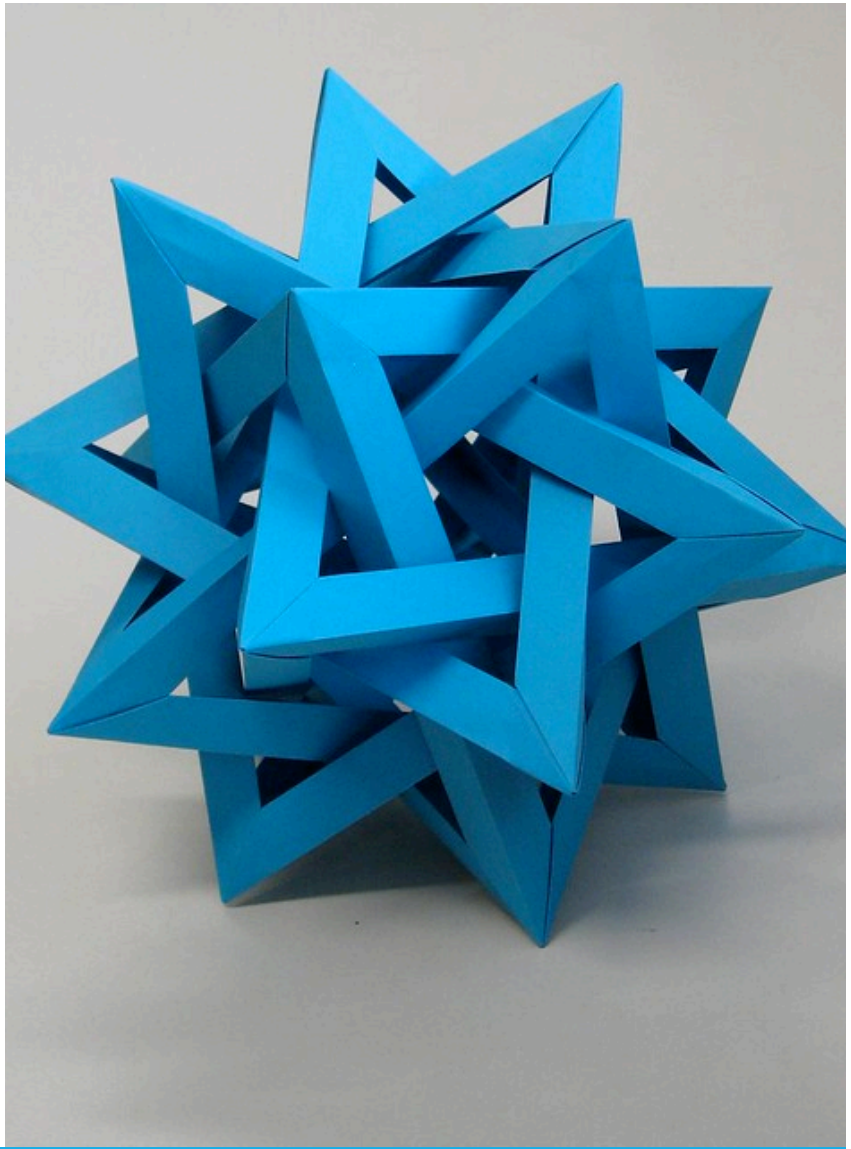
POLITECNICO DI TORINO
Dipartimento di Automatica e Informatica

Comparing commercial and open-source VQMs for HD constant bitrate videos

Hodor Project Phase 1 Summary
VQEG December 2020

A COLLABORATION BETWEEN SKY GROUP (AND AFFILIATES) AND FOUR UNIVERSITIES



Executive Summary

- In August 2019, Sky initiated a collaborative research project to understand how different picture quality metrics perform based on asset characteristics and encoding technologies.
- The collaboration work consisted of Sky Group (and affiliates) and four universities (all of whom are part of the Video Quality Experts Group (VQEG))
- The scope of work reported here focuses on the performance of eight picture quality metrics for HD constant bitrate (CBR) videos.
- The results comprise both proprietary and open source metrics
- VMAF seems to show a performance significantly better than that of the proprietary metrics
- Proprietary metrics show performance that overcomes that of other open source metrics
- The VQMS' disagreement allows to characterise the accuracy of the VQM
- The way a PVS is encoded might determine the difficulty of accurately assessing its visual quality objectively

Hodor project: overview

Goal

- To characterise the performance of commercial/proprietary video quality metrics (PVQMs)
- Objectively determine processed *video sequences (PVSs)* whose visual quality is difficult to assess objectively.

Scope

Codecs	AVC	HEVC
Resolution	Adaptive streaming 1080	4K
Rate control	CBR	VBR
	Proprietary optimisations	Per-title/ Constant quality
Objective metric	Commercial Metric 1	Commercial Metric 2
	VMAF	SSIM/MS-SSIM
	VIF / PSNR	SSIM / MS-SSIM

Dataset construction

- Apple's HLS specification using the Elemental transcoder
- For sports, the frame rates for sports content were interpolated to 50 or 59.94 frames per second (fps) depending on the region
- For movies, the frame rate was same as source
- Video bitrates ranged from 365kbps to 7800kbps
- Scenes were carefully selected according to recognised guidelines, and scene duration were 10 seconds

Dataset construction

- A total of 368 PVSs were generated
  - 47 sources: Sports/Movies
  - 8 Hypothetical Reference Circuits

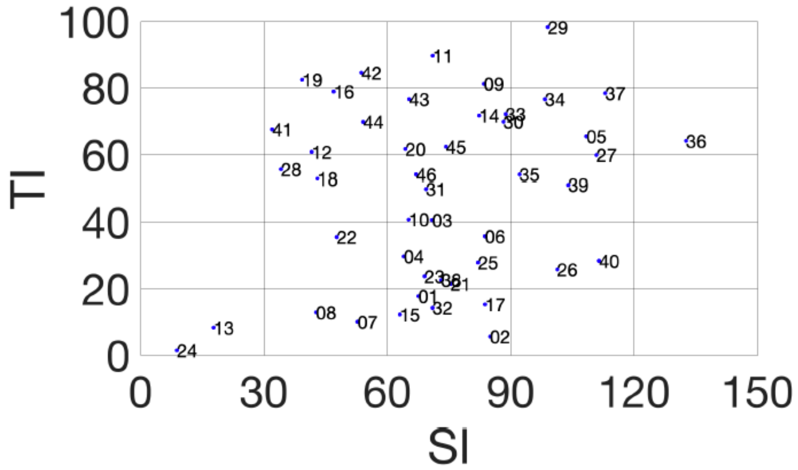


Fig. 1. Assessing the heterogeneity of the 46 SRCs used to generate the PVSs contained in the dataset in terms of the spatial and temporal activity index. The labels indicate the different SRCs

- 6 Open-source VQMs and 2 PVQMs
  - Open-source: PSNR, SSIM, MS-SSIM, VIF, XPSNR, VMAF
  - Proprietary: PVQM1 and PVQM2

Findings

- VMAF seems to show a performance significantly better than that of the PROPRIETARY METRICS
- PROPRIETARY METRICS show performance that overcomes that of other OPEN SOURCE METRICS
- The VQMS’ DISAGREEMENT allows to characterize the accuracy of the VQM
- The way a PVS is encoded might determine the difficulty of accurately assessing its visual quality objectively

Disagreement measure

- All scores from the different VQMs were normalised to the VMAF scale (i.e. 0 to 100)
- If a pair of metrics differs in their predicted scores by more than 7 points, then we say this pair of metrics disagree
- If the metrics differ by less than 7 points, we say the metrics agree. This is because a change of more than 7 VMAF points would be perceptible and noticeable to a viewer
- Pairwise comparisons of eight metrics will result in 28 comparisons on a single transcoded video
- This approach allowed us to identify transcoded videos where metrics disagreed the most

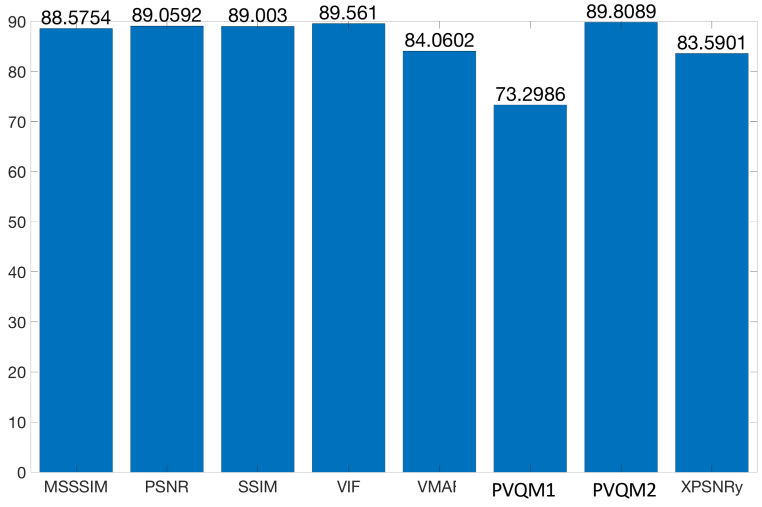
$$D_{pvs} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}\left(\left| f_i\left(vqm_i^{pvs}\right) - f_j\left(vqm_j^{pvs}\right)\right| > \delta_1\right)}{\binom{n}{2}}$$

Dataset construction

For legal reasons Images were removed

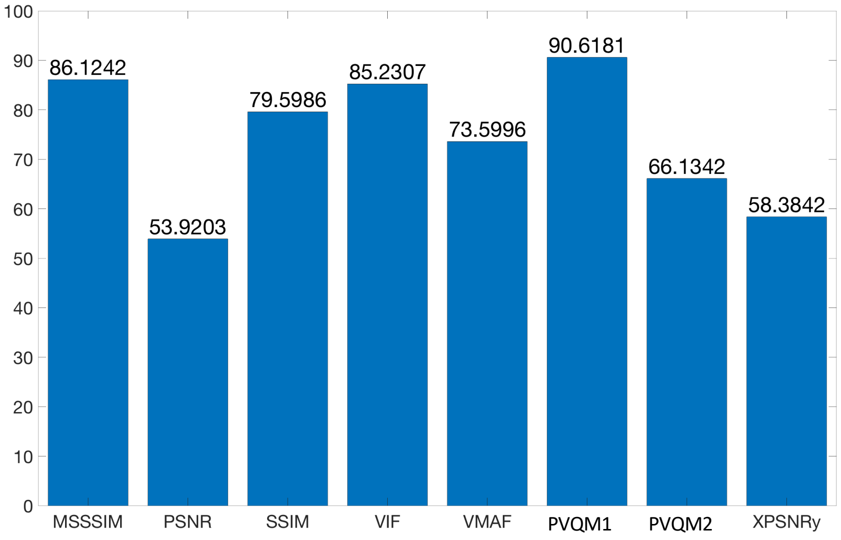
Disagreement measure: Example

PVS 1



LOW DISAGREEMENT

PVS 2



HIGH DISAGREEMENT

Small scale subjective test

- 83 Processed Video Sequences (PVSs)
- 30 PVSs with low VQM disagreement
- 53 PVSs with high VQM disagreement
- 16 subjects
- Method: DSIS
- Experiment conducted in 2 labs (Italy, Germany)

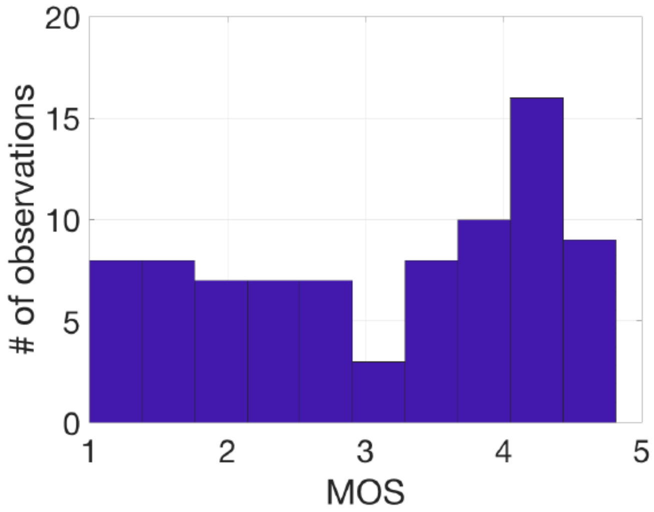
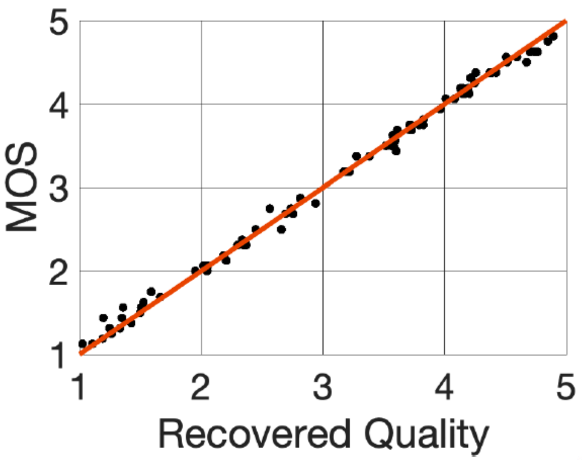
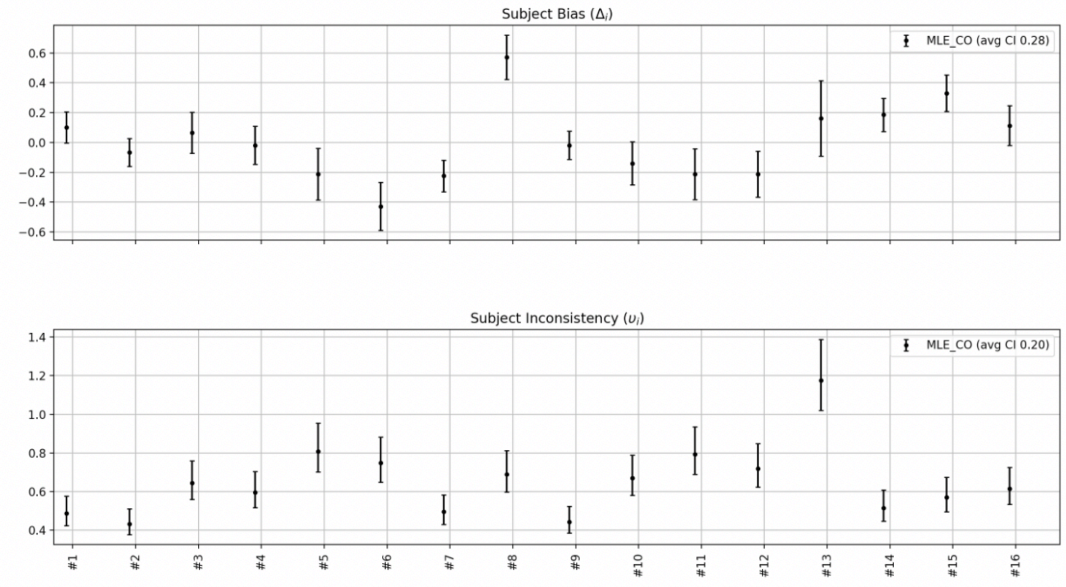


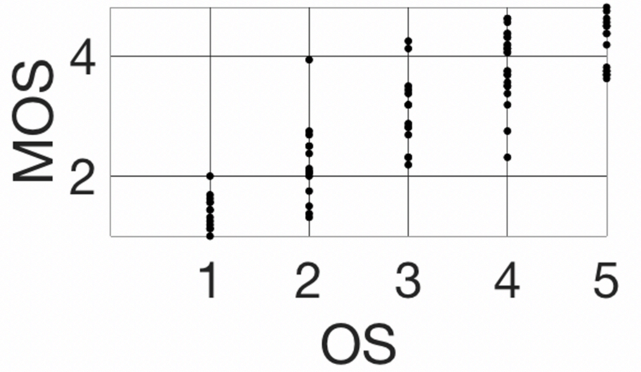
Fig. 3. The histogram of the MOS values shows a distribution that is not far from a uniform one. This is fundamental, since a different distribution of subjective scores could significantly bias the analysis' conclusions.

Applying SUREAL

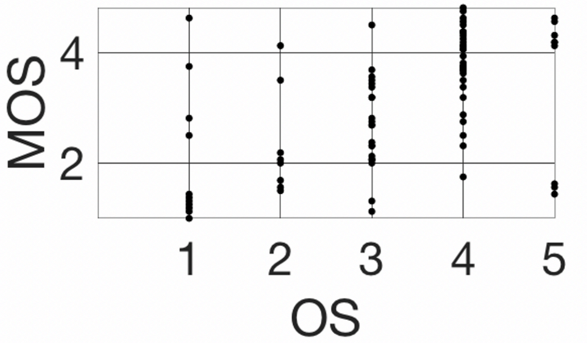




- BT.500 suggested to reject subject #3



Subject #3



Subject #13

Results: VQMs Comparison

TABLE II
COMPARING ALL VQMS IN TERMS OF ACCURACY

Metric	PLCC	SROCC	RMSE
PSNR	0.43	0.61	1.05
SSIM	0.49	0.57	1.02
MSSIM	0.65	0.72	0.88
VIF	0.69	0.68	0.85
XPSNR	0.80	0.81	0.70
PVQM1	0.79	0.76	0.72
PVQM2	0.84	0.84	0.63
VMAF	0.91	0.91	0.50

1
PLCC.: ↑

1
SROCC: ↑

RMSE: ↓
0

VMAF shows higher accuracy

Results: VQMs Comparison

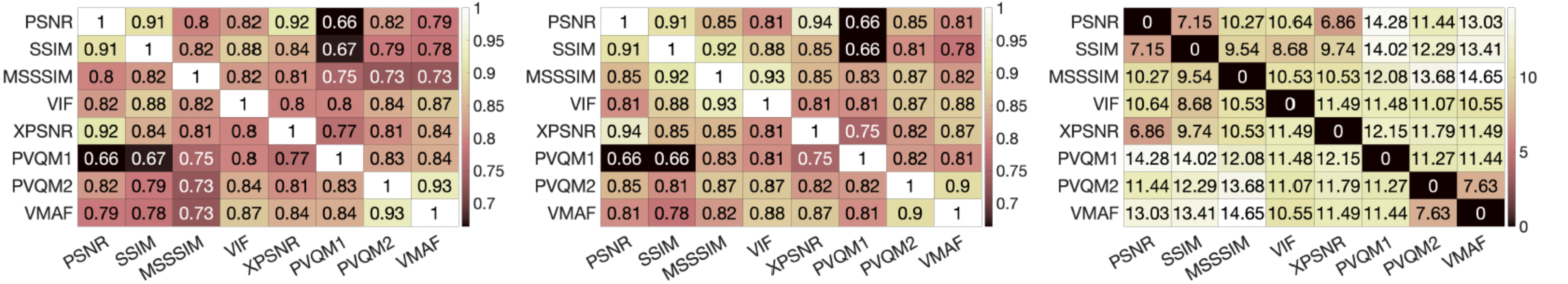
- Is the metric on the row better than the one on the column with statistical significance in term of PLCC ?

	PSNR	SSIM	MSSSIM	VIF	XPSNR	PVQM1	PVQM2	VMAF
PSNR	-	0	0	0	0	0	0	0
SSIM	0	-	0	0	0	0	0	0
MSSIM	1	0	-	0	0	0	0	0
VIF	1	1	0	-	0	0	0	0
XPSNR	1	1	1	0	-	0	0	0
PVQM1	1	1	1	0	0	-	0	0
PVQM2	1	1	1	1	0	0	-	0
VMAF	1	1	1	1	1	1	1	-

Proprietary metrics are not significantly different one from each other but their performance significantly overcomes the one of the PSNR, SSIM and MSSSIM

Results: VQMs Comparison

- All VQMs showed significant alignment on the whole dataset



(a) PLCC

	PSNR	SSIM	MSSSIM	VIF	XPSNR	PVQM1	PVQM2	VMAF
PSNR	1	0.91	0.8	0.82	0.92	0.66	0.82	0.79
SSIM	0.91	1	0.82	0.88	0.84	0.67	0.79	0.78
MSSSIM	0.8	0.82	1	0.82	0.81	0.75	0.73	0.73
VIF	0.82	0.88	0.82	1	0.8	0.8	0.84	0.87
XPSNR	0.92	0.84	0.81	0.8	1	0.77	0.81	0.84
PVQM1	0.66	0.67	0.75	0.8	0.77	1	0.83	0.84
PVQM2	0.82	0.79	0.73	0.84	0.81	0.83	1	0.93
VMAF	0.79	0.78	0.73	0.87	0.84	0.84	0.93	1

(b) SROCC

	PSNR	SSIM	MSSSIM	VIF	XPSNR	PVQM1	PVQM2	VMAF
PSNR	1	0.91	0.85	0.81	0.94	0.66	0.85	0.81
SSIM	0.91	1	0.92	0.88	0.85	0.66	0.81	0.78
MSSSIM	0.85	0.92	1	0.93	0.85	0.83	0.87	0.82
VIF	0.81	0.88	0.93	1	0.81	0.81	0.87	0.88
XPSNR	0.94	0.85	0.85	0.81	1	0.75	0.82	0.87
PVQM1	0.66	0.66	0.83	0.81	0.75	1	0.82	0.81
PVQM2	0.85	0.81	0.87	0.87	0.82	0.82	1	0.9
VMAF	0.81	0.78	0.82	0.88	0.87	0.81	0.9	1

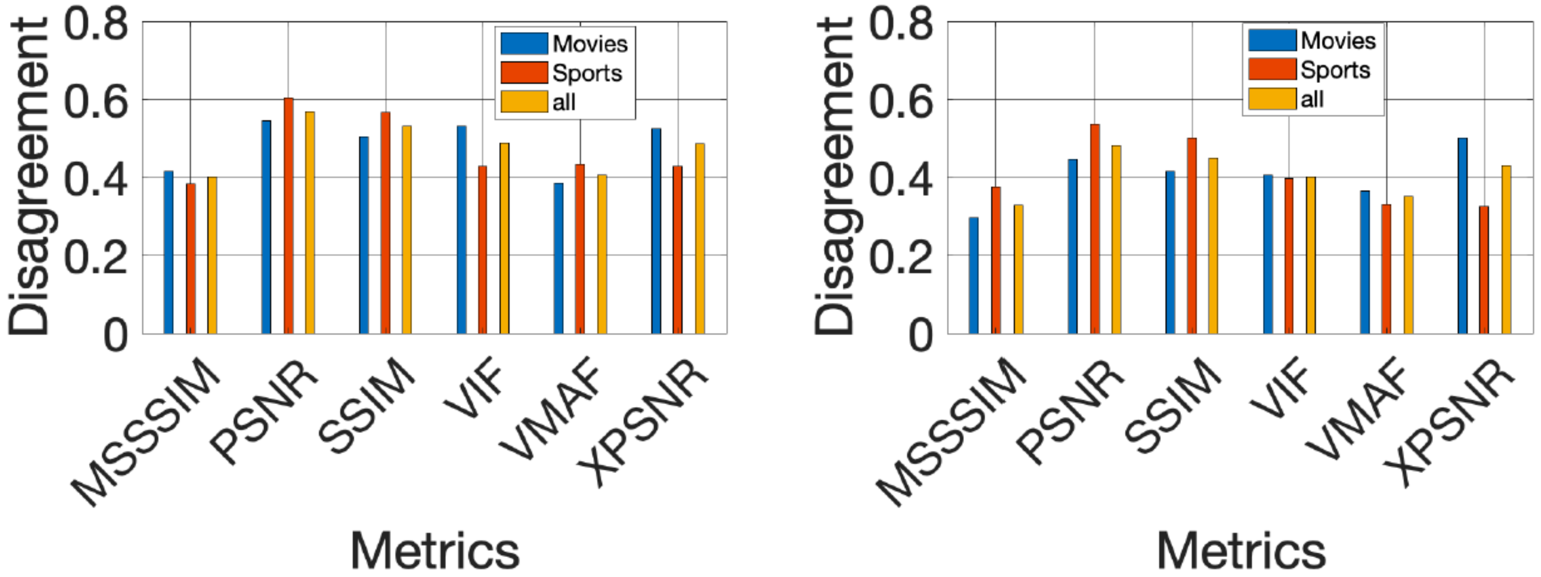
(c) MUTUAL RMSE

	PSNR	SSIM	MSSSIM	VIF	XPSNR	PVQM1	PVQM2	VMAF
PSNR	0	7.15	10.27	10.64	6.86	14.28	11.44	13.03
SSIM	7.15	0	9.54	8.68	9.74	14.02	12.29	13.41
MSSSIM	10.27	9.54	0	10.53	10.53	12.08	13.68	14.65
VIF	10.64	8.68	10.53	0	11.49	11.48	11.07	10.55
XPSNR	6.86	9.74	10.53	11.49	0	12.15	11.79	11.49
PVQM1	14.28	14.02	12.08	11.48	12.15	0	11.27	11.44
PVQM2	11.44	12.29	13.68	11.07	11.79	11.27	0	7.63
VMAF	13.03	13.41	14.65	10.55	11.49	11.44	7.63	0

Fig. 7. Evaluating the correlation and mutual RMSE between all the metrics considered in the study. In general the commercial metrics showed higher correlation to state of the art open-source metrics, as expected.

Results: VQMs Comparison

- Fraction of PVSs on which each PVQM disagrees with the open-source VQMs as function of the content type



(a) PVQM1 vs Open-source VQMs

(b) PVQM2 vs Open-source VQMs

Fig. 8. Evaluating the fraction of PVSs on which the PVQMs disagree with each open-source VQM. The analysis indicates that the PSNR and SSIM are a bit more likely to measure a quality that would be perceptually different than that indicated by the PVQMs especially on Sports content.

Results: VQMs disagreement vs accuracy

- When VQMs disagree, each VQM is likely to show lower accuracy
- In case of agreement VQMs performance approximate that of a subjective test

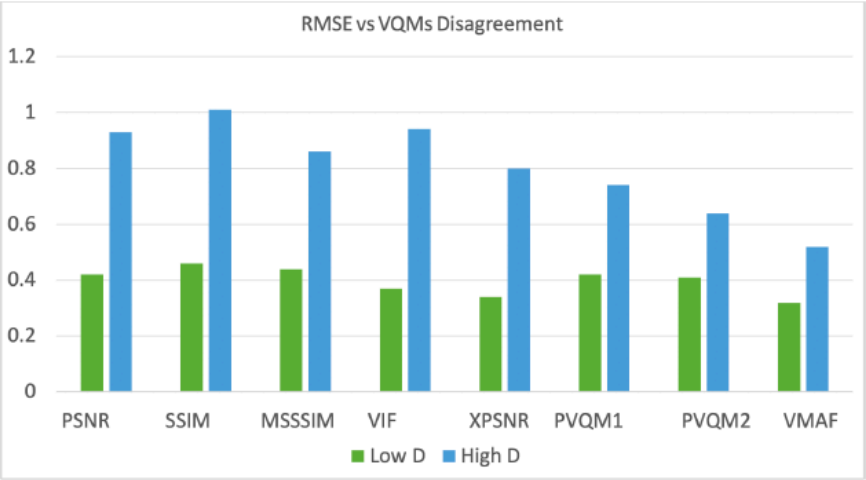
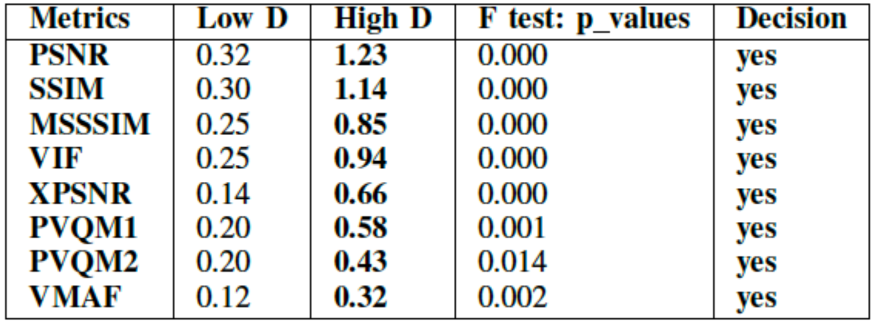


Fig. 9. The video quality metrics' accuracy, in terms of RMSE, for low and high disagreement conditions. Lower RMSE is better. For all the metrics, in case of high disagreement, the score is expected to be affected by larger error.

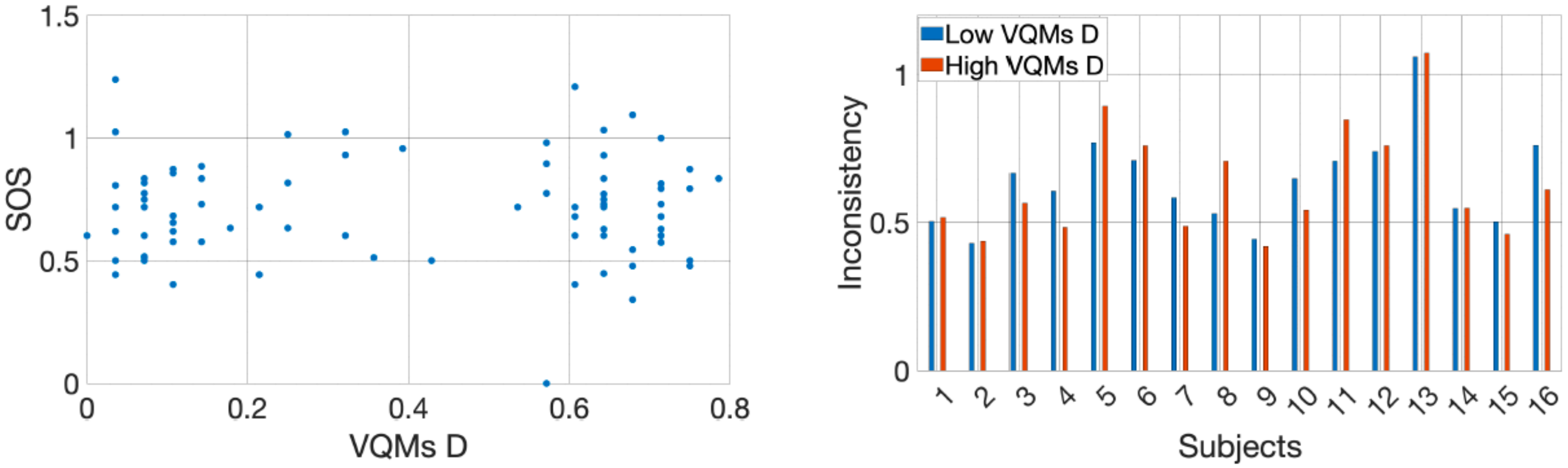
TABLE IV
STATISTICAL ANALYSIS OF THE VARIANCE OF THE MOS PREDICTION ERROR. IN CASE OF HIGH VQMS DISAGREEMENT, EACH METRIC IS EXPECTED TO BE MORE INCONSISTENT WITH STATISTICAL SIGNIFICANCE.



Metrics	**Low D**	**High D**	**F test: p_values**	**Decision**
PSNR	0.32	**1.23**	0.000	**yes**
SSIM	0.30	**1.14**	0.000	**yes**
MSSSIM	0.25	**0.85**	0.000	**yes**
VIF	0.25	**0.94**	0.000	**yes**
XPSNR	0.14	**0.66**	0.000	**yes**
PVQM1	0.20	**0.58**	0.001	**yes**
PVQM2	0.20	**0.43**	0.014	**yes**
VMAF	0.12	**0.32**	0.002	**yes**

Results: VQMs disagreement vs accuracy

- VQMs disagreement makes the objective evaluation of visual quality difficult
- Is it the same in the contest of a subjective test ?
- Are PVSs with high VQMs Disagreement those suitable for a subjective test?



(a)

(b)

Fig. 10. SOS and individual subjects inconsistency as function of the proposed VQMs disagreement. Subjects seem to experience the same difficulty in assessing the quality of a PVS independently on the disagreement of the VQMs scores.

Results: VQMs disagreement & bitstream features

- The VQMs disagreement can, to a certain extent, be predicted from bitstream features
- This establishes a link between the VQMs accuracy and the way the PVS was encoded
- Important bitstream features:
  - bitrate, motion vector, quantization parameter,
  - the percentage of Intra blocks in a slice and the percentage of 2Nx2N Intra coded blocks.

Results: VQMs disagreement & bitstream features

- LM= Linear Model
- RT = Regression Tree
- NN= Neural Network (1 hidden layer, 4 neurons)
- SVR (Gaus) = Support Vector Regression with gaussian kernel
- SVR (rbf) = Support Vector Regression with rbf kernel

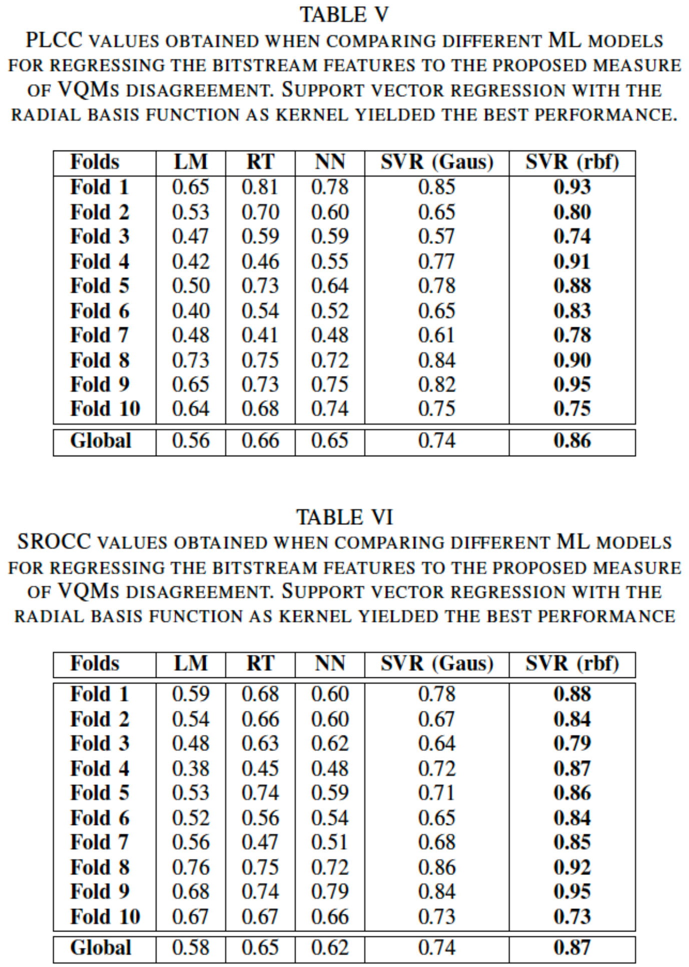


TABLE V

PLCC VALUES OBTAINED WHEN COMPARING DIFFERENT ML MODELS FOR REGRESSING THE BITSTREAM FEATURES TO THE PROPOSED MEASURE OF VQMS DISAGREEMENT. SUPPORT VECTOR REGRESSION WITH THE RADIAL BASIS FUNCTION AS KERNEL YIELDED THE BEST PERFORMANCE.

Folds	LM	RT	NN	SVR (Gaus)	SVR (rbf)
Fold 1	0.65	0.81	0.78	0.85	**0.93**
Fold 2	0.53	0.70	0.60	0.65	**0.80**
Fold 3	0.47	0.59	0.59	0.57	**0.74**
Fold 4	0.42	0.46	0.55	0.77	**0.91**
Fold 5	0.50	0.73	0.64	0.78	**0.88**
Fold 6	0.40	0.54	0.52	0.65	**0.83**
Fold 7	0.48	0.41	0.48	0.61	**0.78**
Fold 8	0.73	0.75	0.72	0.84	**0.90**
Fold 9	0.65	0.73	0.75	0.82	**0.95**
Fold 10	0.64	0.68	0.74	0.75	**0.75**
Global	0.56	0.66	0.65	0.74	**0.86**

TABLE VI

SROCC VALUES OBTAINED WHEN COMPARING DIFFERENT ML MODELS FOR REGRESSING THE BITSTREAM FEATURES TO THE PROPOSED MEASURE OF VQMS DISAGREEMENT. SUPPORT VECTOR REGRESSION WITH THE RADIAL BASIS FUNCTION AS KERNEL YIELDED THE BEST PERFORMANCE

Folds	LM	RT	NN	SVR (Gaus)	SVR (rbf)
Fold 1	0.59	0.68	0.60	0.78	**0.88**
Fold 2	0.54	0.66	0.60	0.67	**0.84**
Fold 3	0.48	0.63	0.62	0.64	**0.79**
Fold 4	0.38	0.45	0.48	0.72	**0.87**
Fold 5	0.53	0.74	0.59	0.71	**0.86**
Fold 6	0.52	0.56	0.54	0.65	**0.84**
Fold 7	0.56	0.47	0.51	0.68	**0.85**
Fold 8	0.76	0.75	0.72	0.86	**0.92**
Fold 9	0.68	0.74	0.79	0.84	**0.95**
Fold 10	0.67	0.67	0.66	0.73	**0.73**
Global	0.58	0.65	0.62	0.74	**0.87**

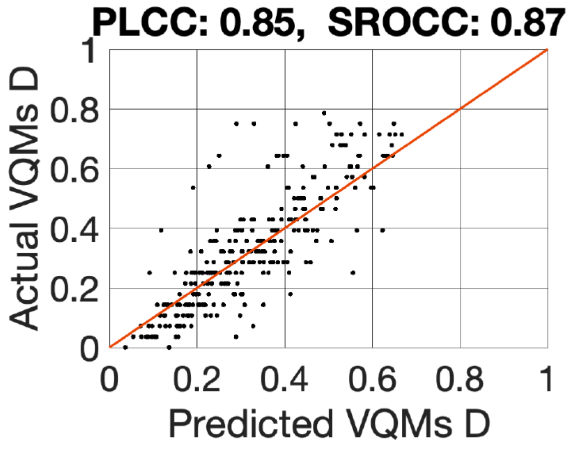


Fig. 11. Accuracy of the final SVR model on all the data. Despite some outliers, in general the model has been able to satisfactory model the metrics disagreement, yielding high linear (0.85) and rank correlation values (0.87).

Thank you for your attention