“Discriminability–Experimental Cost” tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR–HR

Andréas Pastor[1], Pierre David[1,2], Ioannis Katsavounidis[3], Lukáš Krasula[4], Hassene Tmar[3], Patrick Le Callet[1,5]

[1] Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2] Capacités SAS [3] Meta [4] Netflix [5] Institut universitaire de France (IUF)

Key points

Importance of subjective data to validate systems:

- Need precise and accurate estimates
- Need insight on measured subjective uncertainty: MOS confidence interval size
- Be the most cost–efficient for a subjective study budget

From an available budget, how many video sequences can we afford to test knowing the efficiency of a subjective quality assessment protocol and its average required annotation time per observer?

Summary of the experiments: test sequences, displays, environment, subjective quality assessment methodologies, …

Summary of the experiments: test sequences

Sources:

- 4 SDR HD video sequences: 10sec 10bits yuv420 60fps
- 5 SDR UHD video sequences: 10sec 10bits yuv420 30–60fps (1SRC: 60fps)
- 5 HDR UHD video sequences: 10sec 10bits yuv420 60fps

Encoded with Random Access (RA) mode of modern video encoding implementations

Summary of the experiments: display specifications

Device	Sony XG8096	Sony 55X85J	Sony XR-65A95K
Diagonal size	55”	55”	65”
Resolution	3840x2160	3840x2160	3840x2160
Technology	LCD, LED	LCD, LED	QD-OLED
Refresh rate	60Hz	100Hz	100Hz
Peak Luminance	-	-	965 nits*
Usage	HD/UHD SDR	HD/UHD SDR	UHD HDR

* Peak Luminance obtained after calibration

Summary of the experiments: environment

- SDR viewing environment: ITU-T Rec. BT.500
  - TV calibration with color probe over 461 color references
  - D65 white at 120 cd/m²
  - room controlled lighting 15 cd/m²
  - observers placed at 1.6 times the screen height
- HDR viewing environment: ITU-R Rec. BT.2100
  - TV calibration with Calman Home for Sony
  - D65 white at 950 cd/m²
  - room controlled lighting 5 cd/m²
  - observers placed at 1.6 times the screen height

Summary of the experiments: methodologies

DCR methodology: ITU-T Rec P.910

- 11-grade DCR scale from Expert Viewing Procol
- 1 repetition
- 2-second pause transitions
- calibration over 3 PVS: high, mid, low impairment

$$| \, Source \, | \, Coded \; sequence \; A \, | \, Source \, | \, Coded \; sequence \; A \, |$$

Scores	Impairment items	Levels
10	Imperceptible	
9	Slightly perceptible	Somewhere
8		Everywhere
7	Perceptible	Somewhere
6		Everywhere
5	Clearly perceptible	Somewhere
4		Everywhere
3	Annoying	Somewhere
2		Everywhere
1	Severely annoying	Somewhere
0		Everywhere

ITU-R BT.500-14 Expert Viewing Protocol 11-grade DCR scale

Summary of the experiments: methodologies

ACR-HR methodology: ITU-T Rec. P.910

- 5-grade ACR scale
- no repetition
- calibration over 3 PVS: low, mid, high quality
- conversion of raw OS to DMOS

Scores	Quality items
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Summary of the experiments: viewing sessions

*Reminder: 40 PVS for SDR HD, 50 PVS for SDR UHD and HDR UHD each

DCR test viewing sessions:

- For the DCR HD test:
  - 45min sessions: 40 PVS in 1 session – HD 1
  - 30 observers
- For the DCR UHD test:
  - 2x 30min sessions: 25PVS each – UHD 1/2
  - 2x 30 observers
- For the DCR HDR test:
  - 2x 30min sessions: 25PVS each – HDR 1/2
  - 2x 30 observers

DCR Sessions	HD	HD 1
	UHD	UHD 1
		UHD 2
	HDR	HDR 1
		HDR 2
ACR-HR Sessions	HD/UHD	-
	HDR	-

Budget DCR: 30 “45-min” observers + 120 “30-min” observers

Summary of the experiments: viewing sessions

*Reminder: 40 PVS for SDR HD, 50 PVS for SDR UHD and HDR UHD each

Budget DCR: 30 “45-min” observers + 120 “30-min” observers

ACR-HR test viewing sessions:

- For ACR-HR HD + UHD test:
  - 30 min sessions: 44 HD PVS, 3-min break then, 55 UHD PVS* (4–5 hidden-references)
  - 90 observers
- For ACR-HR HDR:
  - 45 min sessions: 55 HDR PVS, 3-min break then again 55 HDR PVS** (5 hidden-references)
  - 45 observers

Budget: 45 “45-min” observers + 90 “30-min” observers

DCR Sessions	HD	HD 1
	UHD	UHD 1
		UHD 2
	HDR	HDR 1
		HDR 2
ACR-HR Sessions	HD/UHD	-
	HDR	-

DCR and ACR-HR DMOS aggregation method

- regular MOS: raw opinion scores average
- BR-SR MOS: outlier rejection technique from ITU-T Rec. P.913
- BISCWIT MOS: Netflix SUREAL “MLE CO-AP2” algorithm from ITU P.913-12.6

For DCR, scaling from 0–10 to 1–5 (when needed):

$$MOS^{j}_{1to5} = 4 \times \frac{MOS^{j} - minScale}{maxScale - minScale} + 1$$

j: PVS ids, minScale = 0, maxScale = 10

ACR MOS to DMOS:

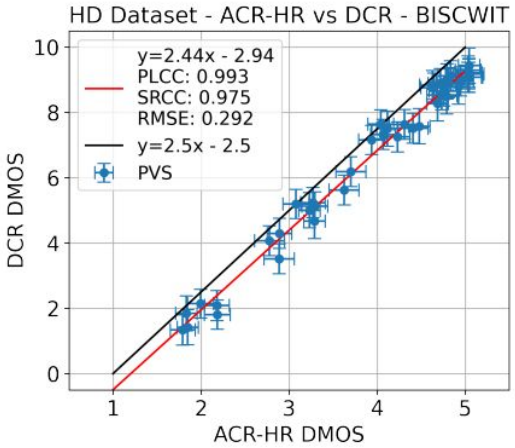
$$DMOS^{j}_{o} = 5 - (MOS^{ref}_{o} - MOS^{j}_{o})$$

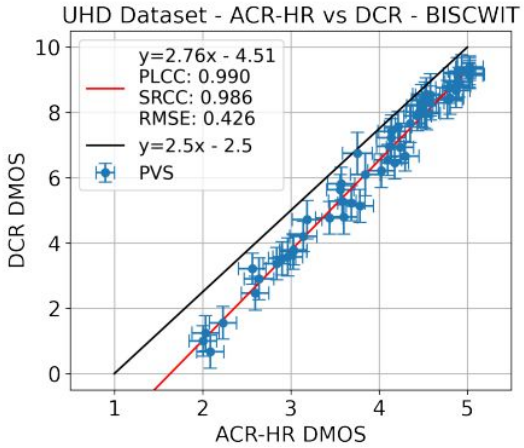
$$DMOS^{j} = \frac{1}{N} \sum_{o=1}^{N} DMOS^{j}_{o}$$

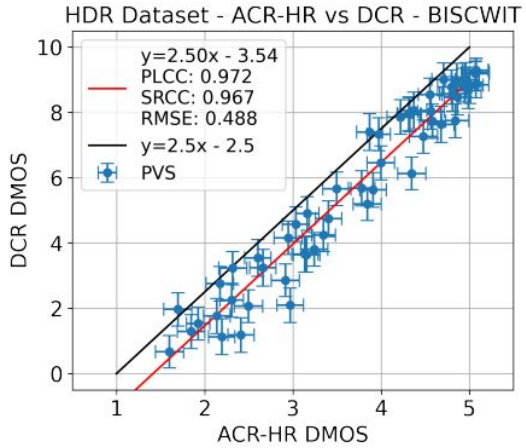
o: observer ids,
j: PVS ids

Results and analysis: usage of the scales, “Discriminability–Experimental Cost” tradeoff

Usage of the scale across HD, UHD, HDR tests







red and black line close = similar overall usage of the scale range

slightly more disagreement in HDR

Naive participants are using both scale ranges similarly with a slight benefit for DCR on UHD

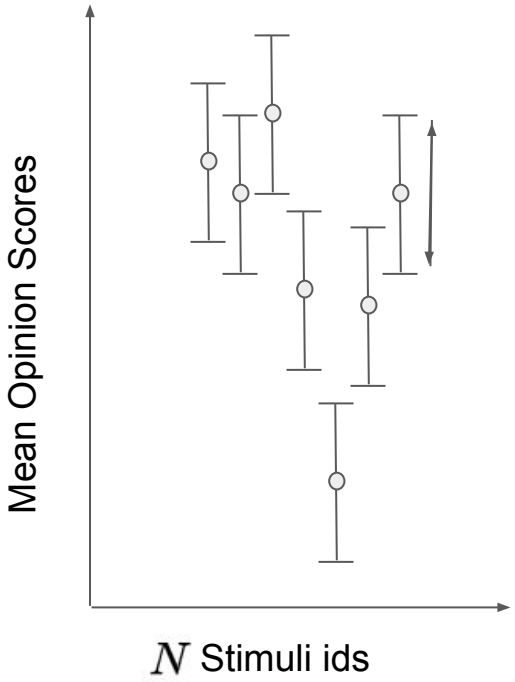
Going deeper into MOS analysis:

Subjective data precision and subjective methodologies efficiency

Data precision: mean MOS Confidence Interval size

Mean CI: average over all the estimated MOS Confidence Intervals - smaller is better

$$Mean_{CI} = \frac{1}{N} \sum_{n=1}^{N} CI_{MOS}^{n}$$



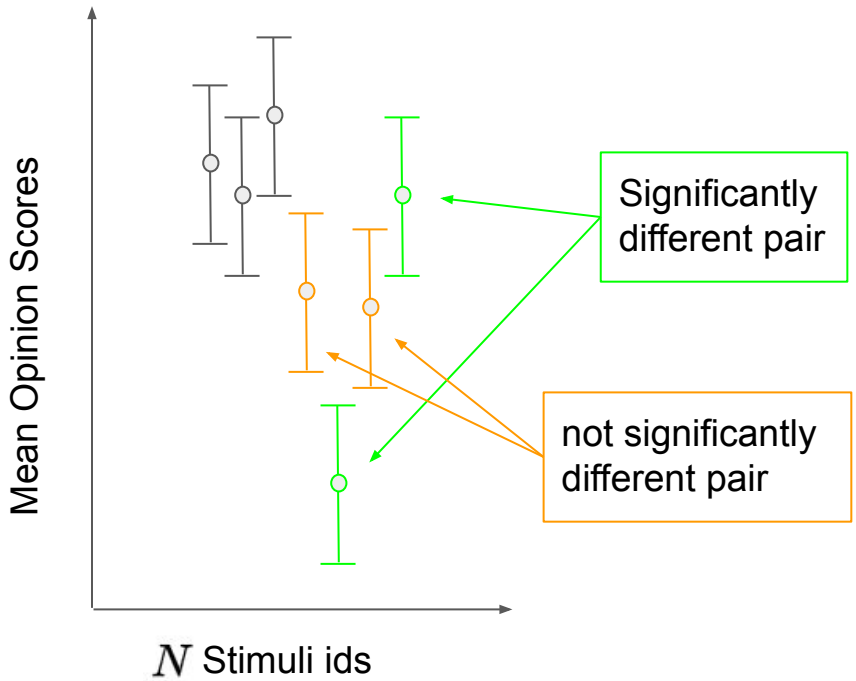
Data precision: discriminability ratio on MOS

T-test analysis on pairs of stimuli MOS

Discriminability ratio: number of significantly different pairs among all the possible ones - higher is better

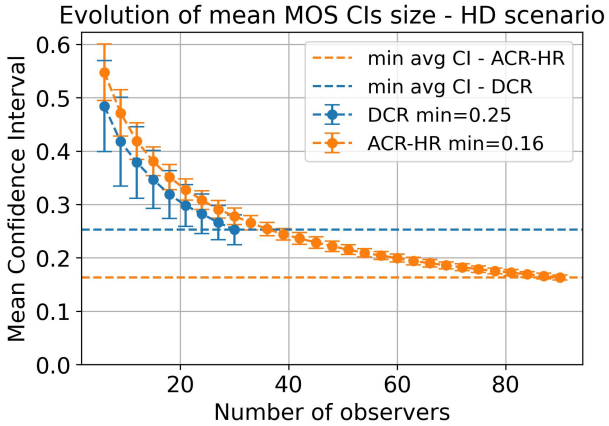
$$D_{ratio} = \frac{1}{M} \sum_{m=1}^{M} Sig_m$$

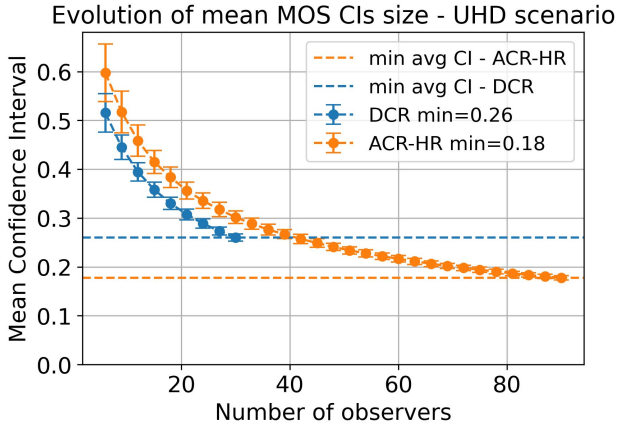
$$M = \frac{N * (N-1)}{2}$$

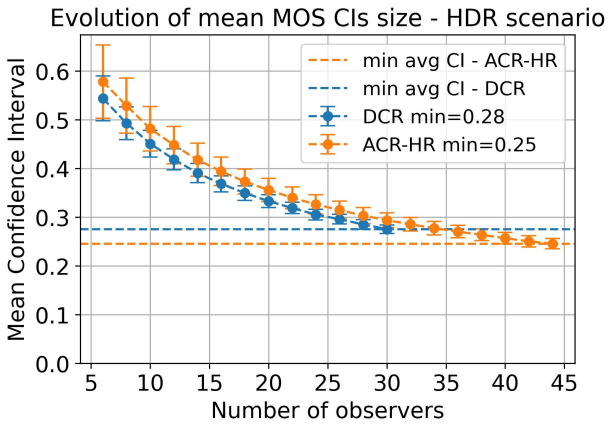


[1] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, “Comparison of subjective methods for quality assessment of 3D graphics in virtual reality,” ACM Transactions on Applied Perception (TAP), vol. 18, no. 1, pp. 1–23, 2020.

Mean MOS Confidence Interval size evolution





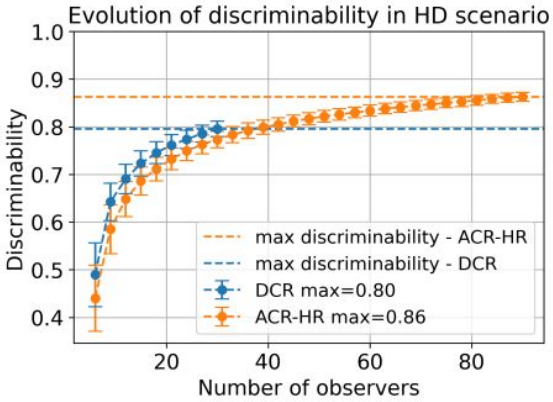


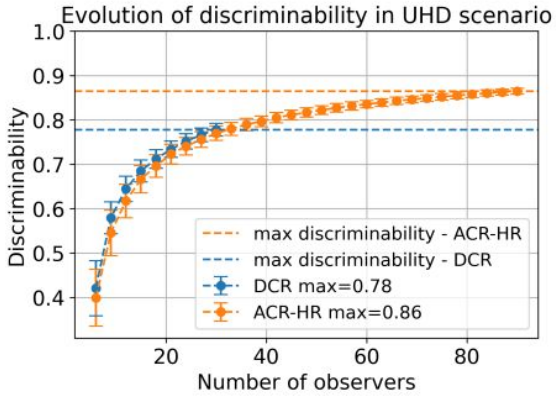
X-axis: selection of K observers with replacement

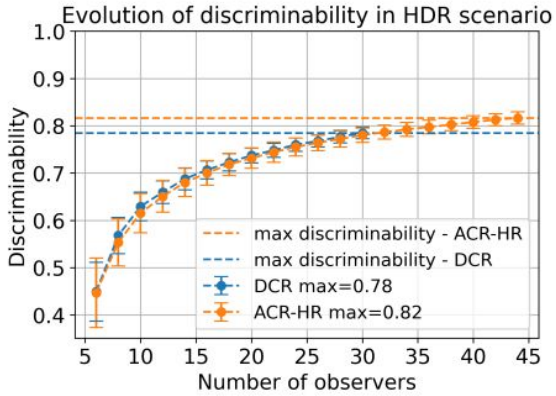
Y-axis: Mean confidence interval obtained over 1000 simulations with K observers

For the 3 scenarios, at same number of observers, DCR test MOS CI are smaller than MOS CI from ACR–HR test.

MOS discriminability evolution







X-axis: selection of K observers with replacement

Y-axis: Discriminability obtained over 1000 simulations with K observers

For HD scenario, DCR achieves slightly greater discriminability than ACR–HR at fixed number of observers. However, for UHD and HDR scenarios, the difference is small, as the two curves overlap in their uncertainty estimates.

“Discriminability–Experimental Cost” tradeoff

Importance of cost/budget in subjective data collection: need to be efficient

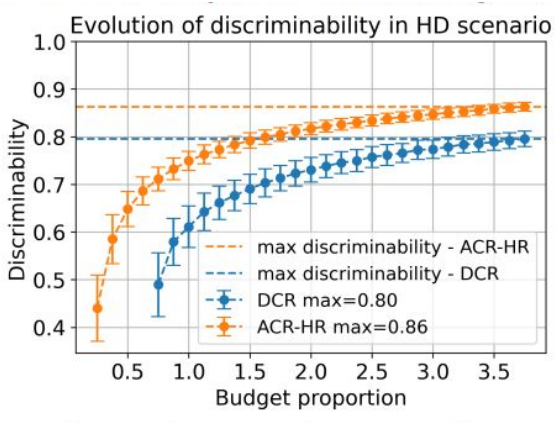
What is the evolution of discriminability with increasing budget proportion?

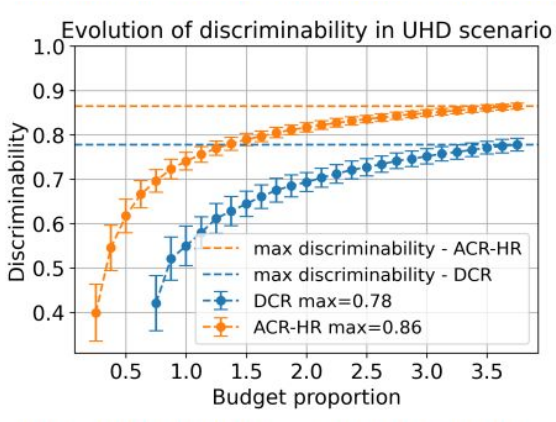
Introduce $B_{prop}^{k,method}$: ratio between cost to recruit K observers with {ACR–HR; DCR} and the cost to recruit 24 observers for an ACR–HR test.

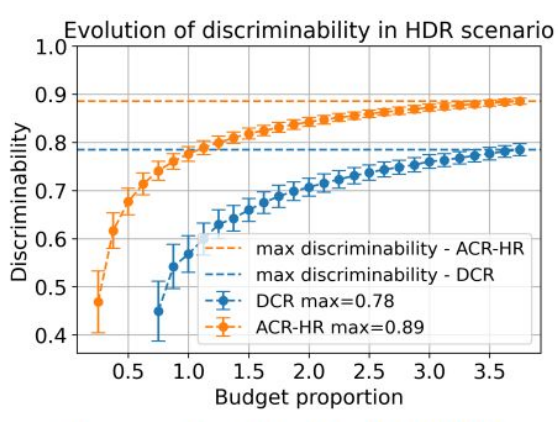
$$B_{prop}^{k,method} = \frac{C_k^{method}}{C_{24}^{ACR-HR}}$$

*24 from ITU recommendations

MOS “Discriminability–Experimental Cost” tradeoff

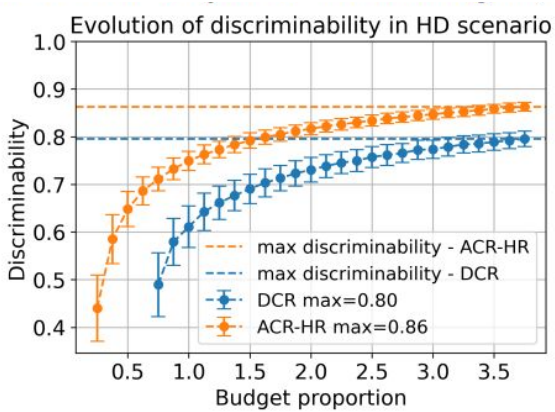


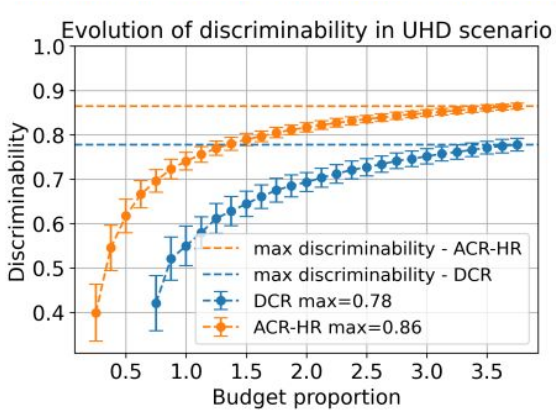


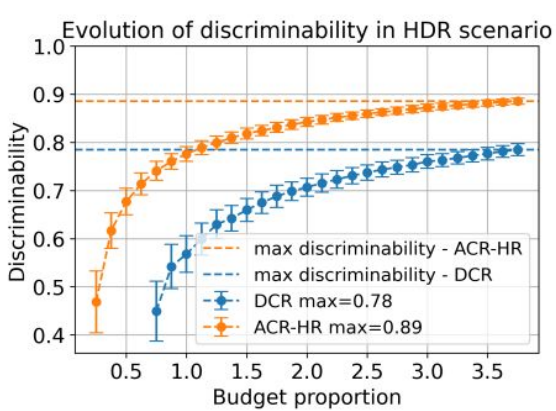


Here, a budget of 3.75 x C_{24}^{ACR-HR} allows to recruit 90 ACR–HR observers or 30 DCR observers

MOS “Discriminability–Experimental Cost” tradeoff



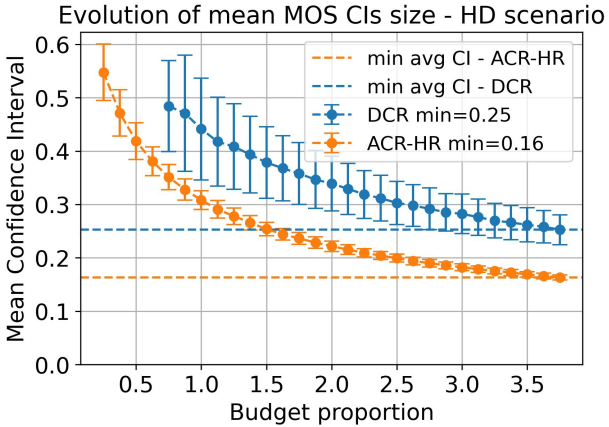


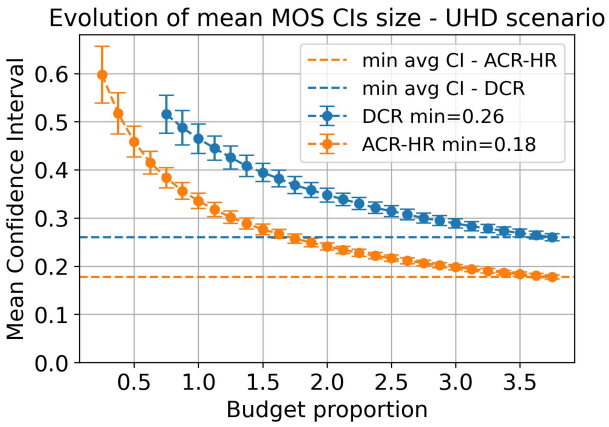


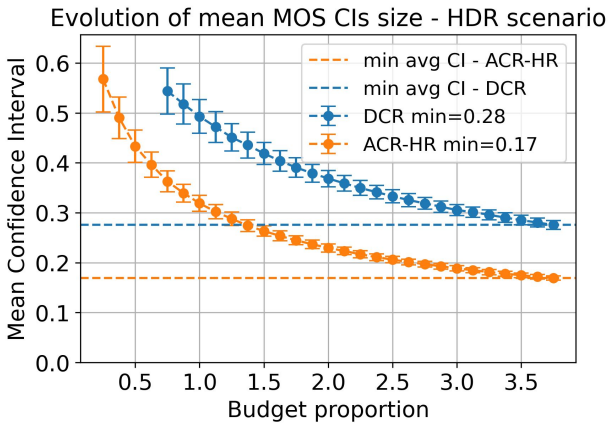
For HD: 3.75 C_{24}^{ACR-HR} budget, ACR–HR test gave us 0.86 discriminability score vs. 0.80 in DCR–EVP test.

30 – 40% of the budget is sufficient with ACR-HR to achieve discriminability obtained with DCR at same budget: ACR-HR is 3 times more efficient than DCR–EVP

MOS “Confidence Interval–Experimental Cost” tradeoff







For HD: 3.75 C_{24}^{ACR-HR} budget, ACR–HR test gave us 0.16 mean CI vs. 0.25 in DCR–EVP test.

30 – 40% of the budget is sufficient with ACR-HR to achieve mean CI obtained with DCR at same budget: ACR-HR is 3 times more efficient than DCR–EVP

Conclusion

- comparison of DCR–EVP and ACR–HR, both can retrieve accurate MOS estimate
- **At identical observer numbers,** DCR with EVP scale gives slightly smaller confidence intervals and better discriminability
- **At an identical budget,** ACR-HR methodology gives better results than DCR with EVP and can reduce experimental cost by a factor of 3

New insight into:

- efficiency of subjective quality assessment protocols on modern codec validation
- accuracy of MOS through discriminability computation and comparison across datasets

More intuitions

We still believe that DCR can be competitive with ACR–HR:

- with classical DCR 5pt scale
- no repetition

Past research works have shown the benefit of DCR over ACR for other multimedia applications (or to achieve higher discriminability. See our next presentation today)

[1] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, “Comparison of subjective methods for quality assessment of 3D graphics in virtual reality,” ACM Transactions on Applied Perception (TAP), vol. 18, no. 1, pp. 1–23, 2020.
[2] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi, “Performance comparisons of subjective quality assessment methods for mobile video,” in 2010 Second international workshop on quality of multimedia experience (QoMEX). IEEE, 2010, pp. 82–87
[3] Taichi Kawano, Kazuhisa Yamagishi, and Takanori Hayashi, “Performance comparison of subjective assessment methods for 3d video quality,” in 2012 Fourth International Workshop on Quality of Multimedia Experience, 2012, pp. 218–223
[4] Ongoing work on 360 video and 360 audio-video sequences