May 4-8, ICASSP 2020
(45th International Conference on Acoustics, Speech, and Signal Processing)

Non-Experts or Experts?
Statistical Analyses of MOS using DSIS Method

NHK

Yasuko Sugito
NHK
(Japan Broadcasting Corporation)

Marcelo Bertalmío

upf. Universitat Pompeu Fabra Barcelona

DSIS Method and MOS

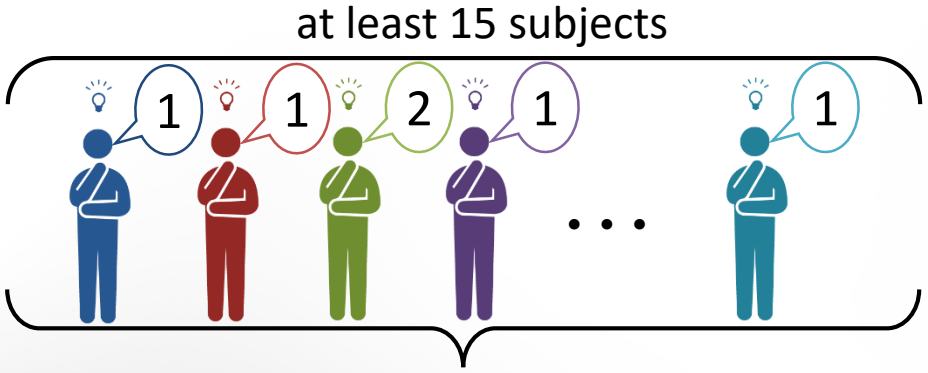
- The double-stimulus impairment scale (**DSIS**) method
  - Described in Rec. ITU-R BT.500 (also in Rec. ITU-T P.910)
  - Frequently applied to subjective assessments of compressed images

reference
(original)

test
(compressed)

Five-grade impairment scale

5	imperceptible
4	perceptible, but not annoying
3	slightly annoying
2	annoying
1	very annoying



the mean opinion score
(**MOS**)

Why Analyze MOS?

- Two main purposes of MOS values
  1. Measure subjective quality of test images
     - Several criteria expressed as MOS values: MOS = 2.5, 3.0, 3.5, and 4.5
     - E.g., image quality is good if its MOS $\geq$ 3.5

☹ **Such criteria are not mentioned in Recs.** (either BT.500 or P.910)

☺ **Discuss statistical meanings of such MOS values**

Why Analyze MOS? 2

- Two main purposes of MOS values
  2. Measure performance of objective image quality metrics

⚠ Subjective experimental conditions should be properly prepared

- Selection of subjects: non-experts or experts?
  - Traditionally believed non-experts are preferable as P.910 (2008) and BT.500-12 (2009)
  - BT.500-14 allows both expert and non-expert subjects depending on purposes

☹ **No sufficient discussion on difference between non-experts and experts**

☺ **Discuss such difference based on analyses of MOS values**

Databases for MOS Analysis

- Used three experimental results using DSIS method
  - One expert (EE) and two non-expert (NE1, NE2) experiments

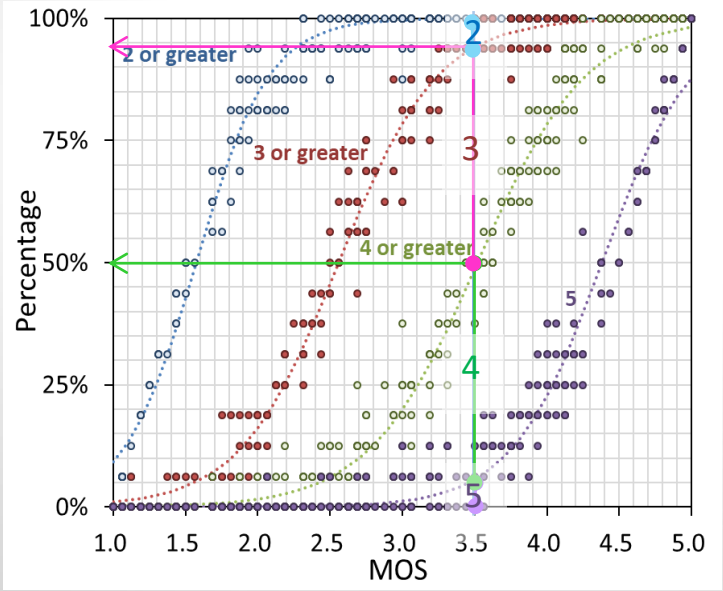
	EE[1]	NE1[2]	NE2[3]
Test images	240 compressed, and 20 original	240 compressed	88 compressed, 6 tone-mapped, and 6 uncompressed
Observers	16 experts	22 non-experts	14 or 15 non-experts
Presentation Method	Original and compressed images side by side for 10 s (SDSCE method)		Original 6 s and compressed 8 s (DSIS method)
Grading Scale	Five-grade impairment scale (DSIS method)		1-100, associated with five-grade scale of DSIS method (Conv. to five-grade score: $\lfloor (OriginalScore - 1)/20 \rfloor + 1$)

- Analyzed relationship between MOS values and score distribution and variance

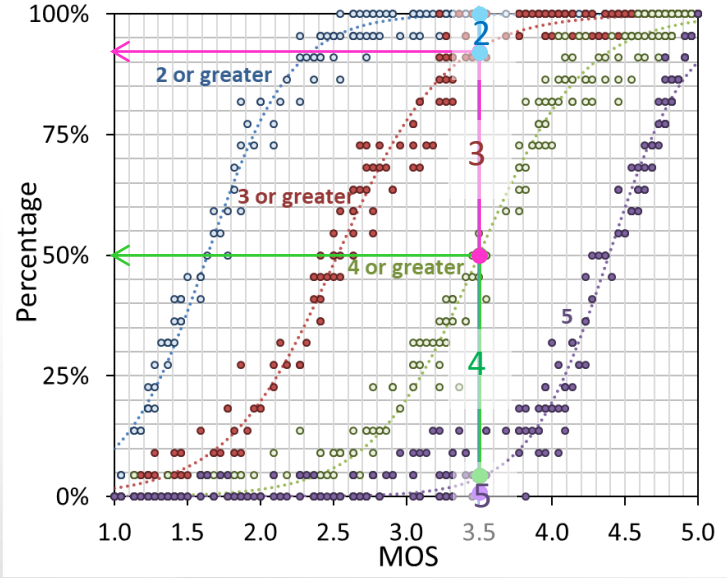
1. Y. Sugito and M. Bertalmio, “Practical use suggests a re-evaluation of HDR objective quality metrics,” 11th QoMEX, Berlin, Germany, 2019.
2. P. Korshunov et al., “Subjective quality assessment database of HDR images compressed with JPEG XT,” 7th QoMEX, Costa Navarino, Messinia, Greece, 2015.
3. E. Zerman et al., “An extensive performance evaluation of full-reference HDR image quality metrics,” Quality and User Experience, vol. 2, no. 1, 2017, p. 16.

Score Distribution per MOS Value

- Relationship between MOS and percentage of scores 2-5, 3-5, 4-5, 5
  - Curve fitting by LSM using logistic function (dotted line) $f(x) = \frac{1}{1 + \exp(-a(x-b))}$



Experts (N=16)



Non-Experts 1 (N=22)

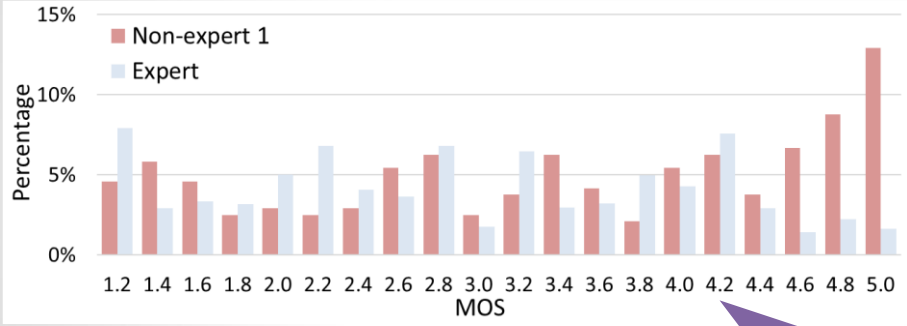
Statistical Meanings of MOS Values

- Slightly difference between non-experts and experts for MOS $\leq$ 3.5

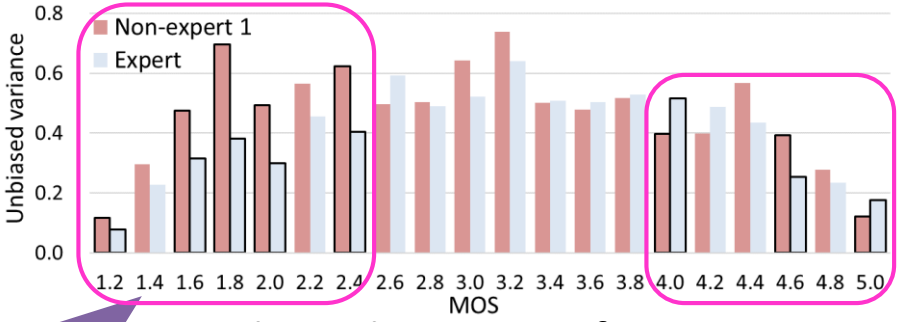
MOS	Criteria	Experts	Non-experts
2.5	lower level below tolerance limit	● Nearly 100% subjects gave scores ≥2 ● ~45% subjects gave scores ≥3	● ~95% subjects gave scores ≥ 2 ● Nearly 50% subjects gave scores ≥3
3.0	lower limit for broadcasting quality (experts)	● Nearly 100% subjects gave scores ≥ 2 ✓ **When MOS < 3.0, there is possibility of 1, very annoying level**	● Nearly 100% subjects, slightly smaller than that of experts, gave scores ≥2
3.5	tolerance limit	● ~95% subjects gave scores ≥ 3 ● ~50% subjects gave scores ≥ 4 ✓ **When MOS > 3.5, > 50% subjects will not consider image as annoying level, and nearly all of remaining subjects will perceive image as barely annoying level**	● > 90% subjects gave scores ≥ 3 ● ~50% subjects gave scores ≥4
4.5	detection limit	● Nearly 100% subjects gave scores ≥ 3 ● > 90% subjects gave scores ≥ 4 ● ~60% subjects gave 5 ✓ 50% subjects gave 5, imperceptible, at MOS~4.4	

Analysis of score variance per MOS value

- Score distribution and variance for each 0.2 range of MOS values for EE and NE1
  - Black-bordered bars: significant difference in variance at 5% sig. level (F-test)



Distribution of MOS values



Unbiased variance of scores

Results of
240 compressed images

- High MOS (> 3.8): variance of non-experts is significantly lower in some cases
- Low MOS (< 2.4): variance of experts always lower; showed significant differences

☺Experts are helpful to determine lower limit of image quality

NHK

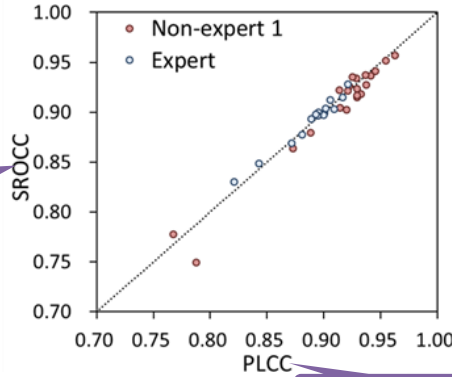
Difference between Non-experts and Experts

- Further consider difference between non-expert and expert subjects
  - Evaluations on compressed images: should be MOS≠5.0
    - 👎Non-experts gave MOS=5.0 (NE1: 3/240 and NE2: 2/88)
    - 👍Experts did not give MOS=5.0 (0/240, Max. MOS=4.81)
  - Evaluations on uncompressed images: should be MOS=5.0
    - 👎Non-experts gave MOS<4.5 (NE2: 2/6)
    - 👍Experts gave MOS=5.0 (5/20, Min. MOS=4.63)

☺**Experts better distinguish difference between original and compressed images**

Difference between Non-experts and Experts 2

- Further consider difference between non-expert and expert subjects
  - Calculated correlations between MOSs and individual scores for NE1 and EE
    - Non-experts: widely spread (PLCC: 0.77-0.96 and SROCC: 0.75-0.96)
    - Experts: consistently high (PLCC: 0.82-0.92 and SROCC: 0.83-0.93)



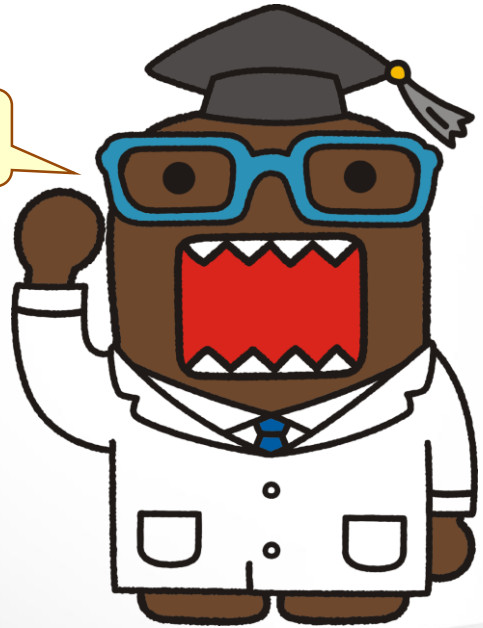
  - Extracted 16 subjects with higher PLCC (0.91-0.96) from NE1
    - Score variance for lower MOS (< 2.4) becomes similar extent to that of experts
    - Score variance for higher MOS (> 4) becomes lower than that of experts
  - Difficult to predict expert trend from results of non-experts

☺**We can perform image quality tests with fewer observers if they are experts**

Conclusions

- Analyzed MOS values of DSIS method
- Showed statistical meanings of MOS values used as criteria of image quality
- Considered difference between non-expert and expert subjects
  - Found that experts can be useful for some purposes
- Type of subjects, non-experts or experts, should be chosen depending on application as described in BT.500
- Continue to analyze other experimental results using DSIS method

Thank you for your attention



E-mail: sugitou.y-gy@nhk.or.jp