Conference on Multimedia Communications, Services and Security. Springer, 2011, pp. 10–18.
15. ITU-T. ITU-T P.800, Methods for subjective determination of transmission quality, 1996.
16. Hamming, R. Error Detecting and Error Correcting Codes. *Bell System Technical Journal* **1950**, *26*, 147–160.

17. Levenshtein, V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* **1966**, *10*, 707.
18. VQiPS. Video Quality Tests for Object Recognition Applications. Public Safety Communications DHS-TR-PSC-10-09, U.S. Department of Homeland Security's Office for Interoperability and Compatibility, 2010.
19. VQiPS. Recorded-Video Quality Tests for Object Recognition Tasks. Public Safety Communications DHS-TR-PSC-11-01, U.S. Department of Homeland Security's Office for Interoperability and Compatibility, 2011.
20. Leszczuk, M.; Kon, A.; Dumke, J.; Janowski, L. Redefining ITU-T P.912 Recommendation Requirements for Subjects of Quality Assessments in Recognition Tasks. In *Multimedia Communications, Services and Security*; Dziech, A.; Czyzewski, A., Eds.; Springer Berlin Heidelberg, 2012; Vol. 287, *Communications in Computer and Information Science*, pp. 188–199. doi:10.1007/978-3-642-30721-8_19.
21. Janowski, L. Task-based subject validation: Reliability metrics. 2012 Fourth International Workshop on Quality of Multimedia Experience, 2012, pp. 182–187. doi:10.1109/QoMEX.2012.6263863.
22. Agresti, A. *Categorical data analysis*; Vol. 482, John Wiley & Sons, 2003.
23. Hoßfeld, T.; Hirth, M.; Redi, J.; Mazza, F.; Korshunov, P.; Naderi, B.; Seufert, M.; Gardlo, B.; Egger, S.; Keimel, C. Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force" Crowdsourcing", 2014.