An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming

Nabajeet Barman*, Steven Schmidt[¥], Saman Zadtootaghaj[‡], Maria G. Martini*, Sebastian Möller[¥]

*Wireless Multimedia & Networking Research Group, Kingston University, London, U.K.

[¥]Quality and Usability Lab, TU Berlin, Germany

[‡]Telekom Innovation Labs, Deutsche Telekom AG, Berlin, Germany

Motivation

- Gaming Videos: Increasing in popularity
- Increasing number of Gaming OTT Providers: Twitch.tv, YouTube Gaming, Hitbox.tv
- Twitch.tv alone consists of approximately 2 million streamers, 15 million daily active users
- Twitch.tv Ranked 4th in terms of total traffic during peak hours: After Netflix, Google and Apple

Motivation ctd…

- Gaming videos consist of synthetic and artificial content
- Streaming Requirements: Real-time, CBR, 1-pass

Motivation ctd…

- Many metrics are based on properties inherent to natural images and videos: SSIM, NIQE, BRISQUE etc.
- Applicability and performance analysis of VQA metrics on gaming videos → Open question !!!!
- Gaming videos: No reference available
  - Performance analysis and design of good No Reference metrics critical for QoE estimation

GamingVideoSET

24 videos, 2 each from 12 different games

(a) Counter Strike: Global Offensive

(b) Diablo III

(c) Dota 2

(d) FIFA 2017

(e) H1Z1: Just Kill

(f) Hearthstone

(g) Heroes of the Storm

(h) League of Legends

(i) Project Cars

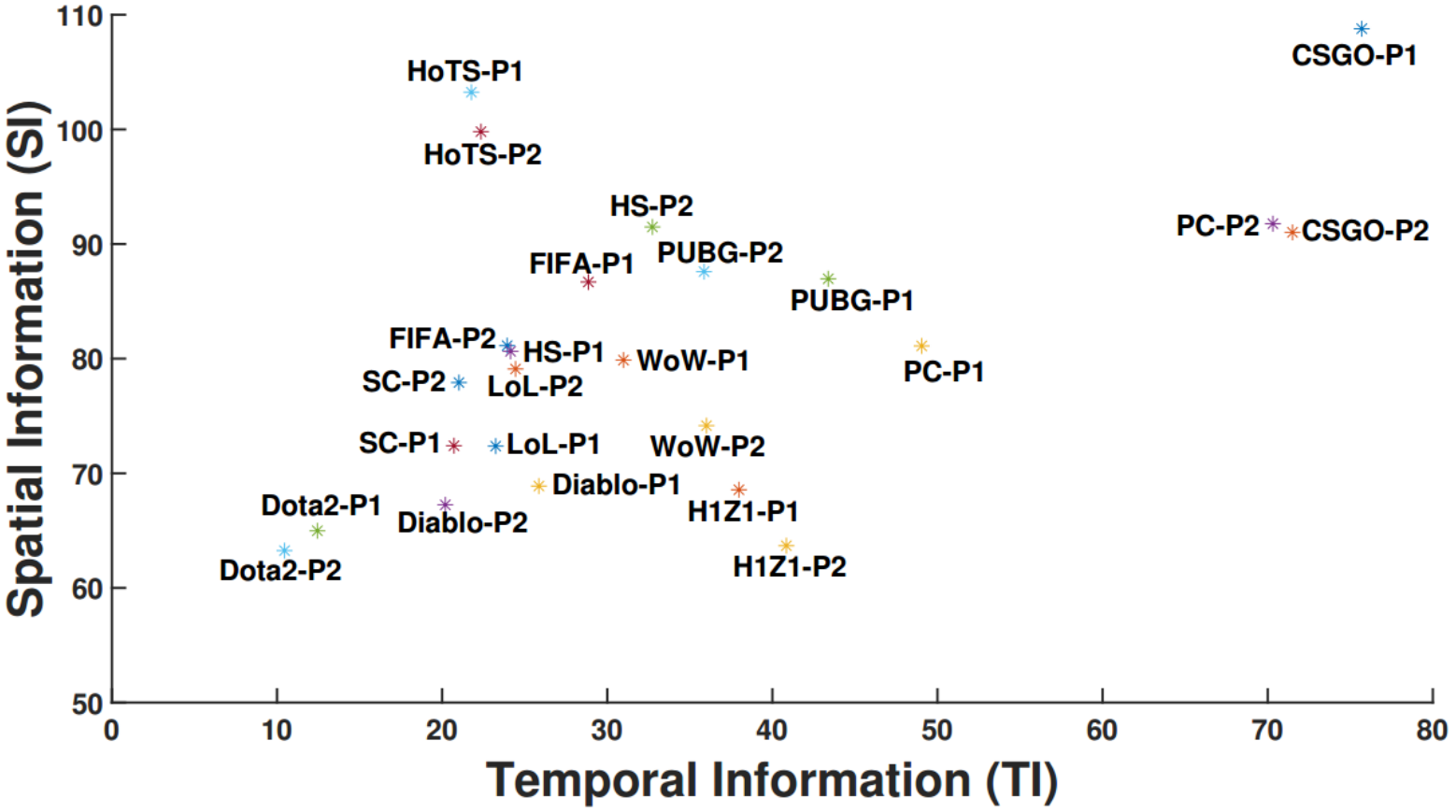
(j) PlayerUnknown's Battleground

(k) StarCraft II

(l) World of Warcraft

N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, “GamingVideoSET: A Dataset for Gaming Video Streaming Applications,” in 2018 16th Annual Workshop on Network and Systems Support for Games (NetGames 2018), Amsterdam, 2018.

Spatial Information (SI) vs. Temporal Information (TI)



Video Encoding Parameters

Parameter	Value
Duration	30 sec
Resolution	1080p, 720p, 480p
Frame Rate	30
Number of Reference Videos	24
Encoder	FFmpeg
Encoding Mode	CBR
Video Compression Standard	H.264, Main 4.0
Preset	Veryfast

Resolution-Bitrate Pairs

Resolution	Bitrate (kbps)
1080p	**600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**
720p	**500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000**
480p	**300**, 400, **600**, 900, **1200**, **2000**, **4000**

The bitrates in bold text refer to the bitrates used in the subjective quality assessment

Evaluated VQA Metrics

Full Reference

- Peak Signal to Noise Ratio (PSNR)
- Structural Similarity Index Metric (SSIM)
- Video Multi-Method Assessment Fusion (VMAF)

Reduced Reference

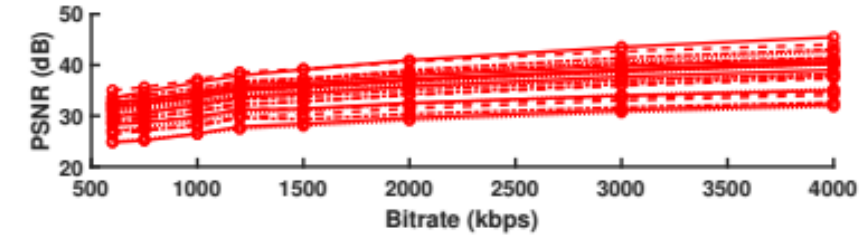
- Optimized version of ST-RRED (Spatio-temporal-reduced reference entropic difference) metric: ST-RREDOpt
- Spatial efficient entropic differencing for quality assessment (SpEED-QA)

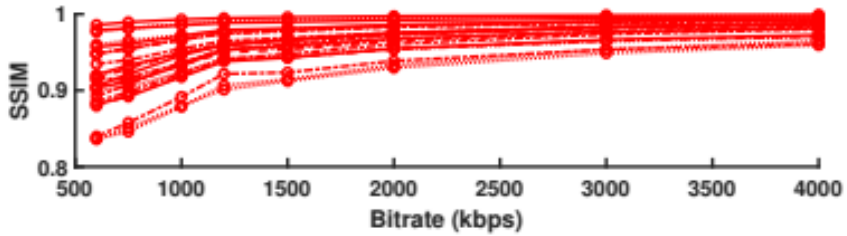
No Reference

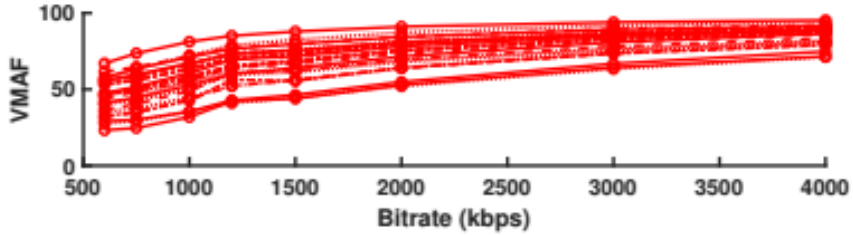
- Blind image quality index (BIQI)
- Blind/referenceless image spatial quality evaluator (BRISQUE)
- Natural Image Quality Evaluator (NIQE)

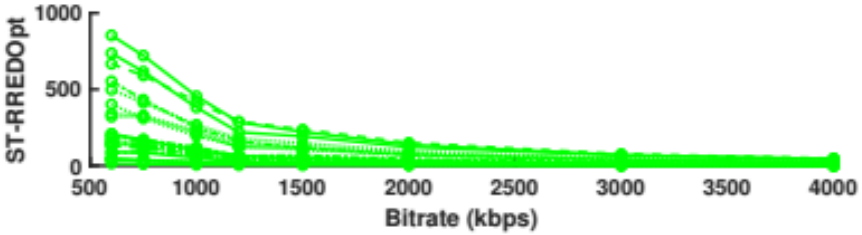
Quality vs. Bitrate Curves

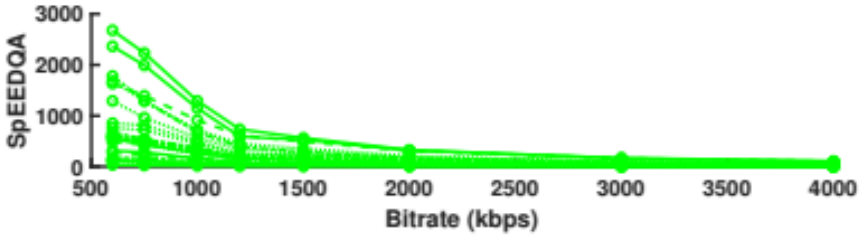
Eight VQA metrics considering all videos for 1080p resolution

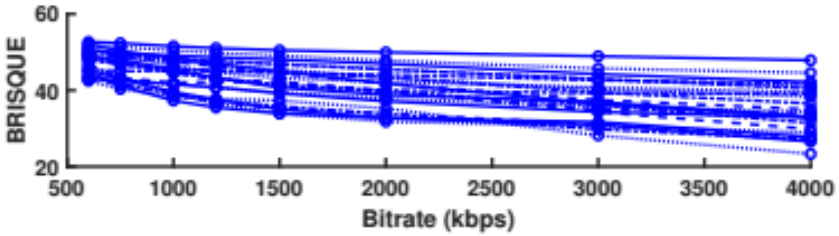


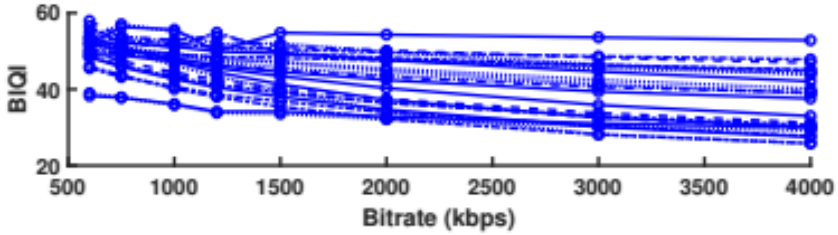


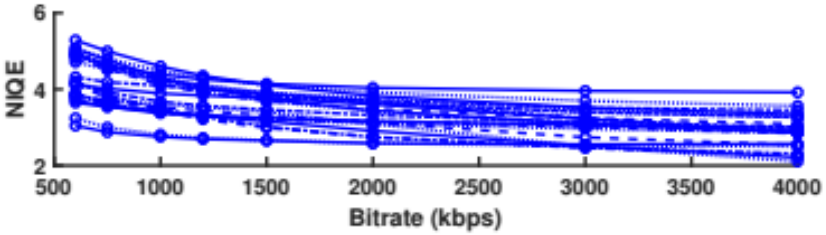












Subjective Test

- Six videos: CSGO, H1Z1, HS, FIFA, LoL and PC
- 3 resolutions, 5 bitrates each: 90 stimuli
- Subjective test methodology: ACR (scale: 1-5)
- Test Environment: ITU-R Rec. BT.500
- Number of test participants: 25
- Display Monitor: 22” FHD ViewSonic
- 720p and 480p videos decoded and rescaled to 1080p

Correlation Analysis

Performance of the VQA metrics scores with respect to MOS

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.57	0.43	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	**0.74**	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	ST-RREDOpt	-0.61	-0.51	-0.82	-0.85	-0.79	-0.92	-0.71	-0.74
	SpEEDQA	-0.63	-0.52	-0.83	-0.87	-0.77	-0.93	-0.71	-0.75
NR Metrics	BRISQUE	-0.57	-0.48	-0.83	-0.89	-0.88	-0.91	-0.49	-0.51
	BIQI	-0.53	-0.51	-0.73	-0.72	-0.81	-0.80	-0.43	-0.46
	NIQE	-0.73	-0.74	-0.85	-0.81	-0.89	-0.90	-0.77	-0.76

- “All Data” refers to the combined data of all three resolution-bitrate pairs
- FR and RR evaluated on rescaled videos. NR Metrics evaluation done on non-rescaled videos

Discussion 1: Comparison of VQA Metrics with respect to MOS

Discussion 1

Comparison of VQA Metrics with respect to MOS

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.57	0.43	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	**0.74**	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	ST-RREDOpt	-0.61	-0.51	-0.82	-0.85	-0.79	-0.92	-0.71	-0.74
	SpEEDQA	-0.63	-0.52	-0.83	-0.87	-0.77	-0.93	-0.71	-0.75
NR Metrics	BRISQUE	-0.57	-0.48	-0.83	-0.89	-0.88	-0.91	-0.49	-0.51
	BIQI	-0.53	-0.51	-0.73	-0.72	-0.81	-0.80	-0.43	-0.46
	NIQE	-0.73	-0.74	-0.85	-0.81	-0.89	-0.90	-0.77	-0.76

- VMAF results in the highest correlation values
- Both RR metrics results in almost similar results →
  - SpEED-QA seven times faster than ST-RREDOpt
- NR metrics
  - BIQI performs the worst
  - BRISQUE and NIQE
    - 1080p and 720p: Almost equal performance
    - 480p and All data: NIQE performs better than BRISQUE

Discussion 2:
Impact of resolution on VQA metrics

Discussion 2

Impact of resolution on VQA metrics

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.57	0.43	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	**0.74**	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	ST-RREDOpt	-0.61	-0.51	-0.82	-0.85	-0.79	-0.92	-0.71	-0.74
	SpEEDQA	-0.63	-0.52	-0.83	-0.87	-0.77	-0.93	-0.71	-0.75
NR Metrics	BRISQUE	-0.57	-0.48	-0.83	-0.89	-0.88	-0.91	-0.49	-0.51
	BIQI	-0.53	-0.51	-0.73	-0.72	-0.81	-0.80	-0.43	-0.46
	NIQE	-0.73	-0.74	-0.85	-0.81	-0.89	-0.90	-0.77	-0.76

- For FR and NR metrics, performance decreases from higher to lower resolution.
- For RR metrics, 720p results in higher correlation value
- Fisher's Z-test to asses the significance of the difference between two resolution correlation values
  - Difference between 1080p and 720p is **not** statistically significant
  - Difference between 720p and 480p **is** statistically significant

Discussion 3: Performance degradation at lower resolutions

Discussion 3

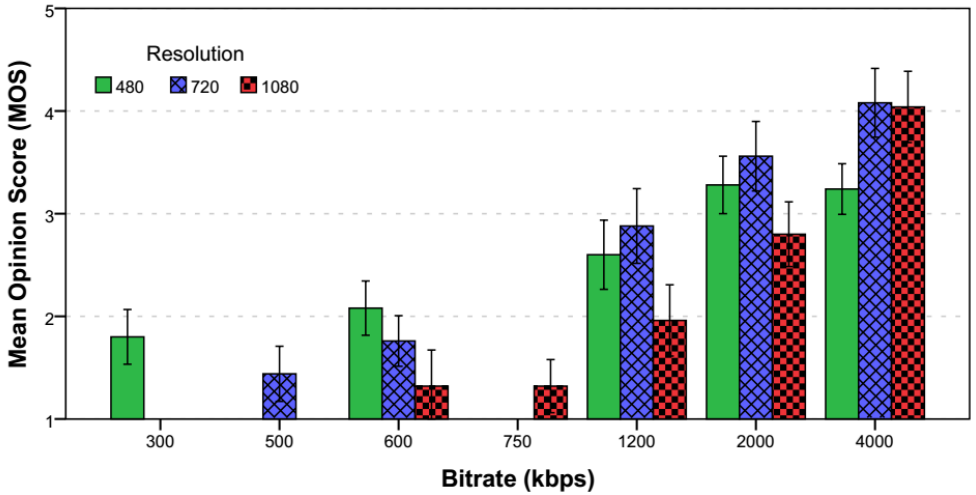
Performance degradation at lower resolutions

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.57	0.43	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	**0.74**	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	ST-RREDOpt	-0.61	-0.51	-0.82	-0.85	-0.79	-0.92	-0.71	-0.74
	SpEEDQA	-0.63	-0.52	-0.83	-0.87	-0.77	-0.93	-0.71	-0.75
NR Metrics	BRISQUE	-0.57	-0.48	-0.83	-0.89	-0.88	-0.91	-0.49	-0.51
	BIQI	-0.53	-0.51	-0.73	-0.72	-0.81	-0.80	-0.43	-0.46
	NIQE	-0.73	-0.74	-0.85	-0.81	-0.89	-0.90	-0.77	-0.76

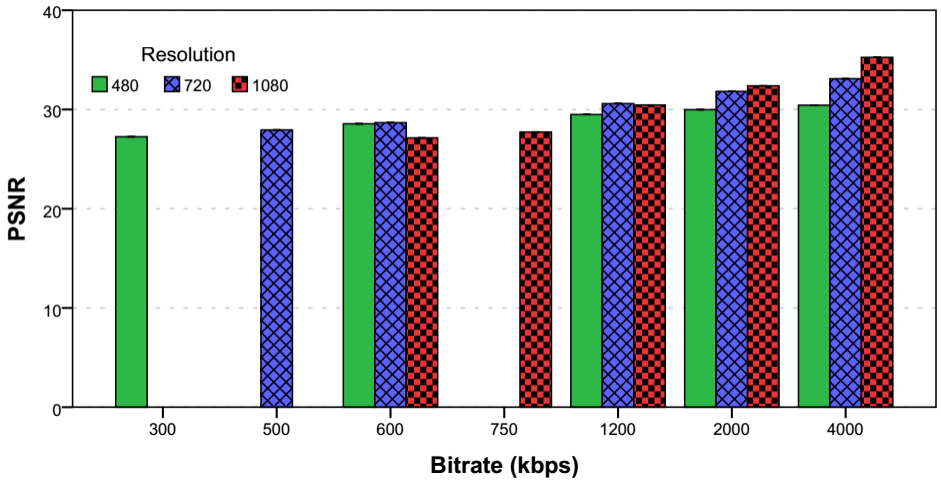
- Performance at 480p is considerably lower compared to the same VQA metric performance for the 720p and 1080p resolutions.
- Decrease in performance for some metrics is higher than others

MOS (with 95% confidence interval), PSNR and VMAF values for the CSGO video sequence

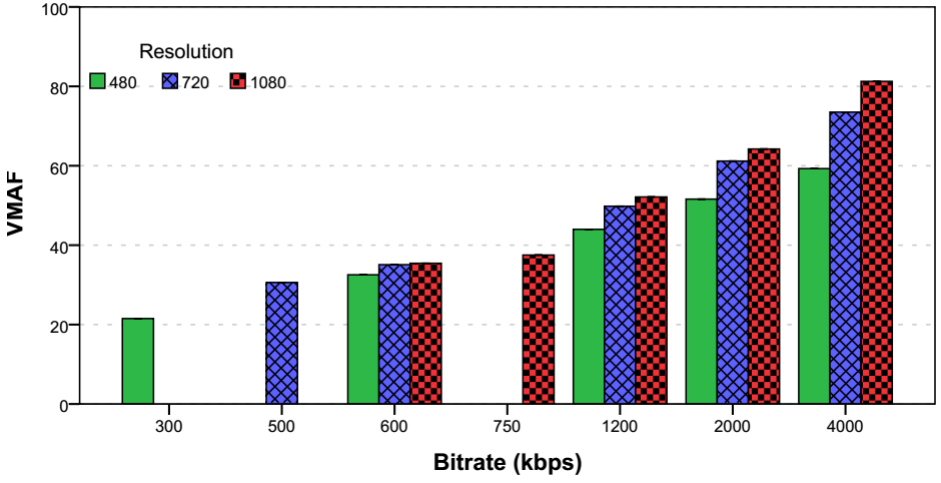
(a) MOS vs. Bitrate (kbps)



(b) PSNR (dB) vs. Bitrate (kbps)



(c) VMAF vs. Bitrate (kbps)



Evaluation over the full dataset

- Use VMAF scores as ground truth
- Evaluate performance of rest seven metrics on the full dataset
  - 24 videos, 24 resolution-bitrate pairs → 576 stimuli

Correlation Analysis

Performance of the VQA metrics scores with respect to VMAF

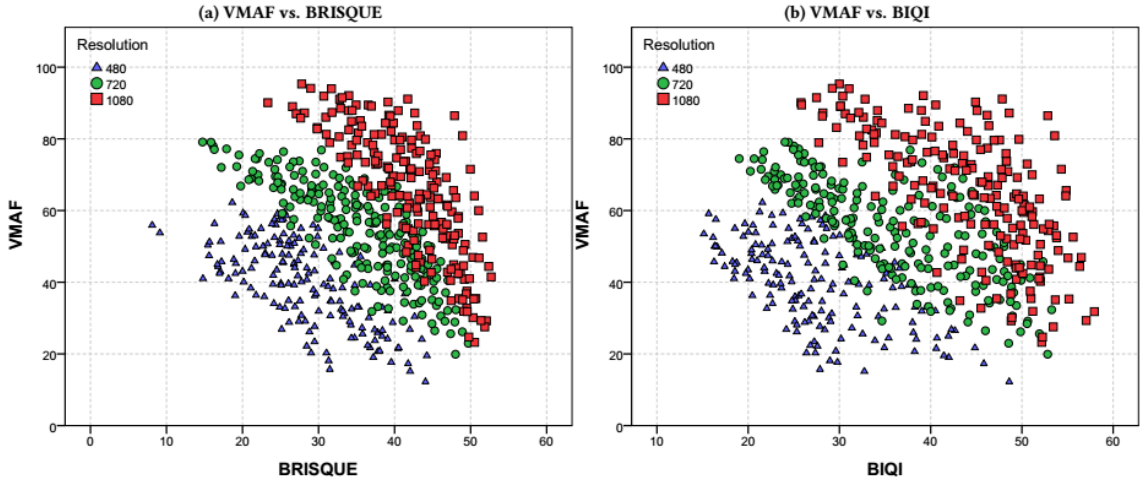
Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.62	0.60	0.79	0.77	**0.91**	0.92	**0.87**	**0.87**
	SSIM	0.56	0.56	0.68	0.70	0.80	0.83	0.70	0.74
RR Metrics	ST-RREDOpt	-0.66	-0.85	-0.74	-0.89	-0.77	-0.91	-0.53	-0.61
	SpEEDQA	-0.68	**-0.88**	-0.76	**-0.92**	-0.77	**-0.93**	-0.55	-0.63
NR Metrics	BRISQUE	-0.68	-0.68	-0.79	-0.79	-0.77	-0.78	-0.14	-0.14
	BIQI	-0.57	-0.54	-0.70	-0.71	-0.67	-0.68	-0.05	-0.05
	NIQE	**-0.75**	-0.77	**-0.81**	-0.81	-0.78	-0.76	-0.42	-0.42

- "All Data" refers to the combined data of all three resolution-bitrate pairs.
- In terms of PLCC:
  - 1080p: PSNR; 720p and 480p: NIQE
- In terms of SROCC:
  - 1080p, 720p and 480p: SpEEDQA
- All Data, PLCC and SROCC: PSNR

Discussion 4: Decreased NR metric performance for "All Data"

NR metrics vs. VMAF scores

Scatter plot of the considering all three resolutions over the whole dataset



→ May not be suitable for spatial resolution based adaptation applications (e.g. Typical HTTP adaptive streaming solutions)

Conclusions

- VMAF results in the highest correlation w.r.t. MOS score in terms of PLCC and SROCC values.
- SSIM and NIQE also performs quite well.
- Performance of all VQA metrics is worse for 480p resolution as compared to 720p and 1080p.
- When considering VMAF values as the benchmark, PSNR results in the highest correlation considering “All data”
- When considering “All data”, the performance of NR metrics decreases significantly.
- VQA metric performance on gaming videos similar to non-gaming videos
- Future work: Investigate shortcomings in existing NR metrics and improve/develop new metrics

Thank you!

QoENet MSCA-ITN Network

European Commission

This presentation is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 643072.