UNIVERSITÉ DE NANTES

NOKIA Bell Labs

UNIVERSITY OF SURREY

VQEG

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

Standard	Full meaning	Stimuli Presentation	Questions / scales	Voting method
ACR	Absolute Category Rating	Single Stimulus	5-grade quality scale ("Bad – Excellent")	Absolute Values
ACR-HR	Absolute Category Rating with Hidden Reference		5-grade quality scale ("Bad – Excellent")	Absolute Values. Differential scores between reference and Impaired versions (DMOS)
SSCQE	Single Stimulus Continuous Quality Rating		Continuouse Scale over time, at certain intervals	Slider/Fader
DSCQS	Double Stimulus Continuous Quality Scale	Double Stimulus	Continuouse Scale over time, at certain intervals	Slider/Fader
DSIS	Double Stimulus Impairment Scale		5-grade scale ("Very Annoying – Imperceptible")	Absolute Values
PC	Pair Comparison		5-grade scale ("Very Annoying – Imperceptible") Preference	Absolute Values. Preference (transformation of values with e.g. BT-model)

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

VS

Immersive Media

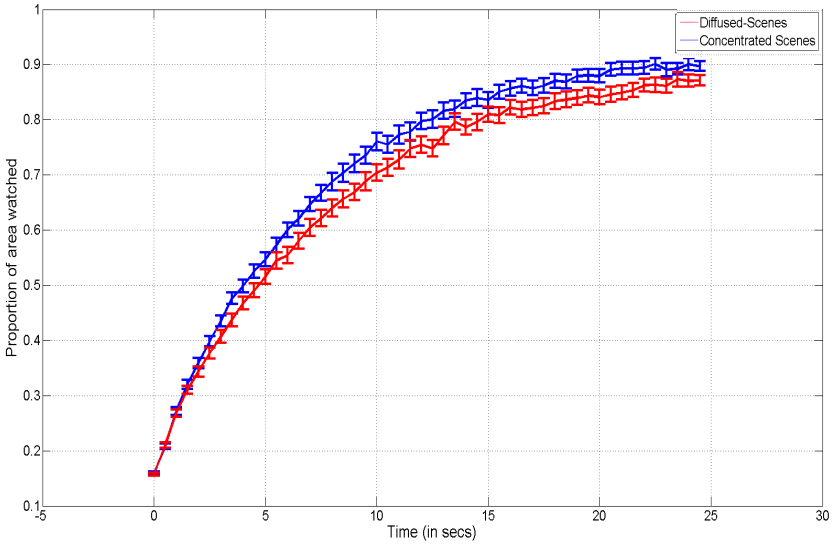
- Content Type
- Encoding
  - Target bitrate
  - Target resolution
  - Video Codec and Implementation
  - Encoding Parameters
- Display Resolution
- Network Impairments
- Immersion
- Presence
- Cyber sickness
- Exploration Behaviour
- Physiological responses
- Audio-Visual quality

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 360º content?

- At least 20 seconds to explore images.



Rai et al. MMSys2017

- 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)

Related work for **short** 360-degree videos
Relevant references

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

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

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

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

- E. Upenik, M. Rerabek, and T. Ebrahimi, “Testbed for subjective evaluation of omnidirectional visual content,” in 2016 Picture Coding Symposium (PCS), 2016, pp. 1–5.
- R. Schatz, A. Sackl, C. Timmerer, and B. Gardlo, “Towards subjective quality of experience assessment for omnidirectional video streaming,” in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017, pp. 1–6.
- A. Singla, S. Fremerey, W. Robitza, P. Lebreton, and A. Raake, “Comparison of Subjective Quality Evaluation for HEVC Encoded Omnidirectional Videos at Different Bit-rates for UHD and FHD Resolution,” in Proceedings of the on Thematic Workshops of ACM Multimedia 2017 - Thematic Workshops ’17, 2017, pp. 511–519.
- A. Singla, S. Fremerey, W. Robitza, and A. Raake, “Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays,” 2017 9th Int. Conf. Qual. Multimed. Exp. QoMEX 2017, 2017.
- M. Xu, C. Li, Z. Wang, and Z. Chen, “Visual Quality Assessment of Panoramic Video,” pp. 1–12, arXiv:1709.06342, Sep. 2017.
- A.-F. Perrin, C. Bist, R. Cozot, and T. Ebrahimi, “Measuring quality of omnidirectional high dynamic range content,” in SPIE Applications of Digital Image Processing XL, 2017, p. 38.
- A. Singla, W. Robitza, and A. Raake, “Comparison of Subjective Quality Evaluation Methods for Omnidirectional Videos with DSIS and Modified ACR,” Hum. Vis. Electron. Imaging, 2018.

References
2/3

- B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, “Subjective and objective quality assessment of panoramic videos in virtual reality environments,” 2017 IEEE Int. Conf. Multimed. Expo Work. ICMEW 201, pp. 163–168, 2017.
- M. Huang *et al.*, “Modeling the Perceptual Quality of Immersive Images Rendered on Head Mounted Displays: Resolution and Compression,” *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 6039–6050, 2018.
- F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, and Y. Wang, “A Subjective Study of Viewer Navigation Behaviors When Watching 360-Degree Videos on Computers,” *2018 IEEE Int. Conf. Multimed. Expo*, pp. 1–6.
- Singla, A., Fremerey, S., Raake, A., List, P. & Feiten, B. (2017). AhG8: Measurement of User Exploration Behavior for Omnidirectional (360°) Videos with a Head Mounted Display.
- C. Ozcinar and A. Smolic, “Visual Attention in Omnidirectional Video for Virtual Reality Applications,” in 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), 2018, no. August, pp. 1–6.
- MacQuarrie, A., & Steed, A. (2017, March). Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In Virtual Reality (VR), 2017 IEEE (pp. 45-54). IEE
- 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
3/3

- M.H. Pinson, M. Sullivan, and A. Catellier, “Immersive audiovisual subjective testing,” Proc. VPQM, 2014
- Garcia, M-N., et al. "Quality of experience and HTTP adaptive streaming: A review of subjective studies." Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on. IEEE, 2014.
- Staelens, N., De Meulenaere, J., Claeys, M., Van Wallendael, G., Van den Broeck, W., De Cock, J., ... & De Turck, F. (2014). Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices. IEEE Transactions on Broadcasting, 60(4), 707-714.
- E. Guervós, J. J. Ruiz, P. Pérez, J.A. Muñoz, C. Díaz, and N. García, “Using 360 VR Video to Improve the Learning Experience in Veterinary Medicine University Degree”, Hum. Vis. Electron. Imaging, 2019
- Gutiérrez, J., Pérez, P., Jaureguizar, F., Cabrera, J., & García, N. (2011, May). Subjective assessment of the impact of transmission errors in 3DTV compared to HDTV. In 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011 (pp. 1-4). IEEE.
- C. Chen, L.K. Choi, G. de Veciana, C. Caramanis, R.W. Heath, and Al.C. Bovik, “A dynamic system model of time-varying subjective quality of video streams over HTTP,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3602–3606, 2013
- Y. Rai, J. Gutiérrez, and P. Le Callet, “A dataset of head and eye movements for 360 degree images,” in Proceedings of the 8th ACM Multimedia Systems Conference, MMSys 2017, 2017.