Nov. 29th 2018, 6th IEEE Global SIP

Performance Evaluation of Objective Quality Metrics on HLG-Based Image Coding

NHK Science & Technology Research Laboratories

Yasuko Sugito
NHK
sugitou.y-gy@nhk.or.jp

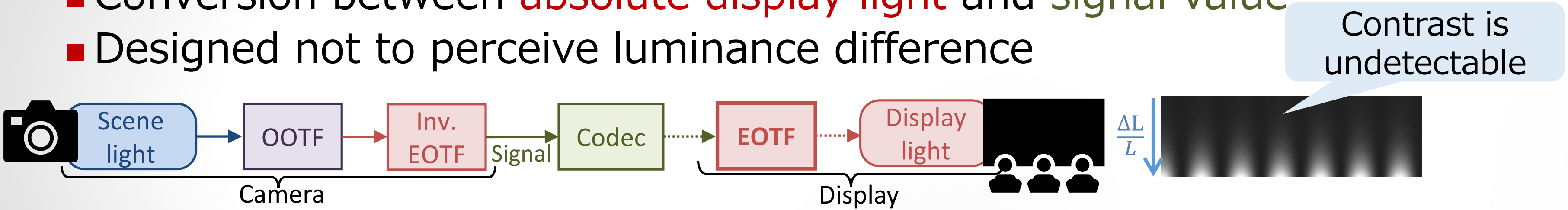
Marcelo Bertalmío

upf. Universitat Pompeu Fabra Barcelona

HDR Image Coding and TF

- High dynamic range (HDR) supports wider range of luminance

1. Contrast sensitivity function (CSF) –based transfer function (TF)
   - Conversion between absolute display light and signal value
   - Designed not to perceive luminance difference


Contrast is undetectable

$\frac{\Delta L}{L}$

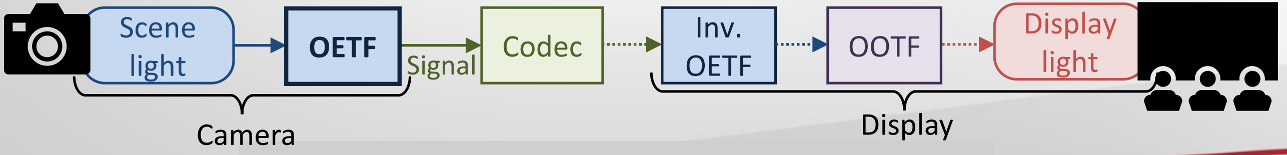
Scene light → OOTF → Inv. EOTF → Signal → Codec → EOTF → Display light

Camera

Display

HDR image coding diagram using perceptual quantizer (PQ) EOTF

2. **Hybrid Log-Gamma (HLG) opto-electronic (OE) TF**
   - Conversion from relative scene light to signal value
   - Designed for backward compatibility with existing SDR displays


Scene light → OETF → Signal → Codec → Inv. OETF → OOTF → Display light

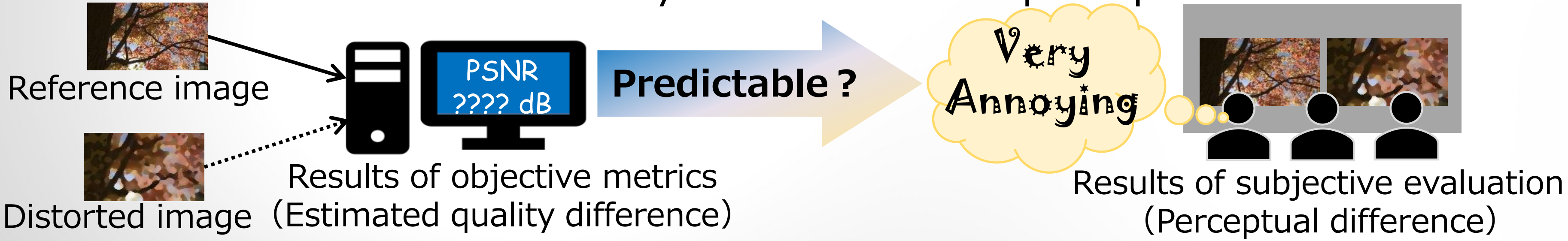
Camera

Display

HLG-based HDR image coding diagram

Objective Quality Metrics

- Objective quality metrics (ex. PSNR) are frequently used for image coding quality assessment
  - Much easier than subjective evaluation experiments
- “Excellent” metrics accurately emulate human perception



Reference image

Distorted image

Results of objective metrics
(Estimated quality difference)

Results of subjective evaluation
(Perceptual difference)

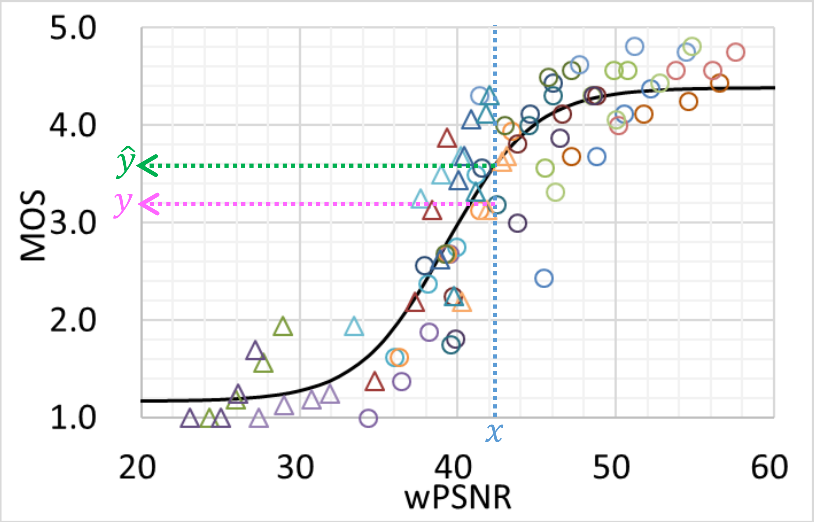
- HDR objective quality metrics have been considered
  - Earlier study* tested metrics for CSF-based image coding
    - HDR-VQM, HDR-VDP-2.2, and PU_MS-SSIM are excellent metrics
- Are these metrics still excellent for **HLG-based image coding**?

*P. Hanhart, M.V. Bernado, M. Pereira, A.M.G. Pinheiro and T. Ebrahimi, “Benchmarking of objective quality metrics for HDR image quality assessment,” EURASIP Journal on Image and Video Processing, 2015(1), pp.1-18, 2015.

Evaluation Method

- Same manner as earlier studies

1. Prepare dataset consists of various distorted images
2. Conduct subjective evaluation experiments, and calculate mean opinion score (MOS): “ground truth data”
3. Calculate objective quality metrics including HLG-based



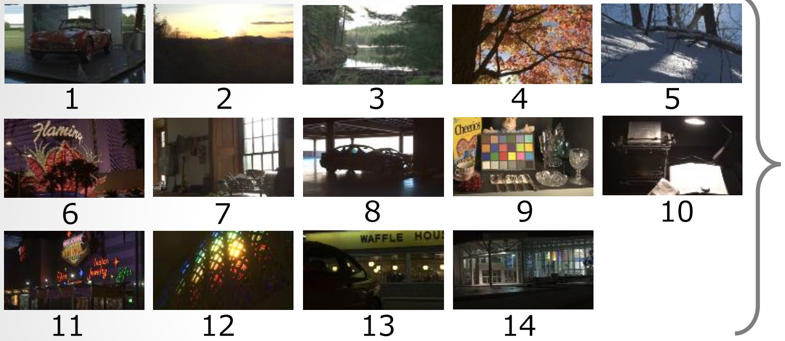
4. Derive logistic function, which calculates predicted MOS $\hat{y}$ from measurement x, with least-square method

$$\hat{y} = a + \frac{b}{1 + \exp(-c(x - d))}$$

5. Assess similarity (correlation coeffs. and mean square error) between true MOS y and predicted MOS $\hat{y}$

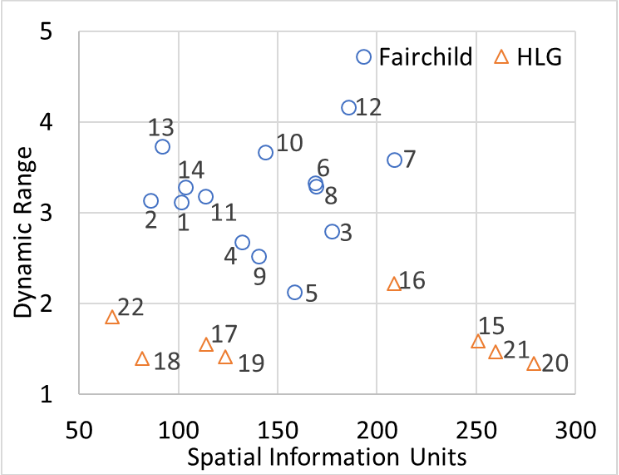
Preparation of Dataset

- 22 various HDR images (cropped 2K)



14 Fairchild images

8 HLG native images



$$Dynamic\ Range = \log_{10}(L_{MAX}/L_{min})$$

- HEVC/H.265 Encoder: HEVC Test Model (HM) 16.17
  - All intra Main 10 (4:2:0/10 bit)
  - Fixed QP: 100, 200, 300, and 400 kbits

Subjective Evaluation Experiments

- 4K HLG monitor (31.1-inch, 1,000 cd/m^2)
  - Viewing distance: 1.5 H (approx. 0.55 m)

- Double stimulus impairment scale method, Variant I (BT.500)
  - Display 2K reference and distorted images side-by-side for 10 s
  - Five-grade scale

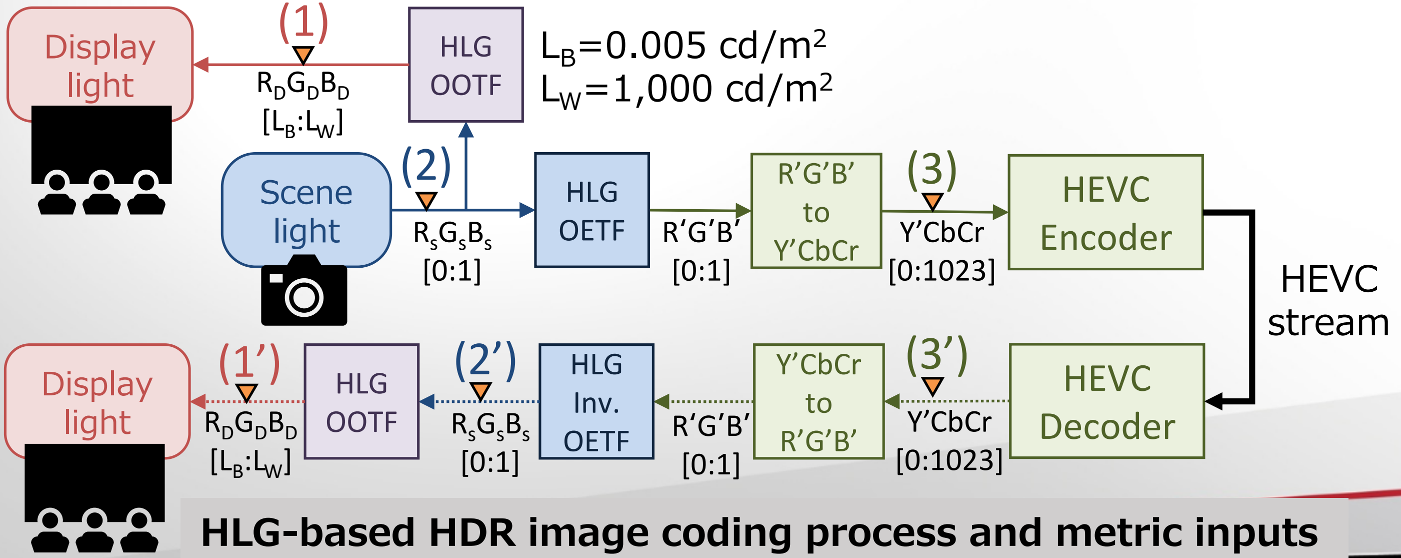
    5 imperceptible
    4 perceptible, but not annoying
    3 slightly annoying
    2 annoying
    1 very annoying
  - Evaluators: 16 video experts

Objective Quality Metrics 1

- Applied 11 types of HDR metrics for luminance component

1. CSF-based metrics: **HDR-VQM**, **HDR-VDP-2.2**, and **PU**_SSIM/**MS-SSIM**
   - Designed to input display light (1)-(1’), Y_D cd/m²
   - Also tested absolute scene light (2)-(2’), Y_{AS} cd/m²
   - **Excellent metrics** in earlier study were included

} 8 types

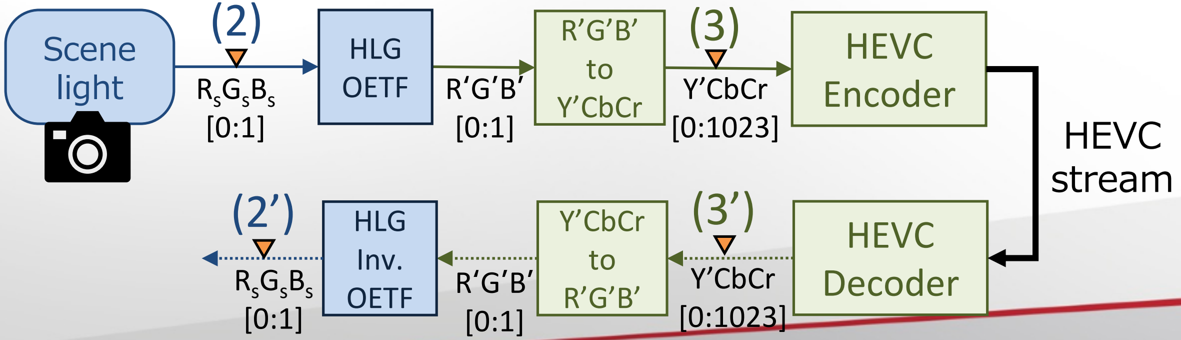


HLG-based HDR image coding process and metric inputs

Objective Quality Metrics 2

- Other 3 types are within HLG-based image coding process

2. HLG-based metrics: HLG_SSIM/MS-SSIM
   - HLG OETF (instead of CSF-based function) + SSIM/MS-SSIM
     - Inputs are scene light (2)-(2’)
3. wPSNR
   - HDR metric used in standardization meeting of VVC
     - PSNR with weight depending on luma value
     - Inputs are HLG Y’CbCr (3)-(3’)



Similarity Results

- HLG_MS-SSIM is the best for HLG-based image coding
- PU_MS-SSIM and HDR-VDP-2.2 show good results
  - HDR-VQM does not

PLCC		SROCC		RMSE	
HLG_M	0.9276	HLG_M	0.9238	HLG_M	0.4463
Y_D_PU_M	0.9175	Y_D_PU_M	0.9164	Y_D_PU_M	0.4751
Y_D_VDP2	0.9163	Y_D_VDP2	0.9146	Y_D_VDP2	0.4783
wPSNR	0.9126	Y_D_PU_S	0.9034	wPSNR	0.4883
Y_D_PU_S	0.8959	wPSNR	0.9009	Y_D_PU_S	0.5307
HLG_S	0.8734	HLG_S	0.8948	HLG_S	0.5817
Y_{AS}_PU_S	0.8613	Y_{AS}_PU_S	0.8545	Y_{AS}_PU_S	0.6068
Y_{AS}_PU_M	0.8599	Y_{AS}_VDP2	0.8421	Y_{AS}_PU_M	0.6097
Y_{AS}_VDP2	0.8460	Y_D_VQM	0.8374	Y_{AS}_VDP2	0.6368
Y_D_VQM	0.8066	Y_{AS}_PU_M	0.8356	Y_D_VQM	0.7060
Y_{AS}_VQM	0.7028	Y_{AS}_VQM	0.7236	Y_{AS}_VQM	0.8497

- Pearson linear correlation coefficient (PLCC)
- Spearman rank order correlation coefficient (SROCC)
- Root mean square error (RMSE)

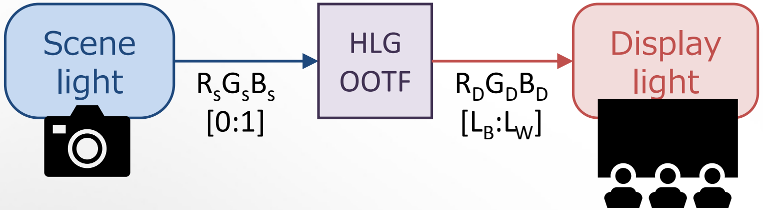
Display Light vs. Scene Light

- Originally, inputs of CSF-based metrics are display light in cd/m^2
- Compared display light Y_D and absolute scene light Y_{AS} inputs

$$Y_{AS} = \alpha Y_S + \beta,\ \alpha = (L_W - L_B),\ \beta = L_B\ where\ L_B = 0.005\ and\ L_W = 1{,}000$$

	PLCC	SROCC	RMSE
Y_D_VQM	0.8066	0.8374	0.7060
Y_{AS}_VQM	0.7028	0.7236	0.8497
Y_D_VDP2	0.9163	0.9146	0.4783
Y_{AS}_VDP2	0.8460	0.8421	0.6368
Y_D_PU_M	0.9175	0.9164	0.4751
Y_{AS}_PU_M	0.8599	0.8356	0.6097
Y_D_PU_S	0.8959	0.9034	0.5307
Y_{AS}_PU_S	0.8613	0.8545	0.6068

Significant difference
($Y_D > Y_{AS}$)


Scene light

$R_sG_sB_s$ [0:1]

HLG OOTF

$R_DG_DB_D$ [L_B:L_W]

Display light

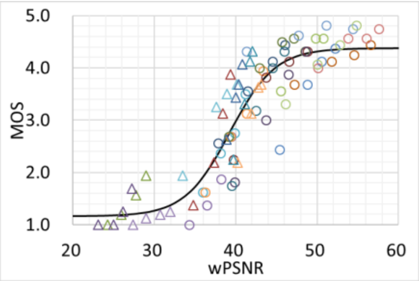
Conversion from scene light to display light

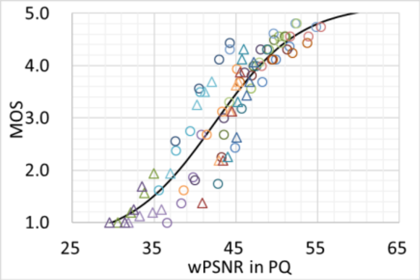
- Scene light inputs are inappropriate for CSF-based metrics

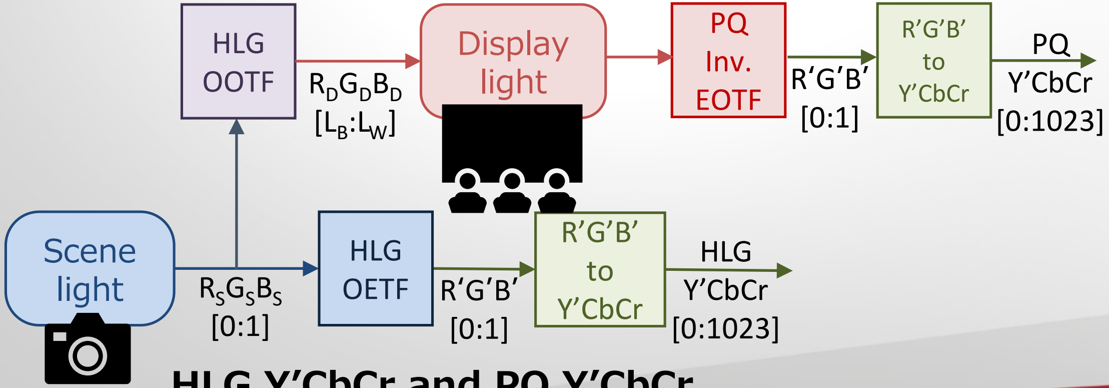
wPSNR in HLG vs. PQ domains

- Compared wPSNR of Y’CbCr in HLG and PQ domains
  - Applying wPSNR after converting to PQ Y’CbCr is mandated for HLG sequences in VVC meeting

	PLCC	SROCC	RMSE
wPSNR	**0.9126**	0.9009	**0.4883**
wPSNR_PQ	0.9084	**0.9110**	0.4995







HLG OOTF

$R_DG_DB_D$ $[L_B:L_W]$

Display light

PQ Inv. EOTF

R’G’B’ [0:1]

R’G’B’ to Y’CbCr

PQ Y’CbCr [0:1023]

Scene light

$R_SG_SB_S$ [0:1]

HLG OETF

R’G’B’ [0:1]

R’G’B’ to Y’CbCr

HLG Y’CbCr [0:1023]

HLG Y’CbCr and PQ Y’CbCr

- No significant difference
- wPSNR of luma Y’ works well in both HLG and PQ domains

Conclusions

- Validated 11 objective metrics for HLG-based image coding
  - Ranking of metrics for HDR coding changes drastically depending on TF used for compression
- Objective metrics should be mindfully selected when comparing image coding methods with different TFs

Future Work

- Continue to study validation with different TFs and objective metrics
  - Explore metrics suit for both HLG- and CSF-based image coding