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 2015H1
  • Conditions for testing
  • Testing methods
  • Multiple choice method and single-answer method
  • Instructing and training subjects
  • Statistical analysis and reporting

• Plans for 2015H2
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
Therefore, instead of quality evaluation, subjective experiment focused on task performance measurement.

For example, test might be measuring probability of:

- For video surveillance – recognition of license plate numbers
- For telemedicine/remote diagnosis – correct diagnosis
- For fire safety – fire detection
- For backup camera – parking car
ITU-T Recommendation P.912

- Problems of quality evaluation procedures & 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
Report for 2015H1
Based on research & 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, 10-17 Jun, Beijing:**
  - Whole Recommendation (conditions for testing)
  - Clause 6 (“Test Methods”)
  - Clause 6.1 (“Multiple Choice Method”)
  - Clause 6.2 (“Single Answer Method”)
  - Clause 7.4 (“Instructions to Subjects”)
  - Clause 8 (“Statistical Analysis & Reporting”)
• Detailed scope of amendments to Recommendation P.912 discussed in following slides...
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 extending P.910, including defining smoother requirements for testing
- Called for introduction of references to P.913, besides references to P.910
Introduction

In Clauses 5, 6, 6.6, 6.7, 7.1, 7.2 and 7.3, Recommendation P.912 states:

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

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 was performed previously under strictly controlled conditions, By 2014 P.913 has been approved partially displacing P.910, including defining smoother requirements for testing.

Proposal

Consequently, the authors propose the introduction of appropriate changes to the Recommendation P.912, involving introduction of references to P.913, replacing references to P.910.
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 video is 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
50 m – Target Positive Recognition
215 m – Target Characteristics
430 m – Target Presence
Testing Methods Cartoon 😊
In Clause 6, Recommendation P.912 states:

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 the video 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.

Unfortunately, especially for less experienced researchers, it is not easy to understand the relationship between parameters such as: number of Cycles Per Degree (CPD), object resolution, and distance between camera and object. Furthermore, CPD – the key parameter, is affected by: resolution of object, and (potentially) distance between camera and object[Leszczuk2011].
Multiple-Choice Method and Single-Answer Method (1/2)

Clauses 6.1 and 6.2 of Recommendation P.912:

(...) “Unsure” may be one of the listed choices.
Multiple-Choice Method and Single-Answer 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 justification by literature reference entry in **Recommendation P.912**
Question(s): 12/9

STUDY GROUP 9 – CONTRIBUTION COM 09-C 111

Source: Poland

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

Author(s) of the Contribution:

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mikolaj Leszczuk, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:leszczuk@kt.agh.edu.pl">leszczuk@kt.agh.edu.pl</a></td>
</tr>
<tr>
<td>Lucjan Janowski, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:janowski@kt.agh.edu.pl">janowski@kt.agh.edu.pl</a></td>
</tr>
</tbody>
</table>

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 presented. "Unsure” may be one of the listed choices.*

It should be noted that subjects tend to abuse the “Unsure”. This problem has been observed when applying a Comparison Category Rating (CCR, table below), as defined in Recommendation ITU-T P.800 [2], in which method, subjects abuse the response “0” (“About the Same”). A similar trend was observed independently in TRVs studies conducted by the authors.
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 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 Recommendation P.912, it is possible to also apply fuzzy logic [Leszczuk]. For scenarios where the result of recognition is an alphanumeric string, such assistance may come by measuring differences between the two strings with the Hamming distance (only for strings of the same length) [Hamming], or its generalization – the Levensthein distance [Levenshtein][Левенштейн]. For example, in practice, to experiment imaged in figure below, it can be regarded with success to consider results containing not more than one error as correct ones [Leszczuk]. It is because even with the wrong result of plate recognition, by correlating it with vehicle database that contains the make and colour of the vehicle, we substantially limit the possibility of ultimate misrecognition.
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
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, preliminary test of instruction & training itself is recommended
Introduction

In Clause 7.4, Recommendation P.912 states:

*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.*
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 added 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
In Clause 8, Recommendation P.912 states:

*The statistical analysis for each method will vary slightly.*

**Single answer**

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.

**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 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.

**Timed task**


**Author(s) of the Contribution:**

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucjan Janowski, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:janowski@kt.agh.edu.pl">janowski@kt.agh.edu.pl</a></td>
</tr>
<tr>
<td>Mikołaj Leszczuk, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:leszczuk@kt.agh.edu.pl">leszczuk@kt.agh.edu.pl</a></td>
</tr>
</tbody>
</table>

**Introduction**

In Clause 8, Recommendation P.912 states:

_The statistical analysis for each method will vary slightly._

**Single answer**

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.

**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 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.

**Timed task**


**Author(s) of the Contribution:**

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucjan Janowski, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:janowski@kt.agh.edu.pl">janowski@kt.agh.edu.pl</a></td>
</tr>
<tr>
<td>Mikołaj Leszczuk, AGH University of Science and Technology, Poland</td>
<td><a href="mailto:leszczuk@kt.agh.edu.pl">leszczuk@kt.agh.edu.pl</a></td>
</tr>
</tbody>
</table>
Plans for 2015H2
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.
• Nothing on impact on choices by buttons’:
  • Order
  • Position
• Research confirms such impact
• Proposing random sequence of buttons
• Proposing adding picture to words – to be easier to find random moved buttons
• Crowdsourcing experiment
  • Uncertain first, preliminary results
  • Users not complain about the random setting
  • No one reported this problem, but reported others
  • Discussion?
Summary

- **10** contributions submitted:
  - 4 at Sep 2014, Geneva
  - 6 at Jun 2015, Beijing

- Next **SG9 meeting** – Jan 2016, Geneva
  - 1-2 more contribution/s (maybe) to be submitted
  - Seeking for final consent (approving, closing)
Thank You!

http://mitsu-project.eu/