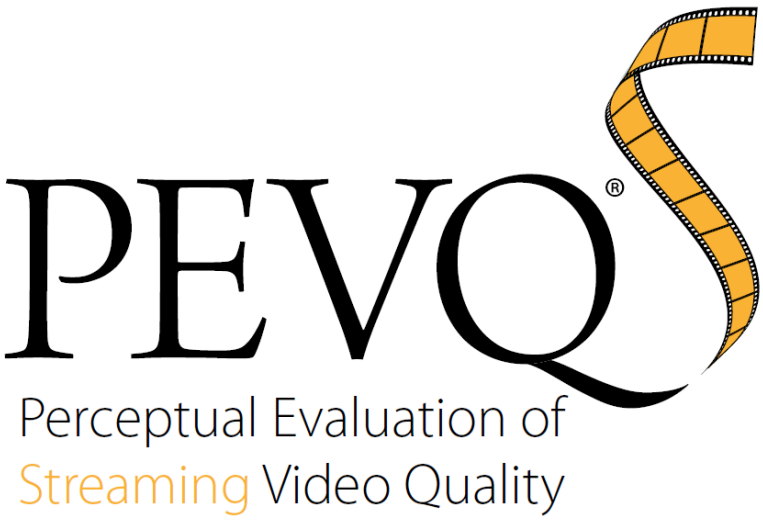


OPTICOM Model Performance Results

AVHD/P.NATS Phase 2 Project

04 June 2021

Author:	Shahid Mahmood Satti OPTICOM GmbH Germany	Tel: +49 9131 53020-0 Fax: +49 9131 53020-20 Email: ss@opticom.de
Author:	Christian Schmidmer OPTICOM GmbH Germany	Tel: +49 9131 53020-0 Fax: +49 9131 53020-20 Email: cs@opticom.de
Author:	Michael Keyhl OPTICOM GmbH Germany	Tel: +49 9131 53020-0 Fax: +49 9131 53020-20 Email: mk@opticom.de

Contents

1. Abstract

This report provides detailed model performance results of the objective video quality models submitted by OPTICOM GmbH in the AVHD-AS/P.NATS phase 2 (AVHD-PNATS2) project. In total six models: two parametric bitstream models, two hybrid no-reference models, one full-reference model and one hybrid full-reference model were submitted by OPTICOM in the AVHD-PNATS2 competition. Parametric models were validated using 26 short sequence databases (video length: 6-9 seconds). Pixel-based and hybrid models were validated for 26 short sequence databases as well as 6 long sequence databases (video length: 1-5 minutes).

2. Introduction

AVHD-PNATS2 was a joint collaboration between ITU-T Study Group 12 and VQEG. The project validated models under ten different model categories, which were defined to cover a broad range of use-cases. The use-cases included video quality monitoring of fully encrypted bitstreams, cases where deep packet inspection is possible to access the video bitstream parameters and unencrypted bitstreams which can be decoded to access the pixel information. Alternatively, the bitstream is either available at the encoding premises or measurement is carried out using pixel information available e.g., from the client side. The models thus have a wide range of applications, from encoding optimization over client-side quality of experience (QoE) assessment up to benchmarking purposes and network/service optimization.

At the ITU-T SG12/Q14 September 2019 interim meeting in Stockholm, all the models submitted to the ten model categories of the competition were validated. In total, winning models were found for five out of ten model categories. These categories were: bitstream mode 0, bitstream mode 1, bitstream mode 3, hybrid no-reference mode 0 and pixel-based reduced-reference models. Winning models are the models which perform statistically significantly better compared to any other model of the same category and winning models of any category of lower complexity. If multiple models were performing statistically equivalent, these models formed a "winning group". Consequently, three models were part of the bitstream mode 0 winning group and two were part of bitstream mode 1 winning group. For bitstream mode 3, hybrid no-reference mode 0 and pixel-based reduced-reference models the winning groups contained exactly one model each [3]. Details of the winning model categories and the range of RMSE values of the winning models were reported to VQEG at the October 2019 VQEG meeting held in Shenzhen, China [3].

This document presents the model performance of all six models submitted by OPTICOM to the AVHD-PNATS2 competition. At the ITU-T SG12 meeting held from 26th Nov. to 5th Dec. 2019 in Geneva, three out of the five winning model categories previously determined were standardized. These were: bitstream mode 3, hybrid no-reference mode 0 and a pixel-based reduced-reference model. Due to the fact that a merging of the winning candidates of bitstream mode 0 and bitstream mode 1 could not be achieved no standards are defined for these cases. Detailed model performance of three standardized models can be found in [1] and [2].

Three out of the six models submitted by OPTICOM were part of the winning groups of their respective model categories. Those were: bitstream mode 0, bitstream mode 1 and the hybrid no-reference model. The Hybrid no-reference model was standardized as ITU-T P.1204.5 [4]. The performance of all six submitted models is described in this document.

3. Model Input

In this section the input/output of different model categories, for which OPTICOM submitted models, are explained.

3.1 Bitstream Mode 0 (BSM0)

This category represents the bitstream models which take into account the following input parameters from a short video sequence or video segment of length 6-9 seconds (no audio) – from here onwards referred to as video segment.

1. Video encoder (H.264, H.265, VP9)
2. Encoded resolution (240p up to 2160p)
3. Encoded bitrate (90 Kbps – 20 Mbps)
4. Encoded framerate (15 to 60 fps)
5. Video encoder profile (optional)
6. Video display device (Tablet/Mobile, PC-Monitor/TV)
7. Video display resolution (Tablet/Mobile: 1440p, PC-Monitor/TV: 2160p)

3.2 Bitstream Mode 1 (BSM1)

This category represents the bitstream models which take into account the following input parameters of a video segment.

1. The complete BSM0 input – see Sec 3.1
2. Encoded video frame sizes in bytes
3. Video frame type (I or Non-I frames)

3.3 Hybrid No-reference Mode 0 (HYN0)

This category represents the bitstream models which take into account the following input parameters of a video segment.

1. The complete BSM0 input – see Sec 3.1
2. Degraded video – video segment decoded and upscaled to the display resolution.

3.4 Hybrid No-reference Mode 1 (HYN1)

This category represents the bitstream models which take into account the following input parameters of a video segment.

1. The complete BSM1 input – see Sec 3.2
2. Degraded video – video segment decoded and upscaled to the display resolution.

3.5 Pixel-based Full-Reference (PXFR)

This category represents the pixel-based models which take into account the following input parameters of a short video (no audio).

1. Original short reference video at native resolution
2. Degraded video –video segment decoded and upscaled to the display resolution.

3.6 Hybrid Full-Reference Mode 0 (HYF0)

This category represents the hybrid models which take into account the following input parameters of a video segment.

1. The complete BSM0 input – see Sec 3.1
2. Full-reference input – see Sec 3.5

3.7 Long-term Quality Prediction

For pixel-based and hybrid model categories (Sec. 3.3–3.6), additionally the following input is available to predict the long-term audio-visual quality. A long video sequence contains a number of audio and video segments. The length of a long sequence is typically between 1 and 5 minutes.

1. Position and length of each stalling event
2. Length of initial-loading
3. Number of segments of varying resolution, bitrate or framerate appended together, which represents the adaptivity in an adaptive streaming scenario. See Sec. 3.3–3.6 for the parameters which define a segment for each model category.
4. Per-segment audio codec and audio bitrate information.

4. Model Output

4.1 All Models (Sec. 3.1 – Sec. 3.6)

All models (bitstream, pixel-based and hybrid) were trained and validated to predict the video quality of a video segment of length 6-9 seconds. The model output is a single quality prediction score on a 5-point ACR quality scale.

4.2 Pixel-based and Hybrid Models (Sec. 3.3 – Sec. 3.6)

Pixel-based and hybrid models were also trained and validated for long video sequences. The model output is a single audio-visual quality prediction score produced for each long (1-5 minutes) test condition on a 5-point ACR quality scale.

5. Model Description (Short Overview)

This section gives a brief summary of model description of the six models submitted by OPTICOM. We start with OPTICOM's HYN0 model. As being already an ITU-T standard (P.1204.5 [4]), the detailed model description is publicly available..

5.1 HYN0

The hybrid no-reference model presented in this section has a parametric logistic a-like function [1] which for a given video encoder maps an average bitrate-based feature x to an intermediate quality prediction S,

$$S = a * \left(\frac{1 - \exp(-d * (x - c))}{1 + \exp(-b * (x - c))} \right)$$

where x is computed for each video segment of fixed resolution and framerate. Note that the above function without the numerator is exactly the logistic function, where the values of a, b and c determine the saturation point, decay rate and shift of the quality curve with respect to x. The additional numerator term is introduced to add a faster decay of the curve towards lower values of x, where the constant d determines the decay factor of this additional decay term.

Assuming that the encoded bitrate of the video segment is defined in kilobits per second, then x is defined as:

$$x = log_{10}(bitrate * \exp(h_0 * (r - 1)))$$

where r has a different value for each chroma subsampling format and. $h_0 > 0$ is a video codec-specific constant. Factors a, b and c of the first equation are further functions of the three quantities, namely, the encoded framerate, encoded resolution and the content complexity.

Impact of the Encoded Framerate: a, c are increasing functions of the encoded framerate, while b is a decreasing function of the encoded framerate (fps), defined as:

$$a' = a_0 - a_f * \left(\frac{60}{fps}\right),\ b' = b_0 + b_f * \left(\frac{60}{fps}\right),\ c' = c_0 - c_f * \left(\frac{60}{fps}\right),$$

where $a_f > 0$, $b_f > 0$ and $c_f > 0$ are codec-specific constants. a_0, b_0 and c_0 are codec-specifc initial positive values.

Impact of the Encoded Resolution: a, b and c are increasing functions of the encoded resolution, defined as:

$$a'' = a' - a_s * log_{10}(u_a * (s - 1)), \qquad b'' = b' - b_s * log_{10}(u_a * (s - 1)),$$

$$c'' = c' - c_s * log_{10}(u_c * (s - 1))$$

where s is the resolution scale factor, defined as the ratio between display resolution and encoded resolution. All constants (a_s, b_s, c_s, u_a, u_b, u_c) in the above equations are codec-specific positive values.

Impact of Content Complexity: a is a decreasing function, while b and c are increasing functions of the video segment content or source complexity ($Ccomplexity$), defined as

$$a = a'' - a_k * Ccomplexity, \qquad b = b'' + b_k * Ccomplexity,$$

$$c = c'' + c_k * Ccomplexity$$

where $a_k > 0$, $b_k > 0$ and $c_k > 0$ are codec-specific constants.

The model has two sets of model coefficients, one set for the PC-Monitor/TV displays and the other for Tablet/Mobile displays. This is logical as subjects may assess the quality differently on different devices. A final linear mapping accounts for slight variation in quality prediction between PC-Monitor and TV, and the Tablet and Mobile cases. For more details on algorithm description see [1], [4].

5.2 BSM0

The BSM0 model is a simplified version of the HYN0 model. BSM0 input does not allow any estimate of the source complexity, so accordingly, the BSM0 model was created by removing the content complexity feature $Ccomplexity$ from HYN0 and replacing it with a constant value. The final set of model coefficients for BSM0 is different from that of HYN0. This is due to the fact that a dedicated retraining was performed to fit the BSM0 model to the training data..

5.3 BSM1

Similar to BSM0, BSM1 is a simplified version of HYN0. The simplification was achieved by computing the source complexity feature based on frame-type and frame-size information. In particular, the ratio of average non-I frame to average I frame size is the most meaningful feature. A low value (closer to 0.0) of this feature indicates a lower complexity video segment and a high value (closer to 1.0) indicates a higher complexity video segment. Like BSM0, the final set of model coefficients for BSM1 is different from that of HYN0.

5.4 HYN1

This model takes into account both HYN0 and BSM1 type of complexity calculations. Other than that, the model structure is very similar to the HYN0 model structure. The final set of model coefficients for HYN1 is different from that of HYN0 however.

5.5 PXFR

OPTICOM's PXFR model is an enhanced version of the PEVQ [5], which has very well-known performance and was previously standardized in ITU-T standards [6] and [7]. In specific, per-frame features like edge-information, blur, contrast, noise, etc., are computed at multiple resolution levels of the video (UHD, HD, SD). These features are then combined using feature scaling and temporal pooling to yield the predicted quality output.

5.6 HYF0

This model is a slight variation of PXFR model, where additionally a normalized encoded bitrate information is calculated and used as an additional parameter for training the model. The final set of model coefficients for HYF0 is different from that of the PXFR model however.

Note: For HYN0, HYN1, PXFR and HYF0 models, a long-term quality prediction function is used to predict the session MOS. The definition of this long-term prediction function is detailed in the appendix 2 of ITU-T P.1204.5 [4].

6. Model Performance

This section contains the following model performance figures for the six OPTICOM models.

- Per-database RMSE
- Weighted average RMSE
- Per-database Pearson correlation
- Per-database scatter plots
- Model qualification thresholds

5.1 Per-database RMSE

The models were evaluated based on one single statistical metric, i.e., the root mean square error (RMSE), aggregated across all databases [1]. The calculation of the RMSE for a model v and database k can be expressed as

$$RMSE_{k,v} = \sqrt{\frac{1}{N_k - 2}\sum_{i=1}^{N_k}\left(s_i - s'_{v,i}\right)}$$

where s_i is the subjective score for the ith sample in the considered test, the score $s'_{v,i}$ denotes the objective score of the model v for the ith sample, and N_k the number of samples in the test k.

In total 26 short and 4 long databases were used for model validation. Out of these, 13 short sequence and 2 long sequence databases were known to the models at the time of model submission and these were used as training databases (P2STR and P2LTR databases). 13 short and 4 long sequence databases were created after the model submission (P2SVL and P2LVL databases) and used for validation.

Table 1: Per-database RMSE values for training databases. P2S (short databases), P2L (long databases).

Database	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
P2STR01	0.5391	0.3996	0.4118	0.4118	0.4597	0.4406
P2STR02	0.6118	0.5419	0.5456	0.5456	0.5848	0.5297
P2STR03	0.5864	0.5544	0.4805	0.4805	0.4137	0.3959
P2STR04	0.4396	0.3985	0.3637	0.3637	0.3680	0.3459
P2STR05	0.5820	0.5042	0.5028	0.5028	0.4663	0.4139
P2STR06	0.5273	0.4130	0.4355	0.4355	0.4825	0.4443
P2STR08	0.4821	0.4640	0.4234	0.4234	0.4788	0.4137
P2STR09	0.5174	0.4378	0.3985	0.3985	0.4452	0.3867
P2STR10	0.4941	0.3848	0.3272	0.3272	0.4349	0.3465
P2STR11	0.6312	0.5238	0.4432	0.4432	0.4034	0.3437
P2STR12	0.5871	0.4859	0.4662	0.4662	0.4088	0.3492
P2STR13	0.4663	0.3930	0.3634	0.3634	0.4365	0.3805
P2STR14	0.4981	0.4648	0.4094	0.4094	0.4452	0.3984
P2SVL15	Not Validated	Not Validated	0.3184	0.3184	0.3500	0.3287
P2SVL17	Not Validated	Not Validated	0.3654	0.3654	0.3104	0.2912

Table 2: Per-database RMSE values for unknown databases. P2S (short databases), P2L (long databases).

Database	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
P2SVL01	0.5157	0.4801	0.3761	0.3990	0.4384	0.3935
P2SVL02	0.4995	0.4898	0.4712	0.4391	0.4940	0.4741
P2SVL03	0.5635	0.5157	0.4360	0.4296	0.4662	0.4485
P2SVL04	0.6723	0.5391	0.5036	0.4558	0.5654	0.5021
P2SVL05	0.5479	0.5162	0.4398	0.4411	0.5111	0.4679
P2SVL06	0.6093	0.5542	0.4586	0.4783	0.4600	0.4312
P2SVL07	0.4895	0.4955	0.4187	0.4230	0.4738	0.3872
P2SVL08	0.5118	0.4151	0.4057	0.3884	0.4708	0.4515
P2SVL09	0.5321	0.5040	0.4647	0.4688	0.4827	0.4230
P2SVL10	0.6291	0.6268	0.5586	0.5768	0.5432	0.5252
P2SVL11	0.5435	0.5073	0.4260	0.4352	0.4542	0.4136
P2SVL12	0.5579	0.5169	0.4957	0.4893	0.4658	0.4489
P2SVL13	0.5126	0.4214	0.4148	0.4001	0.4515	0.4433
P2LVL15	Not Validated	Not Validated	0.5702	0.5668	0.5458	0.5232
P2LVL18	Not Validated	Not Validated	0.5431	0.4625	0.5041	0.4767
P2LVL19	Not Validated	Not Validated	0.4758	0.5260	0.3982	0.3745
P2LVL23	Not Validated	Not Validated	0.6705	0.6797	0.7394	0.7156

5.2 Average Model Performance

For the computation of average RMSE, weighted averaging of per-database RMSE values reported in Table 1 and Table 2 were computed. Each known (training) database was given the weight of 0.1 and an unknown database was given the weight of 0.9. The average RMSE of model v was computed using the following equation

$$RMSE_{avg,v} = \sqrt{\frac{1}{W}\sum_{k=1}^{M} w_k \, . \, RMSE_{k,v}^2}$$

w_k denotes the weight of database k and W denotes the sum of weights of all databases.

As informed before, parametric models were evaluated for short sequence databases only. Pixel-based and hybrid models were evaluated for short as well as long sequence databases. Table 3 and Table 4 report the avg. RMSE for the two evaluations.

Table 3: Average RMSE for short sequence databases.

	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
Avg. RMSE	0.554	0.505	0.452	0.448	0.481	0.444

Table 4: Average RMSE for short + long sequence databases.

	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
Avg. RMSE	Not validated for long dbs	Not validated for long dbs	0.478	0.474	0.498	0.464

5.3 Per-database Pearson Correlation

Per-database Pearson correlation values are tabulated for the six models below.

Table 5: Per-database Pearson correlation values for training databases.

Database	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
P2STR01	0.79	0.89	0.88	0.90	0.85	0.87
P2STR02	0.82	0.86	0.86	0.89	0.84	0.87
P2STR03	0.79	0.81	0.86	0.85	0.90	0.91
P2STR04	0.93	0.94	0.95	0.95	0.95	0.96
P2STR05	0.80	0.85	0.85	0.87	0.88	0.90
P2STR06	0.81	0.89	0.87	0.90	0.84	0.87
P2STR08	0.90	0.91	0.92	0.92	0.90	0.93
P2STR09	0.84	0.89	0.91	0.91	0.88	0.91
P2STR10	0.86	0.92	0.94	0.94	0.90	0.93
P2STR11	0.81	0.87	0.91	0.91	0.93	0.95
P2STR12	0.80	0.87	0.88	0.88	0.91	0.93
P2STR13	0.89	0.92	0.93	0.94	0.90	0.93
P2STR14	0.83	0.85	0.89	0.88	0.86	0.89
P2SVL15	Not Validated	Not Validated	0.92	0.90	0.90	0.91
P2SVL17	Not Validated	Not Validated	0.92	0.92	0.94	0.95

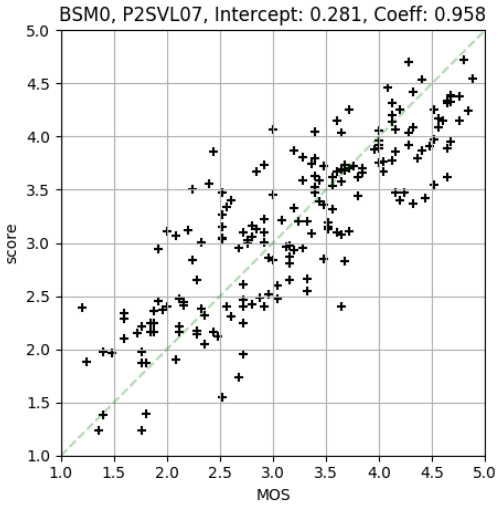
Table 6: Per-database Pearson correlation values for unknown databases.

Database	BSM0	BSM1	HYN0	HYN1	PXFR	HYF0
P2SVL01	0.81	0.85	0.91	0.90	0.88	0.90
P2SVL02	0.81	0.84	0.84	0.87	0.83	0.85
P2SVL03	0.71	0.78	0.85	0.86	0.82	0.84
P2SVL04	0.79	0.88	0.89	0.91	0.86	0.89
P2SVL05	0.85	0.89	0.92	0.92	0.89	0.91
P2SVL06	0.83	0.88	0.91	0.91	0.91	0.92
P2SVL07	0.84	0.86	0.90	0.90	0.87	0.92
P2SVL08	0.83	0.90	0.91	0.92	0.87	0.88
P2SVL09	0.78	0.83	0.85	0.85	0.84	0.88
P2SVL10	0.73	0.75	0.81	0.79	0.81	0.83
P2SVL11	0.83	0.88	0.91	0.91	0.90	0.92
P2SVL12	0.65	0.75	0.75	0.78	0.79	0.81
P2SVL13	0.76	0.86	0.87	0.87	0.83	0.84
P2LVL15	Not Validated	Not Validated	0.81	0.81	0.83	0.84
P2LVL18	Not Validated	Not Validated	0.87	0.91	0.89	0.90
P2LVL19	Not Validated	Not Validated	0.87	0.84	0.91	0.92
P2LVL23	Not Validated	Not Validated	0.82	0.81	0.77	0.79

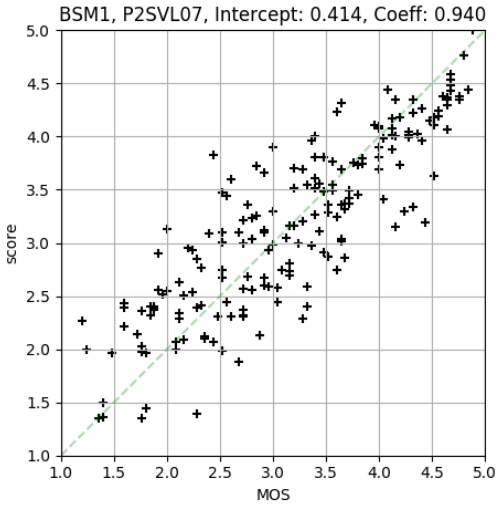
5.4 Scatter Plots

In this section three scatter plots for three unknown short databases (P2SVL07, P2SVL11, P2SVL04) for all six models are reported. Since the lowest complexity BSM0 model yielded lowest RMSE for P2SVL07, median RMSE for P2SVL11, and highest RMSE for P2SVL04. In this sense these three databases cover the difficulty range of the validation dataset.

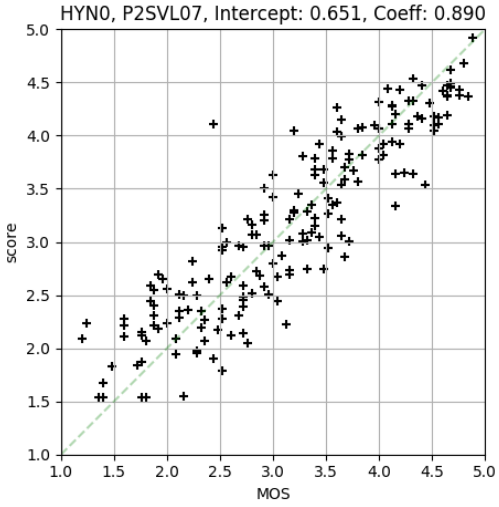
Database 1 – P2SVL07:



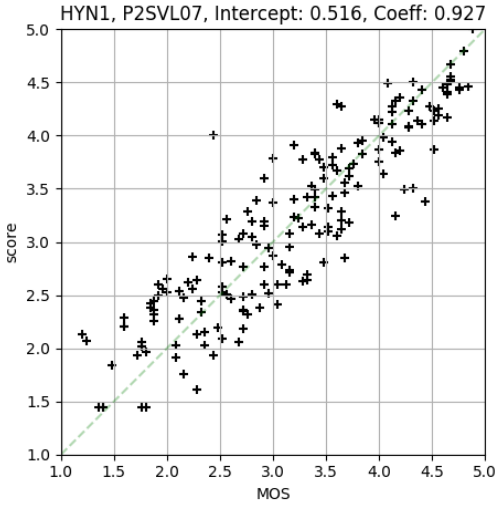
(a) BSM0, RMSE=0.4895, Corr=0.84



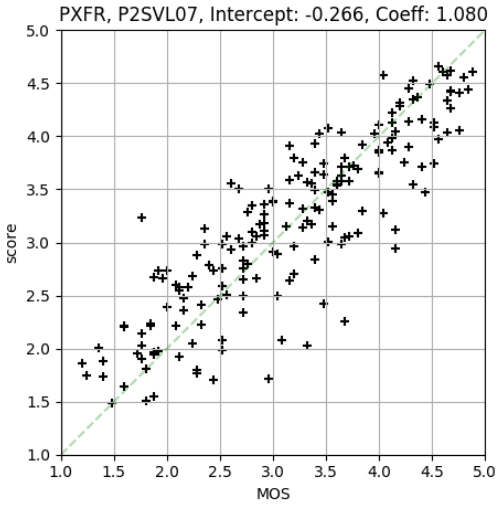
(b) BSM1, RMSE=0.4955, Corr=0.86



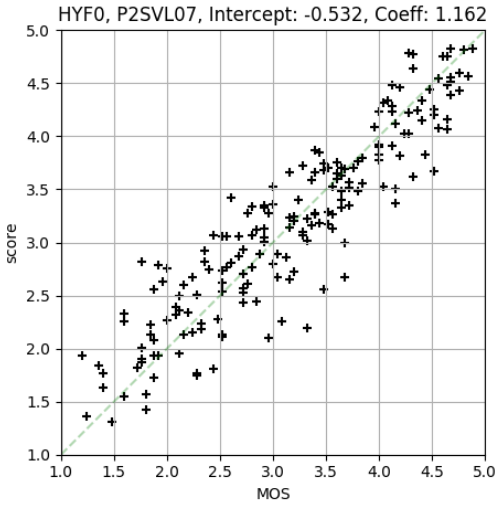
(c) HYN0, RMSE=0.4187, Corr=0.90



(d) HYN1, RMSE=0.4230, Corr=0.90



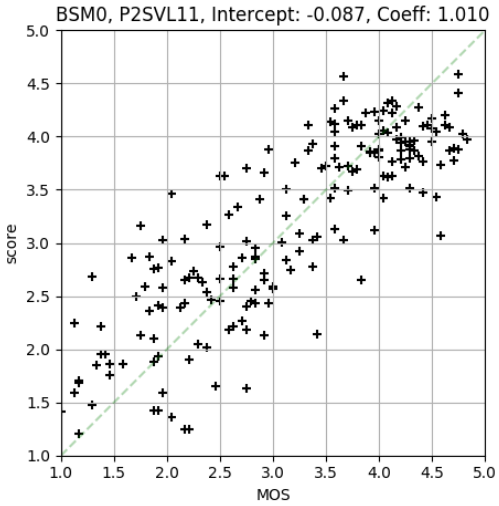
(e) PXFR, RMSE=0.4738, Corr=0.87



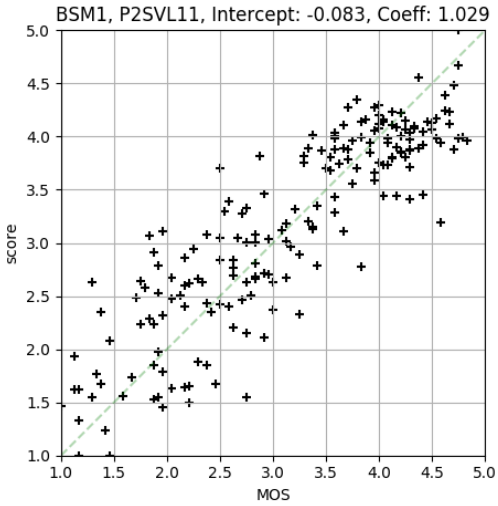
(f) HYF0, RMSE=0.3872, Correl=0.92

Figure 1: P2SVL07 scatter plots for six video quality models.

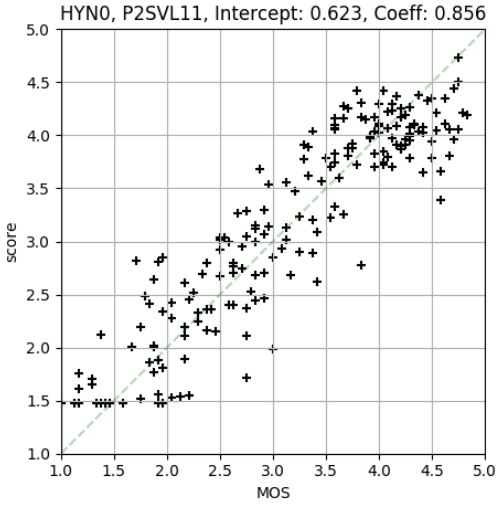
Database 2 – P2SVL11:



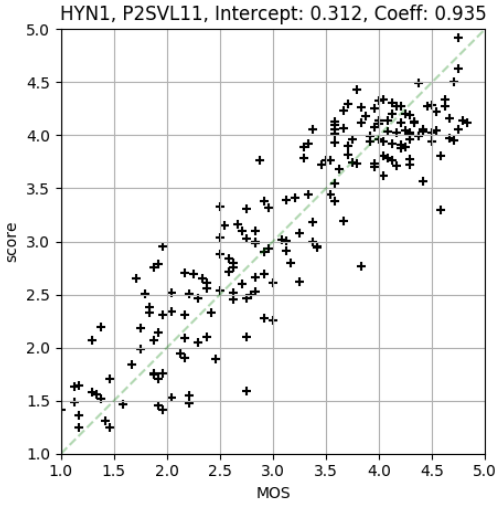
(a) BSM0, RMSE=0.5435, Corr=0.83



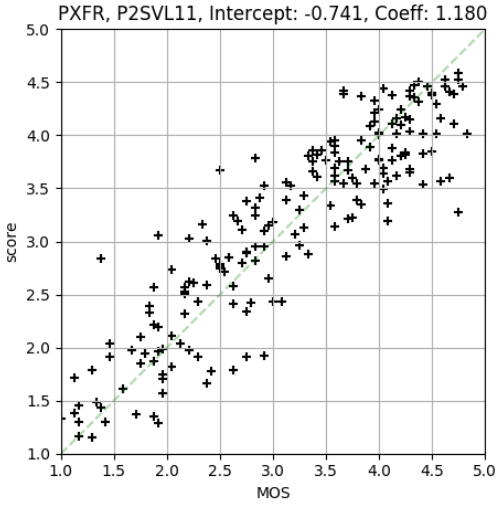
(b) BSM1, RMSE=0.5073, Corr=0.88



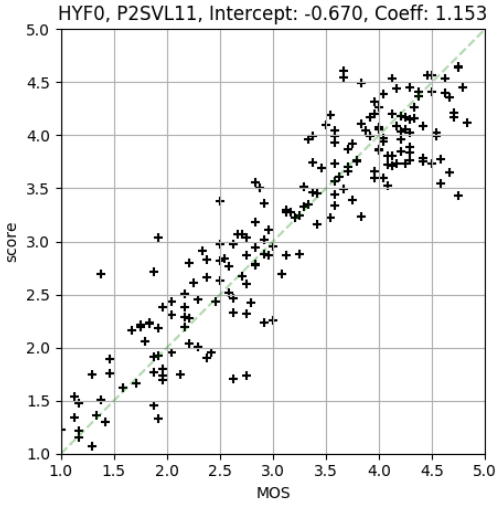
(c) HYN0, RMSE=0.4260, Corr=0.91



(d) HYN1, RMSE=0.4352, Corr=0.91



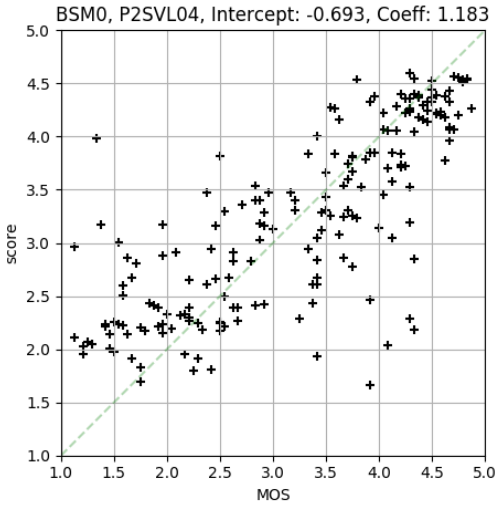
(e) PXFR, RMSE=0.4542, Corr=0.90



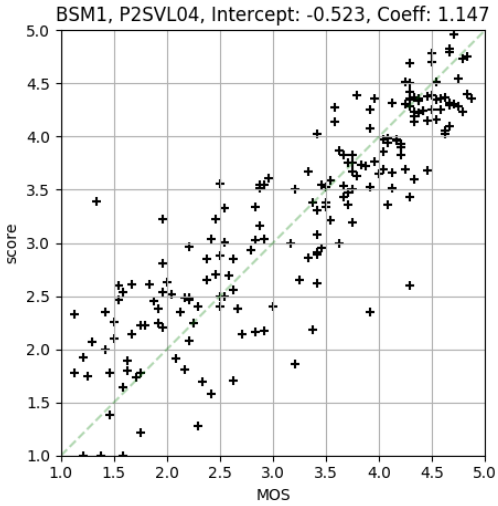
(f) HYF0, RMSE=0.4136, Corr=0.92

Figure 2: P2SVL11 scatter plots for six video quality models.

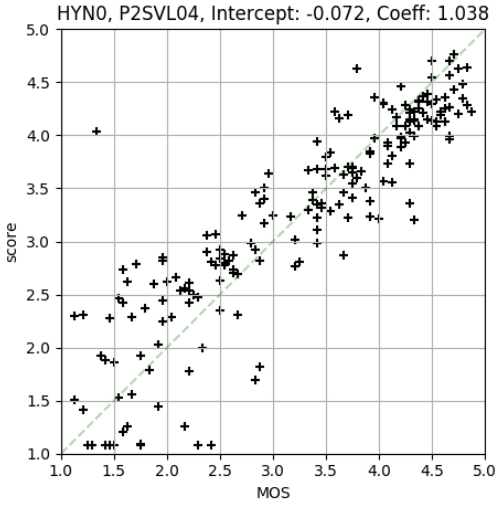
Database 3 – P2SVL04:



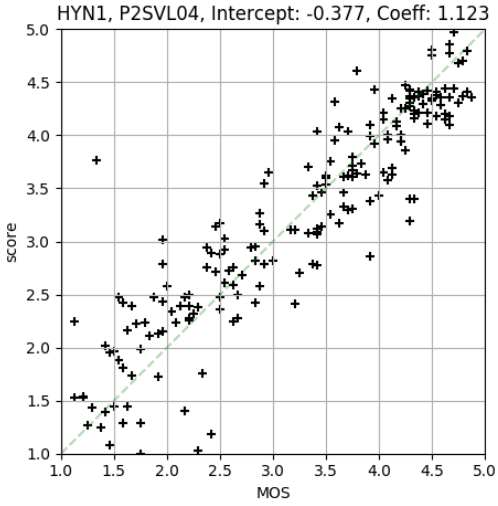
BSM0, RMSE=0.6723, Corr=0.79



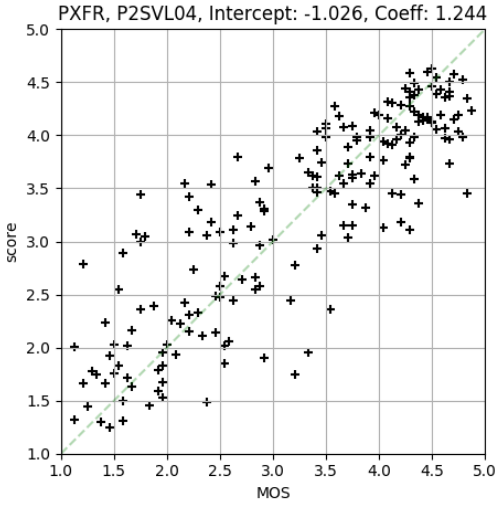
BSM1, RMSE=0.5391, Corr=0.88



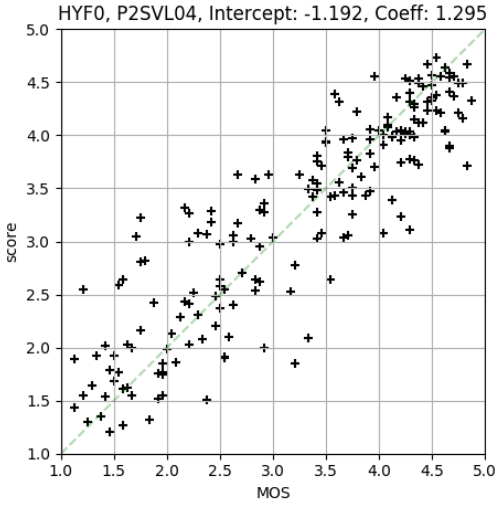
HYN0, RMSE=0.5036, Corr=0.89



HYN1, RMSE=0.4558, Corr=0.91



PXFR, RMSE=0.5654, Corr=0.86



HYF0, RMSE=0.5021, Corr=0.89

Figure 3: P2SVL04 scatter plots for six video quality models.

Figure 1 depicts the scatter plots for the lowest RMSE database P2SVL07. Compared to BSM0, the scatter plot for BSM1 looks more evenly spaced around the unity line. These leads to a slightly better Pearson correlation value compared to BSM0. However, RMSE of BSM1 is slightly worse compared to BSM0. This is an a-typical disagreement between RMSE and correlation. The performance highly improves for HYN0. However, HYN1 does not lead to any further improvement over HYN0. PXFR model leads to a suboptimal prediction for this database as it performs far worse compared to the HYN0. HYF0 on the other hand leads to far better prediction compared to all other models.

Figure 2 shows the scatter plots for the median RMSE database P2SVL11. In general, this database yields similar rank order of RMSE value across models as of P2SVL07. One exception being the BSM1 which understandably yields better RMSE compared to BSM0.

Figure 3 depicts the scatter plots for highest RMSE database P2VL04. This database proves extremely difficult for BSM0 as BSM0 yielded a relatively higher value of RMSE compared to other databases. However, moving from BSM0 to BSM1, then to HYN0, then to HYN1 yields significant performance improvements. PXFR is deficient in this case also. HYF0 improve the performance of PXFR but it is largely deficient to HYN0/HYN1.

5.5 Model Qualification Thresholds

All the models in the competition were judged on the basis of their average RMSE value and certain significance value which determines if two models are statistically equivalent or not. The exact values presented in the tables below were previously published in [3].

In particular, for a certain model category, the lower threshold (lower_th) is set by the best model (lowest average RMSE value model) of that category. The upper threshold (upper_th) is an offset with respect to the lower_th which depends on the value of the lower_th and on the number of samples used for model validation. Any candidate model of that category falling in between lower_th and upper_th was considered equivalent to the best model of that category [2]. This defines the intra-category qualification criterion – criterion 1.

The lower_th and upper_th for model categories where OPTICOM submitted the models, along with their avg. RMSE performance are reported in Table 7. Table 7 reports the threshold values for short models/databases only.

In addition to meeting the aforementioned significance requirement, each model in a category needs to meet the minimum performance requirement criterion. The upper_th for a model category needs to be smaller than lower_th of the any lower complexity model category (see 2nd column in Table 7). This defines the inter-category qualification criterion – criterion 2. The model complexity here is defined in terms of the input given to the model. That is:

- BSM1 Is higher complexity than BSM0. The baseline model provides the minimum requirement for BSM0.
- HYN0 is higher complexity than BSM0 and pixel-based no-reference model.
- HYN1 is higher complexity than both BSM1 and HYN0. Note that, in this case HYN0 gives the minimum performance requirement for HYN1 as lower_th for HYN0 is smaller than lower_th of BSM1.
- PXFR is more complex than pixel-based reduced-reference model.
- HYF0 is higher complexity compared to both PXFR and BSM0, where PXFR being a lower RMSE category provides the minimum performance requirement for HYF0 models.

Table 7: Significance thresholds and minimum average RMSE requirements for short video models.

Model Type	Minimum Required Avg. RMSE	Passing Threshold Avg. RMSE (*) (upper_th)	Best Model Avg. RMSE (*) (lower_th)	OPTICOM Model Avg. RMSE
BSM0	0.610 (baseline)	0.570	**0.554**	**0.554**
BSM1	0.554	0.520	**0.505**	**0.505**
HYN0	0.554	0.466	**0.452**	**0.452**
HYN1	0.452	0.468	**0.448**	**0.448**
PXFR	0.444[1]	0.447	0.434	0.481
HYF0	0.434[2]	0.458	**0.444**	**0.444**

(*) values as reported to VQEG and published by AVHD / P.NATS Phase 2 proponents in [3].

Similar to Table 7, Table 8 reports the threshold values for short and long models together.

Table 8: Significance threshold and minimum average RMSE requirement for short and long video models.

Model Type	Minimum Required RMSE	Best Model RMSE (*) (lower_th)	OPTICOM Model RMSE
HYN0	0.610 (baseline)	**0.478**	**0.478**
HYN1	0.478	**0.474**	**0.474**
PXFR	0.457	Not disclosed for legal reasons	0.498
HYF0	0.457	Not disclosed for legal reasons	0.464

(*) values as reported to VQEG and published by AVHD / P.NATS Phase 2 proponents in [3].

[1] Lowest Avg. RMSE achieved by a pixel-based reduced-reference model in the competition for short databases.

[2] Lowest Avg. RMSE achieved by a pixel-based full-reference model in the competition for short databases.

Evaluation Results:

Unfortunately only very limited comparisons of model performance can be published due to an NDA between the proponents, but based on information already in the public domain ([1], [2] and [3]), the following findings can be reported:

- For short databases, OPTICOM's models BSM0, BSM1, HYN0, HYN1 and HYF0 have performance identical to the model defining the best performance *lower_th* in each category.
- For short plus long databases combined, OPTICOM's models HYN0, HYN1 have performance identical to the model defining the best performance *lower_th* in each category.
- OPTICOM's candidate models for BSM0, BSM1 and HYN0 met the two performance criteria for short databases.
- Additionally, HYN0 also meets the two criteria for both short and short+long databases.
- HYN1 meets the criterion 1 however not criterion 2 for both short and short+long databases.
- PXFR model does not meet both critera for both short and short+long databases.
- HYF0 meet criterion 1 but not criterion 2 for short databases. For short+long databases, HYF0 does not meet both criteria.

Based on these conclusions OPTICOM's BSM0, BSM1, HYN0 models were part of the winning group of the AVHD-PNATS2 competition.

7. Summary

The VQEG report [2] of the AVHD-PNATS2 project which was a joint effort of ITU-T SG12/Q14 and VQEG only described the performance figures of the models which were standardized as an outcome of the AVHD-PNATS2 project. The present report in contrast details the model performance figures of all six models submitted by OPTICOM GmbH to the AVHD-PANTS2 competition. A detailed insight into the candidate models of OPTCOM is provided using per-database performance, scatter plots and analysis based on average RMSE figures. Two bitstream models (BSM0, BSM1) and one hybrid no-reference (HYN0) model fulfilled the qualification criteria to be considered as part of the winning group.

8. References

[1] A. Raake *et al*., "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204," in *IEEE Access*, vol. 8, pp. 193020-193049, 2020, doi: 10.1109/ACCESS.2020.3032080.

[2] AVHD-AS/P.NATS Phase 2 Status update – Dec. 2020 VQEG meeting, ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Stockholm_Dec20/AVHD-ASP.NATS_Phase_2_Model_Performance_Report_v02.pdf

[3] AVHD-AS/P.NATS Phase 2 Status update – Oct. 2019 VQEG meeting. ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Shenzhen_Oct19/VQEG_AVHD_2019_October_AVHD-AS_P.NATS_overview.pptx

[4] ITU-T Rec. P.1204.5, ”Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information”. https://www.itu.int/rec/T-REC-P.1204.5

[5] Percentual Evaluation of Video Quality – OPTICOM GmbH http://www.pevq.com/index.html

[6] ITU-T J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”

[7] ITU-T J.341, “Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference”

9. Contact Information

OPTICOM

Dipl.-Ing. M. Keyhl GmbH

Nägelsbachstrasse 38

D - 91052 Erlangen

GERMANY

Phone: +49 (0) 91 31 - 5 30 20 - 0

Fax: +49 (0) 91 31 - 5 30 20 - 20

EMail: info@opticom.de

Webseite: http://www.opticom.de

Further information:

http://www.pevq.com