Towards High Resolution Image and Video Quality Assessment in the Crowd

Rakesh Rao Ramachandra Rao, Steve Göring, Alexander Raake

Audiovisual Technology Group, Technische Universität Ilmenau, Germany
VQEG

December 19, 2023

TECHNISCHE UNIVERSITÄT
ILMENAU

Motivation

- Lab studies → time-consuming and expensive
- Non-feasibility of lab studies due to external factors, e.g. COVID-19
- Need for large groundtruth for video quality model development
- Applicability of crowdsourcing studies for quality assessment
  - Focus: high-resolution images/videos **UHD-1/4K**
  - Adaptation of the test design
  - Comparison with lab test required

Proposed Approach

- Challenges
  - Lack of control on the appropriate hardware for seamless playout and display device
  - Varying test environment (lighting, viewing distance): not handled in this study
- Potential solutions
  - Displaying the crop of the most salient regions in a scene
  - Alternatives to playing out lossless versions of videos, e.g.: choose a transcoding setting that doesn't affect the visual quality of the encoded video
- Proposed approach
  - Images: use different patches
  - Videos
    - Display a **pre-defined center crop** of losslessly upscaled videos (AVPVS) → to handle varying display devices in crowdsourcing context (c.f. [1])
    - Encode the pre-defined center crop of AVPVS using H.264 with a pre-defined CRF → handle lack of appropriate playout hardware

High-Resolution Image Quality Assessment

Dataset – Overview

- Source images: 39 UHD-1/4K frames extracted from UHD-1/4K videos cropped to 2160 × 2160, different genres
- Encoding: 1-pass HEVC CRF encoding (HEVC chosen as it outperforms JPEG)
- Processed images: 371 images encoded with H.265
- Test methodology: ACR (*ITU-T* 2014)
- # Participants in lab test: 21

Crowdsourcing Test

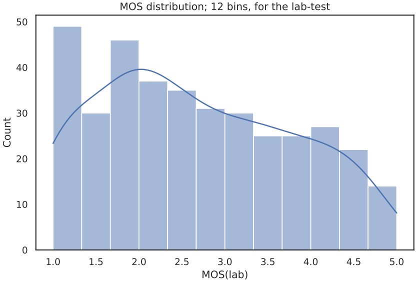
- 2160×2160 image sampled into 4 1080×1080
- Test duration: ≈15 minutes
- Pre-test questionnaire
  - Age range; self-judged visual acuity on an ACR-scale
  - Device type used in test (Phone, Tablet, Laptop, Desktop)
  - Test environment ("Alone in a quiet room", "Some noise and distractions" and "Significant noise and distractions")
- Each participant rated 150 randomly selected patches out of 1484 patches
- No training phase

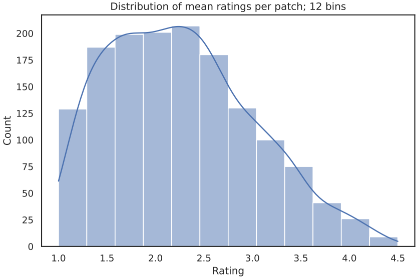
Crowdsourcing Test Results

- Most participants: environment with “less distractions”
- Majority subjects: age range: 18 – 39 years
- # Participants: 238 (recruited via university mailing lists)
- Average ratings per patch: 17

Lab vs. Crowd Test Comparison (1)

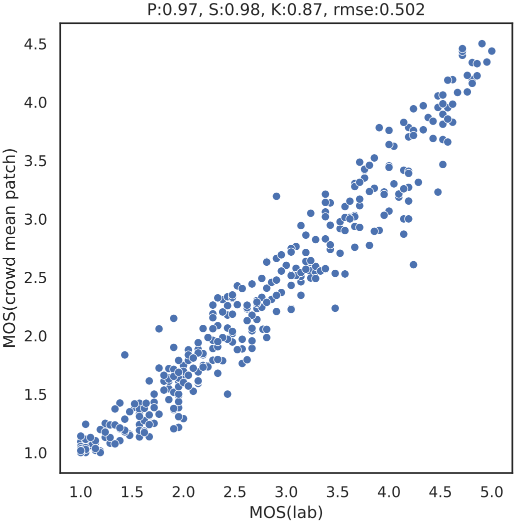
TECHNISCHE UNIVERSITÄT
ILMENAU


MOS distribution; 12 bins, for the lab-test


Distribution of mean ratings per patch; 12 bins

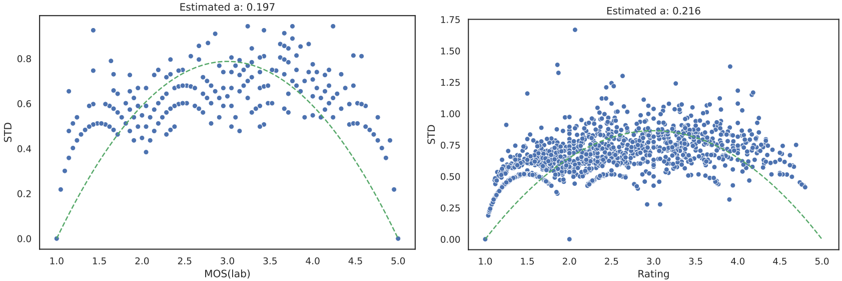
- Participants in crowd test more critical

Lab vs. Crowd Test Comparison (2)



- Correlation: lab and crowd tests: 0.97

Lab vs. Crowd Test Comparison (3)



- SOS analysis: $a_{lab} = 0.197$ and $a_{crowd} = 0.216$

Short-term Video Quality Assessment

Dataset – Overview

test_1 of AVT-VQDB-UHD-1 (*Rao et al.* 2019)

- 540p center crop
- Lab test for comparison
  - Source videos: 6 different videos; each: 10 s ; 3840 × 2160; 60 fps
  - Codecs used: H.264, H.265, VP9
  - Encoding resolutions: 360p to 2160p
- Total number of processed video sequences (PVS): 180
- Participants in lab test: 29
- Outliers in lab test: 0 (Pearson correlation (PCC) $>$ 0.75)

Crowdsourcing Platform and Subject Recruitment

- Used tool: AVrateVoyager[1]
- Subject recruitment
  - $>$ 90% of subjects recruited from university body (staff+students) via email lists
  - Remaining participants from people known to authors

[1] `https://github.com/Telecommunication-Telemedia-Assessment/AVrateVoyager`

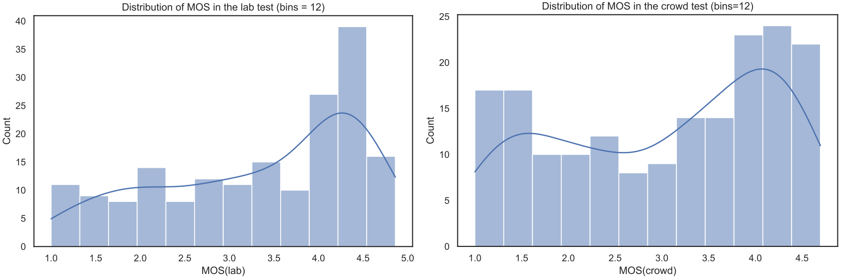
Test Procedure

- Test duration: $\approx$15 minutes
- Pre-test questionnaire
  - Age range; self-judged visual acuity on an ACR-scale
  - Device type used in test (Phone, Tablet, Laptop, Desktop)
  - Test environment (“Alone in a quiet room”, “Some noise and distractions” and “Significant noise and distractions”)
- Checks
  - Minimum device resolution: $720p$
- Each participant rated 30 PVSs randomly selected out of the 180 PVSs
- No training phase

Crowdsourcing Test Results

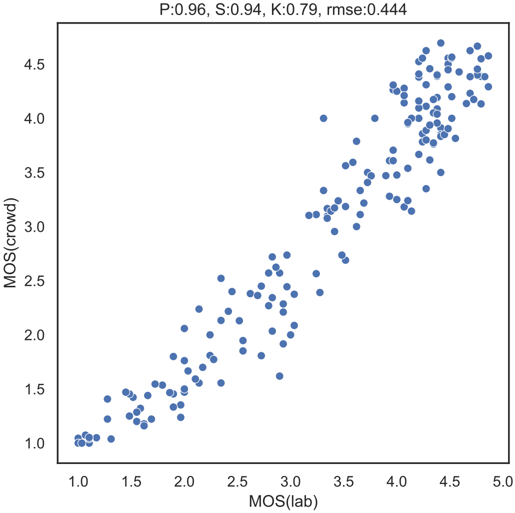
- Most participants: environment with “less distractions”
- Majority subjects: age range: 18 – 39 years
- ≈ 18% of subjects: device with a resolution of full-HD or higher
- # Participants: 175
- Outliers: 19 (PCC > 0.75 as used in lab test)
- Average ratings per PVS: 22.15

Lab vs. Crowd Test Comparison (1)



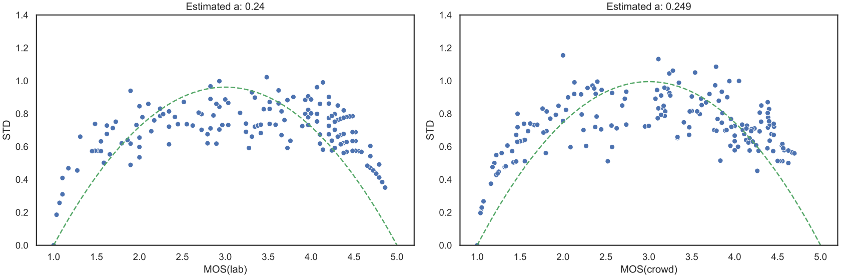
- Participants in crowd test more critical

Lab vs. Crowd Test Comparison (2)



- Correlation: lab and crowd tests: 0.96

Lab vs. Crowd Test Comparison (3)



- SOS analysis: $a_{lab} = 0.240$ and $a_{crowd} = 0.249$

Overall Integral Quality Assessment

Dataset – Overview

test_2 of PNATS-UHD-1-Long (*Ramachandra Rao et al.* 2023)

- $720p$ center crop used
- Lab test for comparison
  - Source videos: 30 different videos; each: 2 min ; 3840×2160
  - Codecs used: H.264, H.265, VP9
  - Encoding resolutions: $360p$ to $2160p$
  - Other impairments: initial buffering, stalling, quality switching
- Total number of processed video sequences (PVS): 30
- # Participants in lab test: 31
- Outliers in lab test: 0 (Pearson correlation (PCC) > 0.75)

Test Procedure

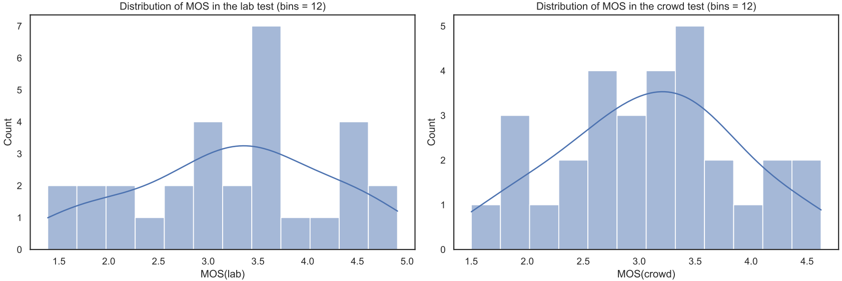
- Used tool: AVrateVoyager[2]
- Test duration: ≈15 minutes
- Pre-test questionnaire + checks: same as short-term video quality test
- Training phase: 1 video *rightarrow* showcasing all impairments
- Test phase: 5 PVSs randomly selected out of the 30 PVSs

[2] `https://github.com/Telecommunication-Telemedia-Assessment/AVrateVoyager`

Crowdsourcing Test Results

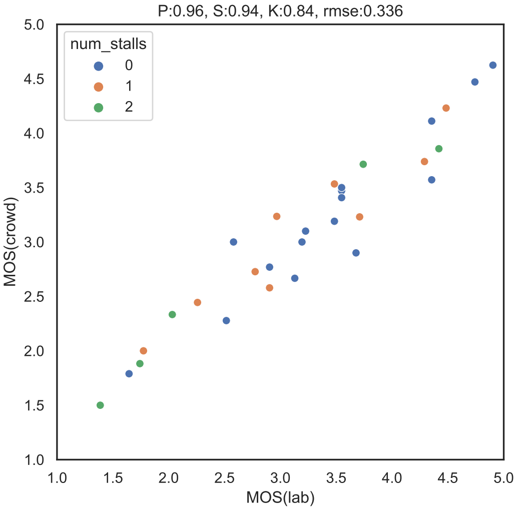
- Participant recruitment via clickworkers
- # Participants: 100
- Most participants: environment with “less distractions”
- $<$ 10% of subjects: device with a resolution of full-HD or higher
- Average ratings per PVS: 17.2

Lab vs. Crowd Test Comparison (1)



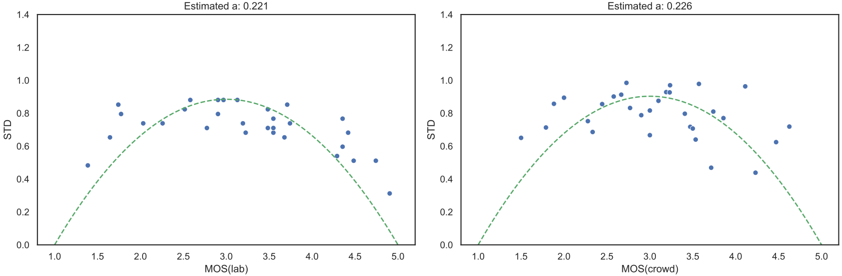
- Participants in crowd test more critical

Lab vs. Crowd Test Comparison (2)



- Correlation: lab and crowd tests: 0.96

Lab vs. Crowd Test Comparison (3)



- SOS analysis [3]: $a_{lab} = 0.221$ and $a_{crowd} = 0.226$

Conclusion

- Proposed method to assess quality of high-resolution images and videos in crowd
- Results show good correlation between lab and crowd tests
  - High PCC; similar SOS parameter values
  - Data publicly available

References I

[1] S. Göring et al. “cencro – Speedup of Video Quality Calculation using Center Cropping”. In: *21st IEEE IEEE ISM*. Dec. 2019, pp. 1–8.

[2] T. Hoßfeld et al. “Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing”. In: *IEEE Transactions on Multimedia* 16.2 (2014), pp. 541–558.

[3] T. Hoßfeld et al. “SOS: The MOS is not enough!” In: *2011 third international workshop on quality of multimedia experience*. IEEE. 2011, pp. 131–136.

[4] V. Hosu et al. “KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment”. In: *IEEE Transactions on Image Processing* 29 (2020).

References II

[5] V. Hosu et al. “The Konstanz natural video database (KoNViD-1k)”. In: *QoMEX*. IEEE. 2017.

[6] ITU-T. *Recommendation ITU-R BT.500-13 – Methodology for the subjective assessment of the quality of television pictures*. Tech. rep. International Telecommunication Union, 2014.

[7] C. Keimel et al. “QualityCrowd — A framework for crowd-based quality evaluation”. In: *2012 Picture Coding Symposium*. 2012, pp. 245–248.

[8] B. Rainer et al. “Quality of Experience of Web-Based Adaptive HTTP Streaming Clients in Real-World Environments Using Crowdsourcing”. In: *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*. VideoNext ’14. Sydney, Australia: ACM, 2014.

References III

[9] R. R. Ramachandra Rao et al. “PNATS-UHD-1-Long: An Open Video Quality Dataset for Long Sequences for HTTP-based Adaptive Streaming QoE Assessment”. In: *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023, pp. 252–257.

[10] R. Rao Ramachandra Rao et al. “AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1”. In: *21st IEEE ISM*. Dec. 2019, pp. 1–8.

[11] F. Ribeiro et al. “CROWDMOS: An approach for crowdsourcing mean opinion score studies”. In: *2011 ICASSP*. 2011.

[12] M. Shahid et al. “Crowdsourcing based subjective quality assessment of adaptive video streaming”. In: *2014 QoMEX*. 2014, pp. 53–54.

References IV

[13] Z. Sinno et al. “Large-Scale Study of Perceptual Video Quality”. In: *IEEE Transactions on Image Processing* (2019).

[14] M. Uhrina et al. “QoE on H.264 and H.265: Crowdsourcing versus Laboratory Testing”. In: *2020 30th International Conference Radioelektronika (RADIOELEKTRONIKA)*. 2020, pp. 1–6.

Thank you for your attention

...... are there any questions?

Back-up

Crowdsourcing for Video Quality Assessment – Overview

- Best practices for crowdsourcing QoE testing (*Hoßfeld et al.* [2])
- Crowdsourcing as a viable alternative for perceptual assessment of image, video and audiovisual content (*Hosu et al.* [4], *Hosu et al.* [5], *Sinno et al.* [13])
- *Keimel et al.* [7], *Ribeiro et al.* [11]: Different crowdsourcing platforms
- *Shahid et al.* [12], *Rainer et al.* [8]: Crowdsourcing in HTTP-based adaptive streaming (HAS) context
- *Uhrina et al.* [14]: Investigation of feasibility of unpaid
  crowdsourcing approach as an alternative for lab-based tests;
  reports a correlation of > 0.92 between lab and “crowd” tests