Methodologies for subjective quality evaluation of short and long 360-degree videos

Jesús Gutiérrez, Pablo Pérez, Femi Adeyemi-Ejeye
VQEG Plenary Meeting, Mountain View, 12-16 Nov 2018
Outline

• Introduction / motivation
• Subjective evaluation of video quality: from 2D to immersive media
• What is short and long?
• Related work for short 360-degree videos
• Related work for long 360-degree videos
Introduction / Motivation

• Need of recommendations/standards for subjective quality assessment of 360-degree videos.
  • Work on defining test plan within VQEG-IMG
  • Contributions to ITU-T SG12/13 G.360-VR
• Some works have been already published using typical methodologies for 2D video.
• Importance of the duration of test content:
  • 10 seconds (e.g., MPEG) → too short?
  • Different factors to evaluate depending on duration? Immersion, sickness, etc.
  • Different methodologies for short and long sequences?
  • What is short and long?
### Subjective evaluation of 2-Dimensional video quality

<table>
<thead>
<tr>
<th>Standard</th>
<th>Full meaning</th>
<th>Stimuli Presentation</th>
<th>Questions / scales</th>
<th>Voting method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR</td>
<td>Absolute Category Rating</td>
<td>Single Stimulus</td>
<td>5-grade quality scale (&quot;Bad – Excellent&quot;)</td>
<td>Absolute Values</td>
</tr>
<tr>
<td>ACR-HR</td>
<td>Absolute Category Rating with Hidden Reference</td>
<td>Single Stimulus</td>
<td>5-grade quality scale (&quot;Bad – Excellent&quot;)</td>
<td>Absolute Values. Differential scores between reference and Impaired versions (DMOS)</td>
</tr>
<tr>
<td>SSCQE</td>
<td>Single Stimulus Continuous Quality Rating</td>
<td>Single Stimulus</td>
<td>Continous Scale over time, at certain intervals</td>
<td>Slider/Fader</td>
</tr>
<tr>
<td>DSCQS</td>
<td>Double Stimulus Continuous Quality Scale</td>
<td>Double Stimulus</td>
<td>Continous Scale over time, at certain intervals</td>
<td>Slider/Fader</td>
</tr>
<tr>
<td>DSIS</td>
<td>Double Stimulus Impairment Scale</td>
<td>Double Stimulus</td>
<td>5-grade scale (&quot;Very Annoying – Imperceptible&quot;)</td>
<td>Absolute Values</td>
</tr>
<tr>
<td>PC</td>
<td>Pair Comparison</td>
<td>Double Stimulus</td>
<td>5-grade scale (&quot;Very Annoying – Imperceptible&quot;) Preference</td>
<td>Absolute Values. Preference (transformation of values with e.g. BT-model)</td>
</tr>
</tbody>
</table>
Subjective evaluation of video quality
Immersive media adds more dimensions

2D
- Content Type
- Encoding
  - Target bitrate
  - Target resolution
  - Video Codec and Implementation
  - Encoding Parameters
- Display Resolution
- Network Impairments

Immersive Media
- Content Type
- Encoding
  - Target bitrate
  - Target resolution
  - Video Codec and Implementation
  - Encoding Parameters
- Display Resolution
- Network Impairments
- Immersion
- Presence
- Cyber sickness
- Exploration Behaviour
- Physiological responses
- Audio-Visual quality

VS
What is short and long?

Stimuli duration

- No standard definition
- For 2D videos
  - In 2009, Interactive Advertising Bureau prescribed long sequences as those longer than 10 mins in length.
  - On Youtube, long sequences are those defined to be longer than 20 mins in length, while short sequences are less than 4 mins
  - SoA subjective tests: long sequences from 1 minute.
- For Immersive media
  - Makers of VR headsets recommend you take a break of 10-15mins after every 30 mins
  - What are the acceptable durations for Long and Short Sequences?
What is short and long?
How much time do observers need to explore 360° content?

- At least 20 seconds to explore images.

- M. Huang et al. TIP2018: Testing different exploration times with images:
  - 10s: Too short
  - 20s: Time “to acclimate to a fixed virtual world”.
  - 40s: Too long for their setup. No improvement over 20s.

- Exploration of videos:
  - “Driven by contents” (F. Duanmu et al. ICME2018) → From “diffused scenes” (exploration like images) to “concentrated scenes” (limited exploration).
  - Limited movements (Singla et al. AhG82017).
  - Repeating the clips “does not necessarily lead to more unique fixation points” (Ozcinar et al. QoMEX2018)
Related work for **short** 360-degree videos

**Introduction**

- Some works published on quality evaluation of 360-degree short videos:
  - Short videos: typically used to develop and evaluate the performance of coding techniques.
    - Videos currently used in MPEG: 10 seconds
  - Mainly only evaluation of audiovisual quality
  - Use of typical methodologies for 2D video: ACR, DSIS, etc.

- **Issues with evaluating short sequences:**
  - Limited immersiveness/interest of the observer on/for the content (even in 2D videos).
  - Videos too short to be explored by the observer?
    - Need of new methodologies? → Modified ACR (Singla et al., ACMMM2017)
## Relevant references

### Relevant work for short 360-degree videos

<table>
<thead>
<tr>
<th>Paper</th>
<th>Objective</th>
<th>Presentation Methodology</th>
<th>Questions / scales</th>
<th>Stimuli duration</th>
<th>Num. Observers</th>
<th>HMD</th>
<th>Voting interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singla et al., ACMMM2017</td>
<td>Coding quality</td>
<td>Modified-ACR</td>
<td>5-grade quality scale</td>
<td>10 s.</td>
<td>30</td>
<td>Oculus Rift</td>
<td>Scale shown on HMD, rating recorded verbally</td>
</tr>
<tr>
<td>Singla et al., HVEI2018</td>
<td>Compare M-ACR and DSIS</td>
<td>M-ACR DSIS</td>
<td>5-grade quality scale</td>
<td>10 s.</td>
<td>30 / 27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xu et al. arXiv2017</td>
<td>Coding quality</td>
<td>ACR</td>
<td>Continuous scale 0-100</td>
<td>12 s.</td>
<td>48</td>
<td>HTC Vive</td>
<td>Slider</td>
</tr>
<tr>
<td>Zhang et al., ICMEW2017</td>
<td>Coding quality</td>
<td>SSCQS SAMVIQ SAMPVIQ</td>
<td>0-5 quality scale</td>
<td>10s</td>
<td>10 / 16 / 23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upenik et al. PCS2016.</td>
<td>Image coding quality (JPEG)</td>
<td>ACR-HR</td>
<td>5-grade quality scale</td>
<td>30 s.</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perrin et al. SPIE2017</td>
<td>HDR quality</td>
<td>PC toggling (switching viewports between reference and test stimuli)</td>
<td>5-grade scale (“worse than”… “Better than”)</td>
<td>x</td>
<td>25</td>
<td>HMD “MergeVR2” and iPhone 6S</td>
<td>Displayed on the voting menu of the testbed</td>
</tr>
</tbody>
</table>
Related work for long 360-degree videos

Introduction

• Very few work on assessing audiovisual quality of long 360 videos
  • AV quality + presence, or just presence-like questions
  • Heterogeneous approach: each work uses its own questionnaires / objectives.
  • Common factors:
    • Each source shown once
    • 1-5 minute sequences
    • 5-50 diverse questions at the end (# depends on # of stimuli per subject)

• Issues with evaluating (2D) long sequences (Garcia 2014, Chen 2013):
  • Hysteresis: past stimulus affect present evaluation
  • Recency: recent events are more relevant than far away events
  • Continuous evaluation: people may forget to evaluate and immerse in the content
  • Number of test sequences per test becomes highly limited
Related work for long 360-degree videos

Content immersion

- For long sequences, factorial design is not possible
  - Not practical (session too long)
  - If people remember stimuli, some QoE factors cannot be assessed (MacQuarrie 2017).
- As an alternative, content-immersive methods are used (Pinson 2014)
  - Put the subject in the frame of mind of using the system for its intended application.
  - Longer and interesting stimuli to engage the subject (e.g., one minute).
  - Match the sensory experience of the target application—not the impairment modality.
  - Each source stimulus is viewed or heard only once by each subject.
- Most existing long-sequence evaluations actually follow it
  - 360 video (all references we have analyzed)
  - 2D video, e.g. P.NATS, see (Raake 2017).
Related work for long 360-degree videos
Within-sequence quality evaluation

- Target: finer-grain measurements, several conditions per sequence.
- We didn't found any reference for 360 video
- Approaches (2D/3D video):
  - Continuous (Staelens 2014): SSCQE, slider where user can select quality continuously.
  - Discrete (Gutierrez 2011): periodic questions to evaluate the previous X seconds of sequence (content is kept playing).
  - Interactive (Borowiak 2014): User can select desired quality by rotating a knob.
- Interaction with content immersion is unknown.
## Related work for long 360-degree videos

### Relevant references

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Schatz et al. QoMEX2017</td>
<td>- Video stalling Normal screen vs HMD.</td>
<td>ACR-HR</td>
<td>- Overall quality, stalling annoyance: 5-grade</td>
<td>60 s.</td>
<td>22</td>
<td>Oculus Rift DK2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Presence (x4): 7-grade (attention, spatial presence, awareness, realistic)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singla et al., QoMEX2017</td>
<td>- QoE and sickness Compare two HMDs</td>
<td>SS (clip + questions)</td>
<td>- Quality evaluation: 5-grade quality scale</td>
<td>60 - 65 s.</td>
<td>28</td>
<td>HTC Vive and Oculus Rift</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- SSQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MacQuarrie &amp; Steed, IEEEVR 2017</td>
<td>- HMD vs TV vs SurroundVideo+ QoE factors</td>
<td>SS (clip + questions)</td>
<td>- Spatial Awareness (object location)</td>
<td>2-5 min</td>
<td>63</td>
<td>Oculus Rift CV1, CAVE, 60” TV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Incidental Memory: 10x open answer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Narrative Engagement (MNEQ)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Enjoyment: 2x 5p Likert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Concern about missing something; 3x 5p Likert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Fear (horror movie): 2x 5p Likert</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guervós et al. HVEI’19</td>
<td>- QoE in learning Veterinary students, real lesson</td>
<td></td>
<td>- Video and overall quality: 5-grade (ACR)</td>
<td>5 min</td>
<td>100</td>
<td>Samsung Gear VR (Galaxy S8+)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Simulator Sickness: 5-grade (Vertigo)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Net Promoter Score: 10-grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Temple Presence Inventory: 40 presence questions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion
Short sequences

What we know
• Length: 10-30 seconds
• Traditional methodologies seem valid
  • M-ACR for very short sequences (e.g., 10 seconds)
• Realistic watching setup (HMD, headphones, video+audio)
• Questions after each clip
• Factors to evaluate: mainly audiovisual quality

Open points
• Effects and need of evaluating other factors (e.g., immersion, cyber-sickness…),
• Validity of typical methodologies:
  • Cross-lab study
Conclusion
Long sequences

What we know
• Length: 1-5 minutes
• Each sequence shown once
  • Therefore Single Stimulus
• Realistic watching setup (HMD, headphones, video+audio)
• Questions after each sequence
• Several factors to evaluate (not only video quality)

Open points
• Narrow down recommended duration?
• Recommend questionnaire
  • Fixed or open?
  • Which factors to evaluate?
• Intra-sequence evaluation? Which method?
  • Focused on a single factor (audiovisual QoE)
  • SSCQE? Other?
References

1/3


References

2/3

- Borowiak, Adam, and Ulrich Reiter. "Long duration audiovisual content: Impact of content type and impairment appearance on user quality expectations over time." Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on. IEEE, 2013.
References