

Quality Assessment for Recognition and Task-based multimedia applications (QART)

Mikołaj Leszczuk and Lucjan Janowski



Presentation plan

- Reminder on QART and Target Recognition Video (TRV)
- Report for 2014H2
 - Source signal
 - Multiple choice method
 - Single answer method
 - Subjects
- Plans for 2015H1
 - Testing methods
 - Multiple choice method (more on that topic)
 - Instructing and training subjects
 - Conditions for testing
 - Statistical analysis and reporting



Reminder on QART and Target Recognition Video (TRV)



VQEG's Subproject: QART

- **Mission:**
“To study effects of resolution, compression and network effects on quality of video used for recognition tasks”
- **Goals:**
 - To perform series of tests to study effects and interactions of
 - Compression
 - Scene characteristics
 - To test existing or develop new objective measurements that will predict results of subjective tests of visual intelligibility
 - Propose subjective test methodology for recognition tasks



Task Recognition Specificity (1/2)

- In many applications video quality not as important as ability to accomplish specific task for which video was created
- Typical examples of such **TRV**:
 - Video surveillance systems
 - Telemedicine/remote medical consultation/diagnosis system
 - Fire safety
 - Backup camera installed in car helping to park
- Quality tests needed
- General idea behind quality tests for TRV: to find threshold at which task can be achieved with certain probability or accuracy



Task Recognition Specificity (2/2)

- Therefore, instead of quality evaluation, subjective experiment focused on task performance measurement
- For example, test might be measuring probability of:
 - For a video surveillance – recognition of license plate numbers
 - For telemedicine/remote diagnosis – correct diagnosis
 - For fire safety – fire detection
 - For backup camera – parking the car



ITU-T Recommendation P.912

- Problems of quality evaluation procedures and measurements for TRV **partially standardized** in ITU Recommendation P.912
- **Title:** “*Subjective Video Quality Assessment Methods for Recognition Tasks*”
- Published: 2008
- Introducing:
 - Basic definitions
 - Methods of testing
 - Psycho-physical experiments

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.912

(08/2008)

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

ITU-T



Report for 2014H2



P.912 Revision (1/2)

- Based on **research and observations with VQEG => introduction of modifications to P.912**
- Formalized procedures for this purpose
- Collaboration with **Polish Ministry of Administration and Digitization**
- Received nomination as delegate of Polish government



P.912 Revision (2/2)

- **ITU-T Study Group 9 (SG9)**
- **SG9 Meeting, 8-12 Sep, Geneva:**
 - Clause 5 (“Source Signal”)
 - Clause 6.1 (“Multiple Choice Method”)
 - Clause 6.2 (“Single Answer Method”)
 - Clause 7.3 (“Subjects”)
- Detailed scope of amendments to Recommendation P.912 discussed in following slides...



Source Signal (1/2)

Clause 5 of Recommendation P.912:

*Test sequences should follow the general principles stated in [ANSI T1.801.01-1995] and [ITU-T P.910], which specify scenes that 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 about, and cover the entire range of possible scenarios for the application area identifying that one is.***



Source Signal (2/2)

- In certain cases, data availability very limited
- High data diversity, e.g.:
 - X-ray diagnosis of bone fractures
 - Licence plate recognition
- Literature including attempts to extrapolate applicability of results
- Proposed introduction of warnings in Clause 5





INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION
STANDARDIZATION SECTOR**

STUDY PERIOD 2013-2016

COM 09-C 069

September 2014

English only

Original: English

Question(s): 12/9

STUDY GROUP 9 – CONTRIBUTION xx

Source: Poland

Title: Proposed Changes to P.912, “Subjective video quality assessment methods for recognition tasks”

Authors

1. Mikołaj Leszczuk, AGH University of Science and Technology, Poland
2. Lucjan Janowski, AGH University of Science and Technology, Poland

Introduction

In Clause 5, Recommendation P.912 states:

Test sequences should follow the general principles stated in [ITU-T P.910] and [b-T1.801.01],

Multiple Choice Method (1/2)

Clause 6.1 of Recommendation P.912:

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.



Please answer clip 1 of 20
What was the person holding?

Gun	Banana	Comb
Hairpiece	Knife	Potato

Multiple Choice Method (2/2)

- Subjects **tending to abuse „Unsure” response**
- Similarly: „0” (*About the Same*), P.800 CCR (*Comparison Category Rating*)
- Missing warning against prudent use of „Unsure”
- Even encouraging its use
- Proposed entry in Recommendation P. 912

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





INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION
STANDARDIZATION SECTOR**

STUDY PERIOD 2013-2016

COM 09-C 070

September 2014

English only

Original: English

Question(s): 12/9

STUDY GROUP 9 – CONTRIBUTION COM 09-C 070

Source: Poland

Title: Proposed Changes to P.912, “Subjective video quality assessment methods for recognition tasks”

Authors

1. Mikołaj Leszczuk, AGH University of Science and Technology, Poland
2. Lucjan Janowski, AGH University of Science and Technology, Poland

Introduction

In Clause 6.1, Recommendation P.912 states:

The number of choices offered to the viewer will depend on the number of alternative scenes being

Single Answer Method (1/3)

Clause 6.2 of Rec. P.912:

*If there is a non-ambiguous answer is 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.***



What was on the license plate?

1	2	3	4	5	6	7	8	9	0
A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				

Single Answer Method (2/3)

- It should be noted that, contrary to Recommendation P.912, it is also possible to apply fuzzy logic
- For alphanumeric results of recognition, assistance may come from measuring differences between two strings using:
 - **Hamming Distance** (for strings of same length)
 - **Levensthein Distance** (Hamming's generalization)
- As example, in practice, results containing no more than one error may be regarded as correct



Single Answer Method (3/3)

- Reduced risk by correlating vehicle identification, with vehicle database, containing also:
 - Make of vehicle, and
 - Color of vehicle
- Proposed description of single choice method to be expanded





INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION
STANDARDIZATION SECTOR**

STUDY PERIOD 2013-2016

COM 09-C 071

September 2014

English only

Original: English

Question(s): 12/9

STUDY GROUP 9 – CONTRIBUTION COM 09-C 071

Source: Poland

Title: Proposed Changes to P.912, “Subjective video quality assessment methods for recognition tasks”

Authors

1. Mikołaj Leszczuk, AGH University of Science and Technology, Poland
2. Lucjan Janowski, AGH University of Science and Technology, Poland

Introduction

In Clause 6.2, Recommendation P.912 states:

If there is a non-ambiguous answer is an identification question, the single answer method may be

Subjects (1/2)

Clause 7.3 of Recommendation P.912:

*Subjects who are **experts** in the application field of the target video recognition should be used.*



Subjects (2/2)

- ITS-NTIA and AGH experiments testing subjects' ability to recognize certain objects
- In first experiment, expert subjects - law enforcement officers, **as with P.912, but...**
- **Observation:**
 - Experiment repeated with non-experts
 - Very similar results obtained, as long as non-experts were compensated for their time



Subjects – Conclusions

- Proposed introduction of entry allowing use of non-expert subjects
- Providing motivated in an appropriate manner (such as being paid for time)
- Only possible for certain areas, since non-experts cannot used in (for example) medical diagnostics



Expert subject

- Costly (practitioner):
 - Police officer
 - Doctor
 - Difficult to hire



Non-expert subject

- Cheap (colleague/friend)
 - Student
 - Retired
 - Easy to hire





INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION
STANDARDIZATION SECTOR**

STUDY PERIOD 2013-2016

COM 09-C 072

September 2014

English only

Original: English

Question(s): 12/9

STUDY GROUP 9 – CONTRIBUTION COM 09-C 072

Source: Poland

Title: Proposed Changes to P.912, “Subjective video quality assessment methods for recognition tasks”

Authors

1. Mikołaj Leszczuk, AGH University of Science and Technology, Poland
2. Lucjan Janowski, AGH University of Science and Technology, Poland

Introduction

In Clause 7.3, Recommendation P.912 states:

Subjects who are experts in the application field of the target video recognition should be used.

Plans for 2015H1



Testing Methods (1/2)

Section 6 of P.912:

The application of TRV is directly related to the ability of the user that recognizes targets at increasing levels of detail. These levels are referred to as Discrimination Classes (DC). When determining the DC for particular scenarios, they must consider that for a set distance from the camera to the object of interest, the DC directly correlates with decreasing resolution of the target, and therefore the object is represented by fewer cycles per degree of resolution. Fewer cycles per degree of resolution also means that the object subtends less of the information content of the video, making identification of the target more difficult.

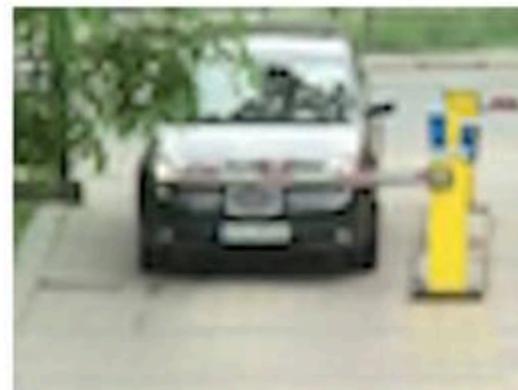
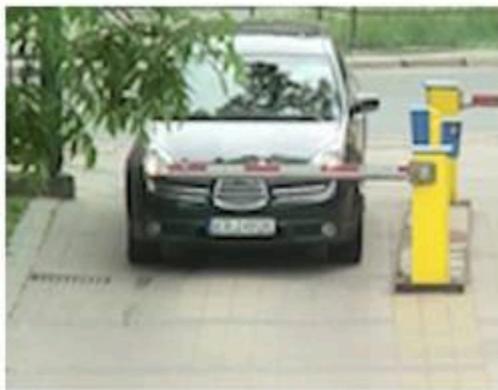


Testing Methods (2/2)

- Not easy to understand relationship between parameters such as:
 - Number of Cycles-Per-Degree (CPD)
 - Resolution of the object, and
 - Distance between camera and object
- CPD – key parameter is CPD, affected by:
 - Resolution of object, and
 - Distance between camera and object (potentially)
- Changes involving easy explanation of parameters proposed



Testing Methods Cartoon 😊



50m 

215 m 

430m 

50 m – Target Positive Recognition



215 m – Target Characteristics



430 m – Target Presence



Multiple-Choice Method (1/2)

Section 6.1 of P.912:

This method is appropriate for all DC levels and target categories (human, object and alphanumeric). For this method, the video is shown above a letter 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 accommodate arise from open questions, and allows for more accurate measurements.



Please answer clip 1 of 20
What was the person holding?

Gun

Banana

Comb

Hairpiece

Knife

Potato

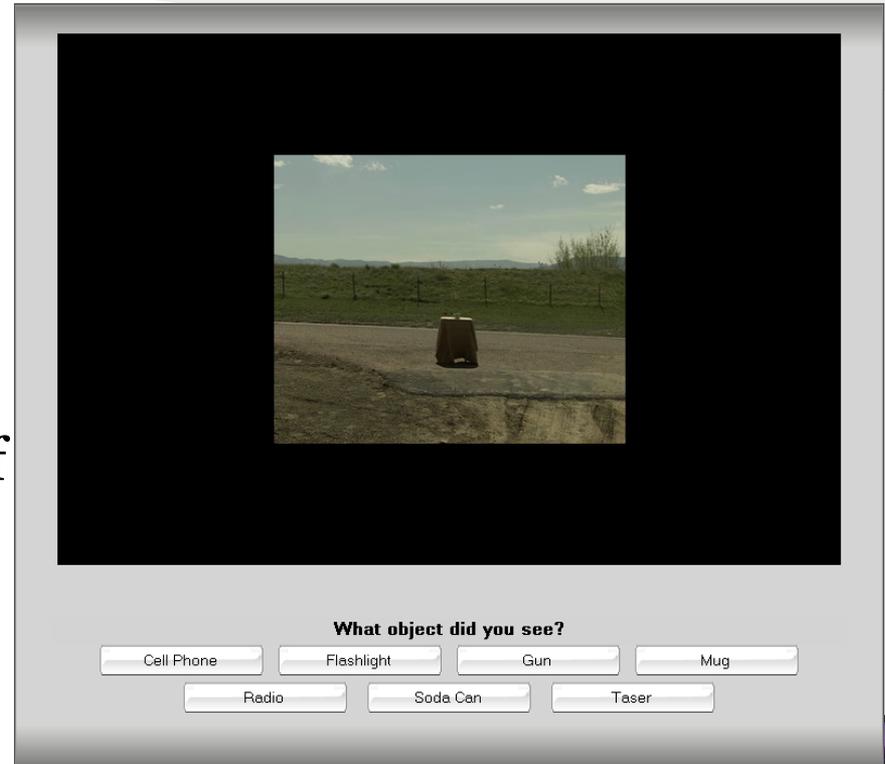
Hairpiece

Knife

Potato

Multiple-Choice Method (2/2)

- Nothing on impact on choices by buttons':
 - Order
 - Position
- Research showing existence of such impact
- Proposing random sequence of buttons
- Proposing adding picture to the words so it will be easier to find randomly moved buttons



Instructing and Training of Subjects (1/4)

Section 7.4 of P.912:

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 the viewer knows that what the task is.

Section 6.2 of P.912:

Care must also be taken to avoid terminology that may differ from participant to participant.



Instructing and Training of Subjects (2/4)

- Issues on interacting with subjects not referred in single Section of P.912
- Unnecessary breakdown of topic
- Call for assembling in one (dedicated) Section 7.4 of P.912



Instructing and Training of Subjects (3/4)

- AGH experiment on recognizing license plates
- Subjects instructed, **compliance with P.912, Sec 7.4, but...**
- Observation:
 - Some subjects recognizing just most obvious characters
 - Others many more of them
- Conclusion:
 - Some subjects assuming to give up on characters difficult to read
 - Others trying hard to read all characters



PLATE NUMBER: I DON'T KNOW COLOR:

BRAND:

 Audi <input type="radio"/>	 BMW <input type="radio"/>	 Citroen <input type="radio"/>	 Daewoo <input type="radio"/>	 Fiat <input type="radio"/>
 Ford <input type="radio"/>	 Honda <input type="radio"/>	 Hyundai <input type="radio"/>	 Kia <input type="radio"/>	 Mazda <input type="radio"/>
 Mercedes <input type="radio"/>	 Nissan <input type="radio"/>	 Opel <input type="radio"/>	 Peugeot <input type="radio"/>	 Renault <input type="radio"/>
 Rover <input type="radio"/>	 Seat <input type="radio"/>	 Skoda <input type="radio"/>	 Subaru <input type="radio"/>	 Suzuki <input type="radio"/>
 Toyota <input type="radio"/>	 Volkswagen <input type="radio"/>	 Volvo <input type="radio"/>	<input type="text"/>	 I don't know <input type="radio"/>

SEND

Instructing and Training of Subjects (4/4)

Proposed changes:

- Adding to training clear examples of correct and incorrect task evaluation
- Objects described by pictures and words
- In case of tests involving specialists, e.g. medical doctors, a preliminary test of the instruction and training itself is recommended



Conditions for Testing (1/2)

**Sections: 5, 6, 6.6, 6.7,
7.1, 7.2, 7.3 of P.912:**

*The Experimenter should
follow the guidelines
outlined in [ITU-T P.910].*



Conditions for Testing (2/2)

- At time of approval P.912 probably most recent on testing conditions to which to refer was P.910 (1998)
- As result, vast majority of tests performed previously under strictly controlled conditions, defined in P.910
- By 2014 P.913 approved largely displacing P.910, including defining smoother requirements for testing
- Calling for introduction of references to P.913, replacing references to P.910



Statistical Analysis and Reporting (1/2)

Section 8 of P.912:

For single answer conditions, where the answers are correct or incorrect, a statistical metric to determine if the subject is performing above the level of chance for answering correctly should be implemented. “Unsure” answers should be pooled with the incorrect answers.

For multiple-choice answers, the probability of an incorrect answer needs to be balanced against the ability to answer the questions correctly. The statistic metric in this situation will require an examination of the stability of the answers within and between subject performance metrics. “Unsure” answers should be pooled with the incorrect answers.



Statistical Analysis and Reporting (2/2)

- Very general statement, we would like to add some specific statistical tools
- For statistical analysis of results, authors shown:
 - Possibility of using logistic function, with equations
 - Possibility of comparing different conditions, with equations
 - Possibility of using Generalized Linear Model (GLZ), just mentioned
 - Proposals for removing outlier's responses from pool of results - standard procedure in other QoE studies



Thank You!

<http://mitsu-project.eu/>



Narodowe Centrum
Badań i Rozwoju



European Union

European Regional
Development Fund
"A way to build Europe"

