Perceptual Quality Assessment of UHD-HDR-WCG Videos

Shahrukh Athar[1], Thilan Costa[1], Kai Zeng[2] and Zhou Wang[1]

[1]University of Waterloo, Canada

[2]SSIMWAVE Inc., Canada

September 24, 2019

Outline

Outline

Perceptual Video Quality Assessment (VQA)

Purpose

Development of quantitative measures that can automatically predict the *perceived* quality of videos

Objective VQA

Types of Objective VQA Models:

- Full-Reference (FR) VQA
- Reduced-Reference (RR) VQA
- No-Reference (NR) VQA

Development of objective VQA models requires subject-rated databases

Limitations of Existing Work

Existing subject-rated databases for High Dynamic Range (HDR) videos [Banitalebi-Dehkordi, 2014], [Narwaria, 2015], [Rerabek, 2015], [Minoo, 2015], [Mukherjee, 2016], [Azimi, 2018] have the following limitations:

- Maximum spatial resolution is Full High Definition (FHD)
- Color gamut of content/displays is limited to BT.709
- Maximum temporal resolution is usually 30 frames per second (fps)
- Fixed distortion levels (bit rates) regardless of content complexity are used
- Evaluation of state-of-the-art FR and NR methods is missing

Outline

Waterloo UHD-HDR-WCG Database

Reference Content Characteristics

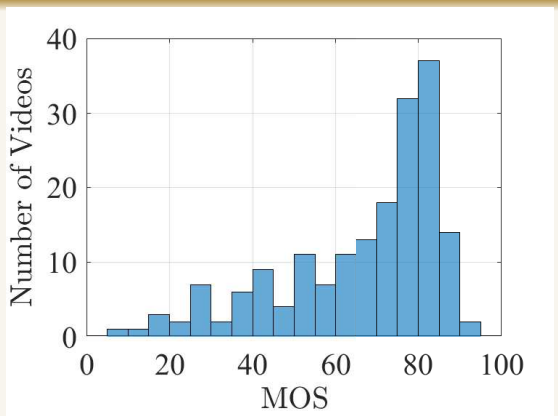
- 14 high-quality reference videos in YUV file format
- Length of each video: 10 seconds
- Ultra High Definition (UHD) resolution (3840 x 2160)
- Bit depth: 10 bits (Luma)
- Wide Color Gamut (WCG): BT.2020 color primaries
- YUV 4:2:0 chroma format
- SMPTE ST 2084 (PQ) transfer function
- Frame rate: 59.94 fps (9 videos) and 24 fps (5 videos)

Waterloo UHD-HDR-WCG Database

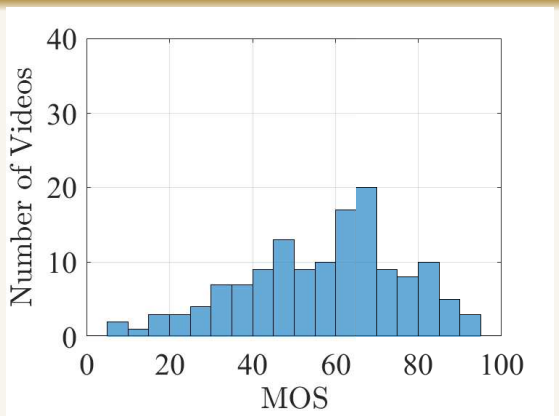
Distorted Content Characteristics

- Focus: To study the impact of compression on UHD-HDR-WCG content
- Two encoders used (H.264 and HEVC)
- Five content-adaptive distortion levels (bitrates) for each encoder
- Overall 140 distorted videos in YUV file format

Impact of Content-Adaptive Distortion Levels



(a) Preliminary FHD database with fixed bitrates



(b) Waterloo UHD-HDR-WCG database with content-adaptive bitrates

Hardware Setup

Canon DP-V2420 Reference Display

- 4K/UHD HDR Mastering monitor
- Screen Size: 24 inch
- Compatible with Academy Color Encoding System (ACES)
- Supports WCG (BT.2020)
- Peak Luminance: 1000 cd/m^2
- Minimum black level: 0.005 cd/m^2
- Supports SMPTE ST 2084 (PQ) transfer function
- Quad 3G Serial Digital Interface (SDI) with throughput of 12 Gbits/s

Hardware Setup

Dedicated Hardware Pipeline

- Maximum throughput requirement: 11.12 Gbits/s
- Workstation
  - Stores the entire database (1.64 TBytes) in a 2 TByte Samsung 960 Pro SSD (read speed up to 3.5 GBytes/s)
  - 32 GBytes 3000 MHz DDR4 RAM (holds each video while playing)
- Blackmagic Design Ultrastudio 4K Extreme 3
  - Connected to the workstation through a Blackmagic Design PCI Express Cable Kit
  - Splits single input data stream into four streams connected to a Quad SDI output interface
  - Output of Ultrastudio connected to the Canon Reference Display
- Customized video playback software developed using Blackmagic Design Software Development Kit (SDK)

Outline

Subjective Study

Salient Features

- 51 subjects aged between 18 and 35
  - 29 males and 22 females
  - 43 naïve and 8 experts
- Single stimulus with hidden reference methodology
- Viewing distance approximately twice the screen height
- Two 30-minutes rating sessions with a mandatory break in-between
- Dark room environment
- Scores range: 0 to 100 (higher for better quality)
- Scoring GUI allowed selection of integers through sliding bar
- Training session preceded the study
  - Five training videos (No overlap with test set)

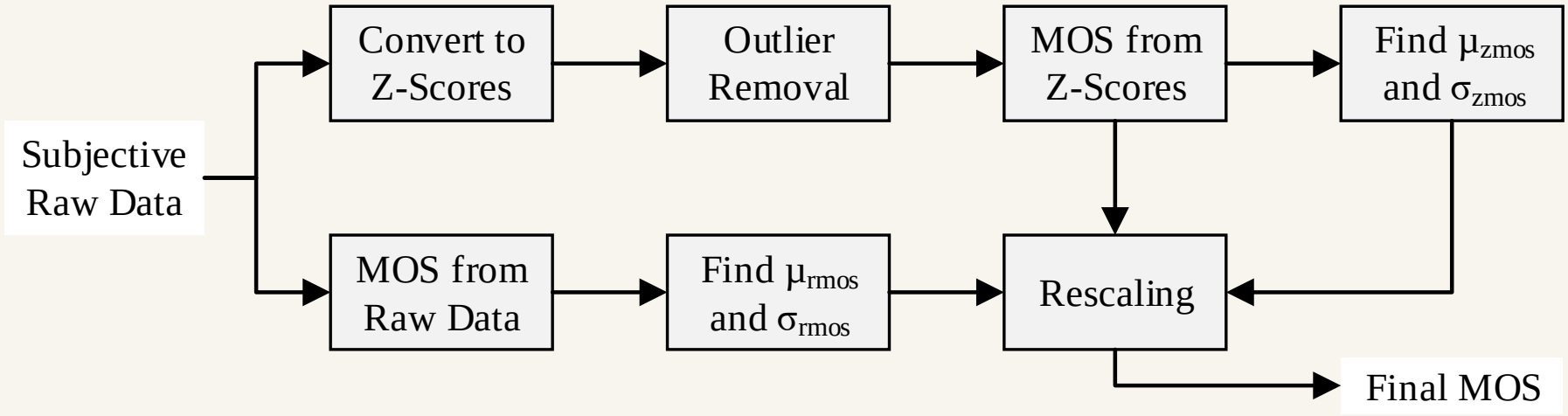
Data Processing

Steps

1. Raw scores converted to Z-scores
   - Accounts for the quality scale variations between subjects
2. Outlier removal procedure according to Rec. ITU-R BT.500-13
   - 9 subjects removed
3. Mean Opinion Score (MOS) for each content computed from Z-scores
4. MOS rescaled to the 0 to 100 range
   - MOS distribution is preserved
   - Maintains overall mean and variance of raw scores

Data Processing

Mean Opinion Score (MOS) Generation Mechanism

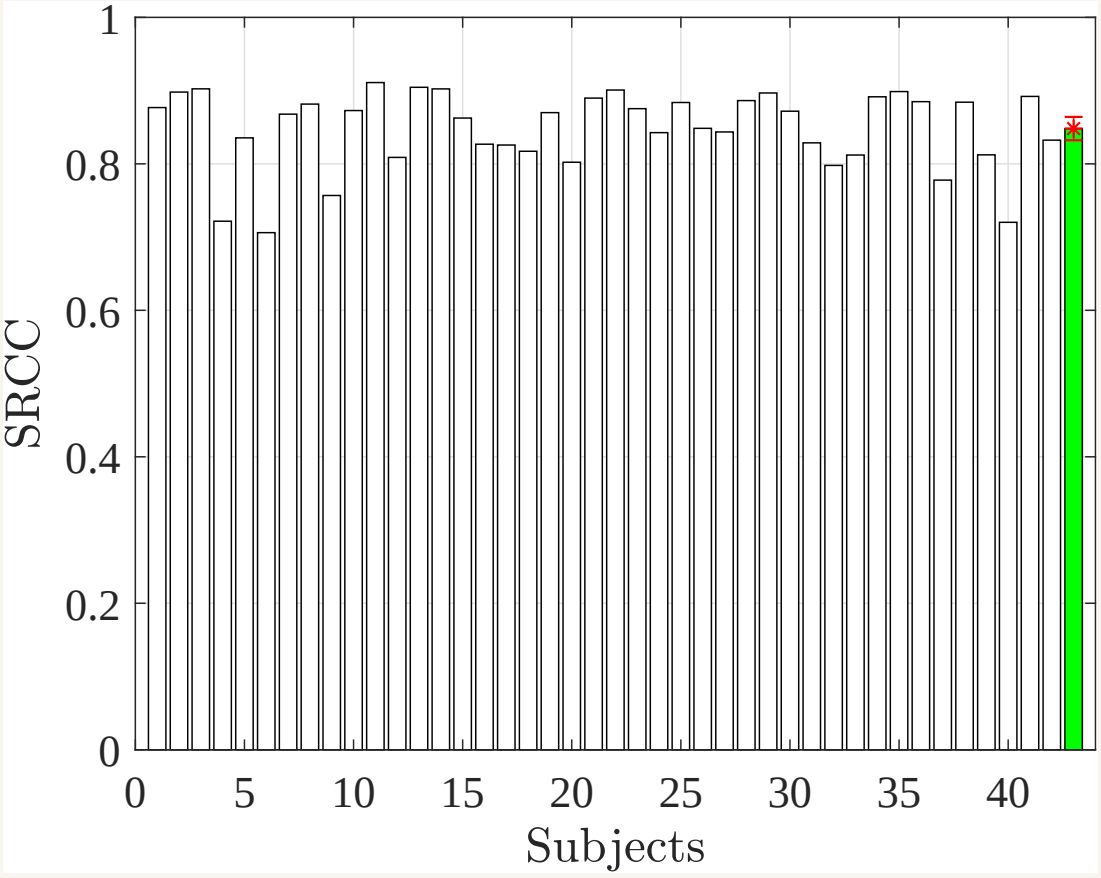


Rescaling

$$MOS = \sigma_{rmos}\left[\frac{MOS_z - \mu_{zmos}}{\sigma_{zmos}}\right] + \mu_{rmos} \quad (1)$$

Data Processing

Spearman Rank Correlation Coefficient between MOS and Individual Subjects



Outline

Performance of Objective Models

Evaluation Criteria

- Prediction Accuracy
  - Pearson Linear Correlation Coefficient (PLCC)
  - Root Mean Square Error (RMSE)
- Prediction Monotonicity
  - Spearman Rank order Correlation Coefficient (SRCC)
- Statistical Significance Testing on prediction residuals
  - Jarque-Bera test to determine Gaussianity of residuals
  - Hypothesis testing through the F-test

Number of Objective Models Evaluated

- 11 FR Models
  - Including HDRVDP2 and HDRVQM (designed for HDR content)
- 7 NR Models

Performance of Objective Models

Category	Method	PLCC	SRCC	RMSE
FR	DSS [Balanov, 2015]	0.7685	0.7456	12.3718
	ESSIM [Zhang, 2013]	0.8512	0.8389	10.1485
	FSIM [Zhang, 2011]	0.8693	0.8564	9.5568
	GMSD [Xue, 2014]	0.7366	0.7045	13.0781
	GSIM [Liu, 2012]	0.8596	0.8453	9.8812
	HDRVDP2 [Mantiuk, 2011]	0.7035	0.6703	13.7423
	HDRVQM [Narwaria, 2015]	0.7783	0.7759	12.1428
	IWSSIM [Wang, 2011]	0.8088	0.7861	11.3730
	PSNR	0.5113	0.4615	16.6185
	SRSIM [Zhang, 2012]	**0.8726**	**0.8630**	**9.4462**
	VIFDWT [Rezazadeh, 2013]	0.6809	0.6748	14.1612
NR	BRISQUE [Mittal, 2012]	0.3622	0.3271	18.0241
	CORNIA [Ye, 2012]	**0.6497**	**0.6296**	**14.7003**
	dipIQ [Ma, 2017]	0.6192	0.5560	15.1845
	HOSA [Xu, 2016]	0.5379	0.5138	16.3015
	LPSI [Wu, 2015]	0.3941	0.3820	17.7718
	NIQE [Mittal, 2013]	0.5286	0.4922	16.4152
	VMEON [Liu, 2018]	0.5776	0.5308	15.7845

Performance Analysis

FR Methods

- SRSIM is the top performing FR method
- Performance of ESSIM, GSIM, and FSIM is statistically indistinguishable from SRSIM
- Above methods are developed for Low Dynamic Range (LDR) content and inherit a similar formulation of signal fidelity measurement from SSIM [Wang, 2004]
- HDR specific FR methods (HDRVDP2 and HDRVQM) do not offer superior performance
- LDR FR methods may be extended for HDR VQA

Performance Analysis

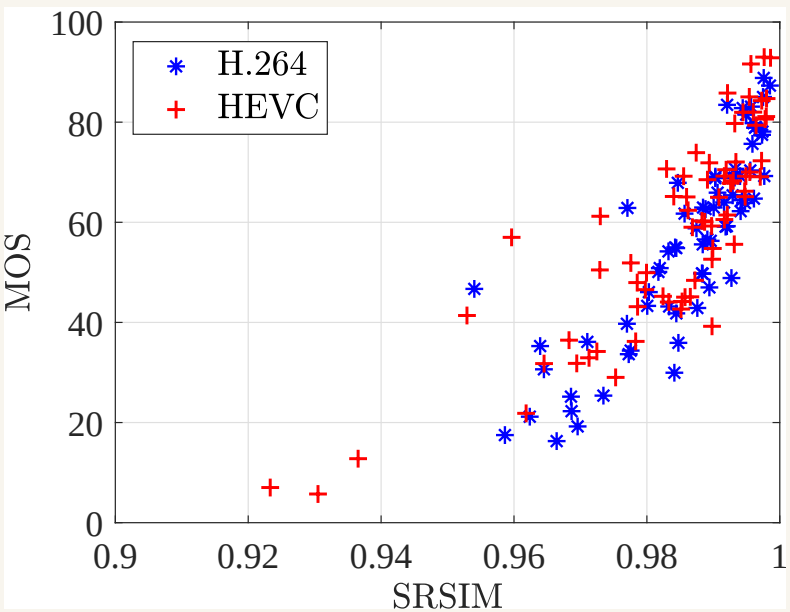
NR Methods

- All NR methods perform inadequately
- CORNIA is the top performing NR method
- All NR methods under test were developed for LDR content
- There is significant room for improvement in HDR specific design innovations

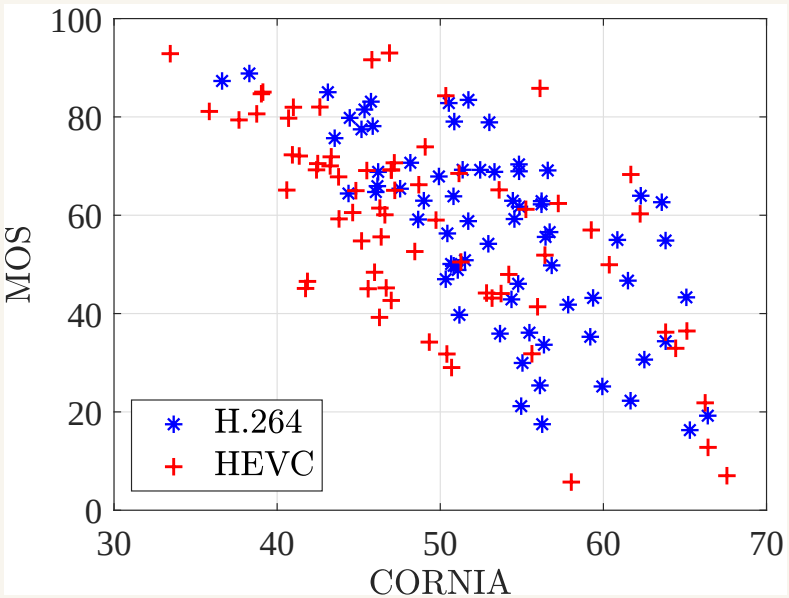
Performance Analysis

Objective Model Performance on Individual Distortion Types

- Models perform similarly on H.264 and HEVC compression
- Example below shows scatter plots for top performing FR (SRSIM) and NR (CORNIA) models



(a)



(b)

Outline

Summary

Contributions

- Constructed a first-of-its-kind Waterloo UHD-HDR-WCG database
- Carried out a first-of-its-kind subjective study on a professional HDR Reference Display with a dedicated hardware pipeline
- Proposed a novel method to process subjective data
  - Accounts for subject quality scale variations
  - Preserves distribution of data and keeps the overall mean and standard deviation of subjective scores unchanged
- Evaluated the performance of 11 FR and 7 NR objective models
  - FR models developed for LDR content can be used as a basis for new UHD-HDR-WCG FR VQA models
  - Substantial room for improvement exists when it comes to NR VQA of UHD-HDR-WCG content

QUESTIONS

References

A. Banitalebi-Dehkordi, M. Azimi, M. T. Pourazad, and P. Nasiopoulos, "Compression of High Dynamic Range Video using the HEVC and H.264/AVC Standards," In *Int. Conf. Het. Netw. Quality, Rel., Security, Robustness*, 2014.

M. Narwaria, M. P. Da Silva, and P. Le Callet, "Study of High Dynamic Range Video Quality Assessment," In *Proc. SPIE Opt. Eng. Appl.*, 2015.

M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and Objective Evaluation of HDR Video Compression," In *Int. Workshop Video Process., Quality Metrics Consum. Electron. (VPQM)*, 2015.

K. Minoo, Z. Gu, D. Baylon, and A. Luthra, "On metrics for objective and subjective evaluation of high dynamic range video," In *Proc. SPIE Opt. Eng. Appl.*, 2015.

R. Mukherjee, K. Debattista, T. Bashford-Rogers, P. Vangorp, R. Mantiuk, M. Bessa, B. Waterfield, and A. Chalmers, "Objective and subjective evaluation of High Dynamic Range video compression," In *Signal Process.: Image Commun.*, 2016.

M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content," In *arXiv preprint arXiv:1803.04815*, 2018.

A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015.

X. Zhang, X. Feng, W. Wang, and W. Xue, "Edge Strength Similarity for Image Quality Assessment," In *IEEE Signal Process. Lett.*, 2013.

L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," In *IEEE Trans. Image Process.*, 2011.

References


W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," In *IEEE Trans. Image Process.*, 2014.


A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," In *IEEE Trans. Image Process.*, 2012.


R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," In *ACM Trans. Graphics*, 2011.


M. Narwaria, M. P. Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," In *Signal Process.: Image Commun.*, 2015.


Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," In *IEEE Trans. Image Process.*, 2011.


L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2012.


S. Rezazadeh and S. Coulombe, "A novel discrete wavelet transform framework for full reference image quality assessment," In *Signal, Image, Video Process. (SIViP)*, 2013.


A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," In *IEEE Trans. Image Process.*, 2012.


P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012.


K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs," In *IEEE Trans. Image Process.*, 2017.

References

J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind Image Quality Assessment Based on High Order Statistics Aggregation," In *IEEE Trans. Image Process.*, 2016.

Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015.

A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," In *IEEE Signal Process. Lett.*, 2013.

W. Liu, Z. Duanmu, and Z. Wang, "End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks," In *ACM Int. Conf. Multimedia*, 2018.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: from error visibility to structural similarity," In *IEEE Trans. Image Process.*, 2004.