|  |  |
| --- | --- |
| **Joint Collaborative Team on Video Coding (JCT-VC)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  23rd Meeting: San Diego, USA, 19–26 February 2016 | Document: JCTVC-W0091 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **HDR CE3: Benchmarking of objective metrics for HDR video quality assessment** | | |
| *Status:* | Input Document | | |
| *Purpose:* | Information | | |
| *Author(s) or Contact(s):* | Name(s) Philippe Hanhart Touradj Ebrahimi | Email: | [philippe.hanhart@epfl.ch](mailto:philippe.hanhart@epfl.ch) |
| *Source:* | Ecole Polytechnique Fédérale de Lausanne (EPFL) | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This contribution reports the performance of several objective metrics for HDR quality assessment. To benchmark the metrics, we used as ground truth the subjective scores collected in two different evaluations conducted at EPFL [1,2]. Three different analyses are performed to assess the performance of the different metrics. Based on the results, we recommend using HDR-VDP-2 or PSNR-DE1000, which shows similar performance for a lower computational complexity and considers color differences.

# Introduction

An important issue for the development of future HDR video compression algorithms is related to the selection of the objective metrics used to measure quality. Currently, there is no agreement on which metrics should be used as there is not enough evidence than one metric outperforms significantly over others. In a previous contribution [3], we have already addressed this problem, but the database used to benchmark the metrics was rather limited. In this contribution, we used the subjective scores collected during the evaluations of the CfE to further benchmark several objective metrics. In particular, we used the 196 MOS and corresponding CI values of a recent evaluation conducted using the DSIS method [2] to measure the correlation between objective and subjective scores using widely used state of the art methods, i.e., the linearity, monotonicity, accuracy, and consistency of the estimation of MOS. To further evaluate the effectiveness of considered metrics, we used the results of our first evaluation conducted using partial pair comparison method [1], which provides a higher discrimination power than the DSIS method. From the 176 paired comparisons, we computed the classification errors of the metrics and we performed a new analysis to determine the ability of the metrics to distinguish visible quality differences.

# Objective quality metrics

The following objective quality metrics were benchmarked in this contribution:

1. Metrics computed in linear domain
2. PSNR-x: PSNR computed on x component
3. PSNR-DEx: PSNR of mean of absolute value of deltaE2000 metric, derived with x as reference luminance value
4. PSNR-Lx: PSNR of mean square error of L component of the CIELab color space used for the deltaE2000 metric, derived with x as reference luminance value
5. avLumaErr: average error in Barten Steps [4,5]
6. avLumaPSNR: PSNR of average error in Barten Steps [4,5]
7. avColorErr: average color error [4,5]
8. avColorPSNR: PSNR of average color error [4,5]
9. HDR-VDP-2 [6]
10. HDR-VQM [7]
11. Metrics computed in PQ-TF domain [8]
12. tPSNR-x: PSNR computed on x component
13. PQ2SSIM
14. PQ2MSSSIM
15. PQ2VIFP: VIF pixel based
16. Metrics computed using multi-exposure [9]
17. mPSNR

SSIM, MS-SSIM, and VIFP were computed using MeTriX MuX Visual Quality Assessment Package[[1]](#footnote-1). For these three metrics, the luminance information was extracted from the RGB values, clipped to the range [0.005,4000] cd/m2, transformed using the PQ EOTF, and normalized to the interval [0,255] before computing the metric. The MATLAB implementations of HDR-VDP-2[[2]](#footnote-2) and HDR-VQM[[3]](#footnote-3) were used. The remaining metrics were computed using the HDRTool software (v0.9) modified by Philips to implement their Luvstar metric. For contents *ShowGirl2* and *WarmNight*, the top and bottom black borders were discarded when computing the metrics. For content *Market3*, the metrics were computed on the first 240 frames that were used in [1]. We assumed that there would not be much differences in the remaining 150 frames, as the cropped part contains rather constant/similar motion between the beginning and the end of the sequence.

*Note that for better orientation in the results, the tPSNR-Y metric corresponding to PSNR after PQ-TF on RGB and YUV conversion was renamed to tPSNR-Yyuv and the tPSNR-Y metric corresponding to PSNR on XYZ after conversion from RGB to XYZ and PQ-TF on XYZ was renamed to tPSNR-Yxyz. Similarly, the tPSNR-Y metric corresponding to PSNR after PQ-TF on RGB and Yu’v’ conversion was renamed to tPSNR-Yyupvp.*

# Statistical analysis

The results of subjective visual experiments are considered as ground truth to evaluate how well an objective quality metric estimates perceived quality. The result of execution of a particular objective metric is an objective quality rating (OQR), which is expected to be the estimation of the MOS corresponding to a video sequence.

## Classification errors

A regression was fitted to the [OQR, MOS] data set to map the objective scores to the subjective ratings. Note that different objective metrics typically have different range of values, so the mapping to a common scale also facilitates the comparison of different models. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a non-linear mapping function was used:

where is the predicted MOS, is the objective metric result, and , , and are the parameters that define the shape of the logistic mapping function and were determined via the least squares method.

The following properties of the estimation of MOS were considered: linearity, monotonicity, accuracy, and consistency. To this end, four different performance indexes were computed between the ground truth and predicted subjective scores. In particular, the Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation (SROCC) were computed to estimate linearity and monotonicity, respectively. Accuracy and consistency were estimated using the root-mean-square error (RMSE) and outlier ratio (OR), respectively. The OR is the ratio of points for which the error between the predicted and actual MOS values exceeds the 95% confidence interval of MOS values.

To determine whether the difference between two performance index values corresponding to two different metrics is statistically significant, two-sample statistical tests were performed on all four performance indexes. In particular, for the PLCC and SROCC, a *Z*-test was performed using Fisher z-transformation. For the RMSE, an *F*-test was performed, whereas a *Z*-test for the equality of two proportions was performed for the OR. No correction was applied to correct for the multiple comparisons.

The statistical tests were performed according to the guidelines of recommendation ITU-T P.1401 [10].

## Classification errors

In recommendation ITU-T J.149 [11], it is suggested to compute the classification errors to evaluate the performance of an objective metric. A classification error is performed when the objective metric and subjective evaluation lead to different conclusions on a pair of video sequences, *A* and *B*, for example. Three types of error can happen

1. *False Tie*, the least offensive error, which occurs when the subjective evaluation says that *A* and *B* are different, whereas the objective scores say that they are identical,
2. *False Differentiation*, which occurs when the subjective evaluation says that *A* and *B* are identical, whereas the objective scores say that they are different,
3. *False Ranking*, the most offensive error, which occurs when the subjective evaluation says that *A* (*B*) is better than *B* (*A*), whereas the objective scores say the opposite.

**Table 1 – Classification errors**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Subjective | | |
|  |  | A > B | A = B | A < B |
| Objective | A > B | *Correct Decision* | *False Differentiation* | *False Ranking* |
| A = B | *False Tie* | *Correct Decision* | *False Tie* |
| A < B | *False Ranking* | *False Differentiation* | *Correct Decision* |

To determine whether two groups of subjective scores are statistically different, a simple binomial test was used as in [1]. The percentage of *Correct Decision*, *False Tie*, *False Differentiation*, and *False Ranking* were then recorded from all possible distinct pairs as a function of the difference in the metric values, .

As increases, more pairs of data points are considered as equivalent by the objective metric. This reduces the occurrences of *False Differentiations* and *False Rankings*, but increases the occurrence of *False Ties*. On the other hand, as tends towards 0, the occurrence of *False Tie* will tend towards 0, while the occurrence of *False Differentiation* will tend towards the proportion of pairs of data points where there was not enough evidence to show a statistical difference in the subjective evaluation.

The relative frequencies are plotted as a function of the significance threshold . Ideally, the occurrence of *Correct Decision* should be maximized and the occurrence of *False Ranking* should be minimized when the tends towards 0. The occurrences of *False Differentiations* and *False Rankings* should decrease as fast as possible as increases. Based on this, different graphs corresponding to different metrics can be compared to determine the best metric for the application under analysis.

## ROC analysis

We present a new analysis inspired from the classification errors and receiver operating characteristic (ROC) analysis. The ROC curve illustrates the performance of a binary classifier system as its discrimination threshold is varied. However, when comparing a pair of video sequences, and , there are three possible outcomes: , , or . Hence, the outcome of the comparison is ternary and a direct ROC analysis cannot be performed. Therefore, we perform three separate ROC analyses where we consider only binary classification:

1. *Different/Similar ROC Analysis*: this analysis illustrates the ability of the metric to discriminate between significant and not significant visual quality differences in a pair of video sequences. In this case, all data points are considered.
2. *Better/Worse ROC Analysis*: considering pairs with significant visual quality differences, this analysis illustrates the ability of the metric to determine which video sequence in a pair has the best visual quality. In this case, only data points corresponding to pairs with significant difference are considered.
3. *Better/Equal-Worse ROC Analysis*: this analysis illustrates the ability of the metric to determine whether a specific video sequence has similar or worse visual quality than a reference video sequence or if it has significantly better visual quality than the reference. In this case, all data points are considered.

Figure 1 illustrates the analysis. First, the dataset is split into pairs with and without significant differences in terms of visual quality based on the pair comparison test. Then, different classes, namely, , , , , and are formed. For each ROC analysis, the histogram of the two corresponding classes is constructed as a function of the metric difference between the two video sequences. Note that for the *Different/Similar ROC Analysis*, the absolute metric difference is used. For the two other ROC analyses, the data is repeated to have both AB and BA pairs. Finally, the discrimination threshold is set on and varied to build the ROC curve. The area under the curve (AUC) can be computed as a simple indicator to easily compare ROC curves.

Macintosh HD:Users:philippe:Downloads:diagram.pdf

**Figure 1 - ROC analysis: creation of the different classes.**

# Results and discussions

Figure 2 depicts the scatter plots of subjective versus objective results for these metrics. As it can be observed, the data points are well concentrated near the mapping curve for HDR-VDP-2, as well as for PQ2VIFP, whereas they are more scattered for the other metrics, especially in the case of PSNR in linear domain and in PQ-TF domain on the Yu’v’ color space, as well as mPSNR, PSNR-DEx, PSNR-Lx, and HDR-VQM, which show higher content dependency. These findings indicate that HDR-VDP-2 and PQ2VIFP have a very high consistency when compared to the other metrics when all contents are considered.

|  |  |  |
| --- | --- | --- |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-R.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-G.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-B.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:mPSNR.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-Yyuv.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-U.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-V.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-YUV.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-R.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-G.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-B.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-RGB.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-X.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-Yxyz.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-Z.png |
| **Figure 2 – Subjective versus objective results. The black curve represents the mapping function. For the data points, each color corresponds to a specific content.** | | |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-XYZ.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-Yyupvp.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-up.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-vp.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:tPSNR-Yupvp.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-DE100.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-L100.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-DE1000.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PSNR-L1000.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:avLumaPSNR.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:avColorPSNR.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:avLumaErr.png |
| Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:avColorErr.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PQ2SSIM.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:PQ2MS-SSIM.png |
| **Figure 2 – Subjective versus objective results. The black curve represents the mapping function. For the data points, each color corresponds to a specific content. *(Continued)*** | | |
|  | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:HDR-VDP-2.png | Description: Data:new:SubjectiveTests:MPEG_HDR_DSIS:MOSvsMetrics:HDR-VQM.png |

**Figure 2 – Subjective versus objective results. The black curve represents the mapping function. For the data points, each color corresponds to a specific content. *(Continued)***

**Table 2 – Linearity, monotonicity, accuracy, and consistency indexes for the different metrics: considering all contents at once.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | PLCC | SROCC | RMSE | OR |
| PSNR-R | 0.2883 | 0.2045 | 1.1212 | 0.6990 |
| PSNR-G | 0.2966 | 0.1815 | 1.1182 | 0.6888 |
| PSNR-B | 0.3016 | 0.1311 | 1.1164 | 0.6837 |
| mPSNR | 0.4325 | 0.4435 | 1.0557 | 0.7347 |
| tPSNR-Yyuv | 0.6332 | 0.6038 | 0.9063 | 0.6633 |
| tPSNR-U | 0.4653 | 0.3861 | 1.0364 | 0.7449 |
| tPSNR-V | 0.4668 | 0.4577 | 1.0355 | 0.7245 |
| tPSNR-YUV | 0.5718 | 0.5667 | 0.9606 | 0.7041 |
| tPSNR-R | 0.5720 | 0.5629 | 0.9605 | 0.6837 |
| tPSNR-G | 0.6292 | 0.5963 | 0.9100 | 0.6429 |
| tPSNR-B | 0.5403 | 0.5063 | 0.9853 | 0.6582 |
| tPSNR-RGB | 0.5767 | 0.5732 | 0.9566 | 0.7296 |
| tPSNR-X | 0.6042 | 0.6104 | 0.9330 | 0.7245 |
| tPSNR-Yxyz | 0.6325 | 0.6093 | 0.9069 | 0.6786 |
| tPSNR-Z | 0.5578 | 0.5188 | 0.9719 | 0.6735 |
| tPSNR-XYZ | 0.6020 | 0.5953 | 0.9350 | 0.7347 |
| tPSNR-Yyupvp | 0.6325 | 0.6093 | 0.9069 | 0.6786 |
| tPSNR-up | 0.3888 | 0.3725 | 1.0788 | 0.7500 |
| tPSNR-vp | 0.3691 | 0.3285 | 1.0882 | 0.7500 |
| tPSNR-Yupvp | 0.3558 | 0.3569 | 1.0943 | 0.7653 |
| PSNR-DE100 | 0.4049 | 0.2633 | 1.0706 | 0.6582 |
| PSNR-L100 | 0.5078 | 0.3353 | 1.0087 | 0.6684 |
| PSNR-DE1000 | 0.4852 | 0.2895 | 1.0238 | 0.6276 |
| PSNR-L1000 | 0.4467 | 0.2582 | 1.0476 | 0.6684 |
| avLumaPSNR | 0.6309 | 0.6031 | 0.9085 | 0.6531 |
| avColorPSNR | 0.4681 | 0.4597 | 1.0347 | 0.7245 |
| avLumaErr | 0.4398 | 0.4281 | 1.0516 | 0.7194 |
| avColorErr | 0.4003 | 0.4147 | 1.0730 | 0.7296 |
| PQ2SSIM | 0.5080 | 0.4137 | 1.0085 | 0.7092 |
| PQ2MS-SSIM | 0.6401 | 0.5938 | 0.8996 | 0.6327 |
| **PQ2VIFP** | **0.7255** | **0.7001** | **0.8059** | **0.5459** |
| **HDR-VDP-2** | **0.8632** | **0.8658** | **0.5911** | **0.4847** |
| HDR-VQM | 0.4618 | 0.3746 | 1.0386 | 0.7092 |

Table 2 reports the linearity, monotonicity, accuracy, and consistency indexes for the metrics computed in the different domains when the mapping is applied on all contents at once (as in Figure 2). Results show that HDR-VDP-2 has the best correlation with human perception of visual quality (with PLCC and SROCC values above 0.86), followed by VIFP computed in the PQ-TF domain. However, the statistical tests demonstrate that HDR-VDP-2 is statistically significantly better than PQ2VIFP on the PLCC, SROCC, and RMSE indexes, whereas there is not sufficient evidence to show statistical differences on the OR index. Nevertheless, the OR of HDR-VDP-2 is statistically significantly lower than for the other metrics. All other metrics have poor correlation with perceived quality and a large prediction error.

**Table 3 – Linearity, monotonicity, accuracy, and consistency indexes for the different metrics: average over all contents.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | PLCC | SROCC | RMSE | OR |
| **PSNR-R** | **0.9569** | **0.9508** | **0.3086** | **0.2289** |
| **PSNR-G** | **0.9559** | **0.9421** | **0.3149** | **0.2256** |
| **PSNR-B** | **0.9509** | **0.9460** | **0.3400** | **0.2883** |
| mPSNR | 0.8397 | 0.8162 | 0.5764 | 0.4622 |
| tPSNR-Yyuv | 0.9115 | 0.8830 | 0.4330 | 0.3222 |
| tPSNR-U | 0.8074 | 0.7837 | 0.6100 | 0.4833 |
| tPSNR-V | 0.7551 | 0.7526 | 0.6995 | 0.5589 |
| tPSNR-YUV | 0.8935 | 0.8667 | 0.4438 | 0.3361 |
| tPSNR-R | 0.8979 | 0.8790 | 0.4498 | 0.3711 |
| tPSNR-G | 0.9136 | 0.8858 | 0.4219 | 0.3222 |
| tPSNR-B | 0.8721 | 0.8482 | 0.5001 | 0.3717 |
| tPSNR-RGB | 0.8916 | 0.8675 | 0.4506 | 0.3306 |
| tPSNR-X | 0.9170 | 0.8965 | 0.4193 | 0.3417 |
| tPSNR-Yxyz | 0.9153 | 0.8889 | 0.4220 | 0.3422 |
| tPSNR-Z | 0.8932 | 0.8773 | 0.4699 | 0.3617 |
| tPSNR-XYZ | 0.9088 | 0.8836 | 0.4310 | 0.3317 |
| tPSNR-Yyupvp | 0.9153 | 0.8889 | 0.4220 | 0.3422 |
| tPSNR-up | 0.6046 | 0.5746 | 0.8209 | 0.5967 |
| tPSNR-vp | 0.6716 | 0.6300 | 0.7458 | 0.5422 |
| tPSNR-Yupvp | 0.6436 | 0.6858 | 0.6818 | 0.5439 |
| PSNR-DE100 | 0.8895 | 0.8781 | 0.4976 | 0.4572 |
| **PSNR-L100** | **0.9581** | **0.9517** | **0.3095** | **0.2189** |
| PSNR-DE1000 | 0.8959 | 0.8824 | 0.4908 | 0.4517 |
| **PSNR-L1000** | **0.9605** | **0.9525** | **0.3010** | **0.2189** |
| avLumaPSNR | 0.9235 | 0.9003 | 0.4026 | 0.3372 |
| avColorPSNR | 0.8414 | 0.8316 | 0.5958 | 0.5244 |
| avLumaErr | 0.9317 | 0.9080 | 0.3872 | 0.3117 |
| avColorErr | 0.8416 | 0.8351 | 0.5939 | 0.5239 |
| PQ2SSIM | 0.9337 | 0.9205 | 0.3815 | 0.3161 |
| PQ2MS-SSIM | 0.9368 | 0.9211 | 0.3790 | 0.3317 |
| PQ2VIFP | 0.9205 | 0.8990 | 0.4213 | 0.3406 |
| **HDR-VDP-2** | **0.9615** | **0.9542** | **0.3002** | **0.2244** |
| **HDR-VQM** | **0.9441** | **0.9295** | **0.3575** | **0.2856** |

In the case of codec optimization, it is more important to know that an increase (decrease) in the metric value computed on a specific content will correspond to an increase (decrease) in visual quality rather than to be able to relate a metric score of any content to an absolute quality level. Thus we have performed the same analysis, but for each content separately. Table 3 reports the linearity, monotonicity, accuracy, and consistency indexes for the metrics computed in the different domains when the mapping is applied on each content separately. In this case, the performance indexes were computed separately on each content and then averaged across contents. Results show that most metrics achieve a relatively high correlation with perceived quality as most correlation coefficients are above 0.8. As the mapping is applied on each content separately, metrics that showed strong content dependency previously achieve significantly better performance in this case. The top performing metrics with PLCC and SROCC values above 0.9, RMSE below 0.4, and OR below 0.3 are: HDR-VDP-2, PSNR-Lx, PSNR computed in the linear domain, and HDR-VQM. Note that because of the relatively low number of data points per content, no statistical tests were performed in this case.

Figure 3 reports the classification errors for each metric separately. Even though the results are reported in the native scale of the metric instead of a common scale, it is still possible to compare the classification errors of the different metrics by looking at the relative ∆*OM* ratio (∆*OM* divided by the maximum value of ∆*OM*) rather than the absolute ∆*OM*.

|  |  |  |
| --- | --- | --- |
| PhD:Publications:SPIE2015:paper:classification errors:PSNR-R.png | PhD:Publications:SPIE2015:paper:classification errors:PSNR-G.png | PhD:Publications:SPIE2015:paper:classification errors:PSNR-B.png |
| PhD:Publications:SPIE2015:paper:classification errors:mPSNR.png |  | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-U.png |
| PhD:Publications:SPIE2015:paper:classification errors:tPSNR-V.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-YUV.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-R.png |
| **Figure 3 – Frequencies of classification error: *Correct Decision* (dark blue), *False Tie* (light blue), *False Differentiation* (orange), and *False Ranking* (red). The dashed lines indicate the ∆OM value that maximizes the *Correct Decision* frequency.** | | |
| PhD:Publications:SPIE2015:paper:classification errors:tPSNR-G.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-B.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-RGB.png |
| PhD:Publications:SPIE2015:paper:classification errors:tPSNR-X.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-Yxyz.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-Z.png |
| PhD:Publications:SPIE2015:paper:classification errors:tPSNR-XYZ.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-Yyupvp.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-up.png |
| PhD:Publications:SPIE2015:paper:classification errors:tPSNR-vp.png | PhD:Publications:SPIE2015:paper:classification errors:tPSNR-Yupvp.png | PhD:Publications:SPIE2015:paper:classification errors:PSNR-DE0100.png |
| **Figure 3 – Frequencies of classification error: *Correct Decision* (dark blue), *False Tie* (light blue), *False Differentiation* (orange), and *False Ranking* (red). The dashed lines indicate the ∆OM value that maximizes the *Correct Decision* frequency. *(Continued)*** | | |
| PhD:Publications:SPIE2015:paper:classification errors:PSNR-L0100.png | PhD:Publications:SPIE2015:paper:classification errors:PSNR-DE1000.png | PhD:Publications:SPIE2015:paper:classification errors:PSNR-L1000.png |
| PhD:Publications:SPIE2015:paper:classification errors:avLumaPSNR.png | PhD:Publications:SPIE2015:paper:classification errors:avColorPSNR.png | PhD:Publications:SPIE2015:paper:classification errors:avLumaErr.png |
| PhD:Publications:SPIE2015:paper:classification errors:avColorErr.png | PhD:Publications:SPIE2015:paper:classification errors:SSIMexr.png | PhD:Publications:SPIE2015:paper:classification errors:MS-SSIMexr.png |
|  | PhD:Publications:SPIE2015:paper:classification errors:HDR-VDP-2.png | PhD:Projects:MPEG HDR CfE:objective:HDR-VQM.png |

**Figure 3 – Frequencies of classification error: *Correct Decision* (dark blue), *False Tie* (light blue), *False Differentiation* (orange), and *False Ranking* (red). The dashed lines indicate the ∆OM value that maximizes the *Correct Decision* frequency. *(Continued)***

Subjective results reported in [1] showed that there were many cases where the Proponent version was providing similar quality when compared to the Anchor. More precisely, in 55% of cases, no statistically significant difference was observed between Proponents and Anchor. This value determines the plateau for the *Correct Decision* and *False Tie* frequencies, i.e., if the threshold on ∆*OM* is set to infinite, all pairs of video sequences are considered as equal for the objective metric, which will lead to a *Correct Decision* frequency of 55%, as 55% of the pairs were evaluated as not statistically different in the subjective evaluations. Similarly, the plateau for the *False Tie* frequency is 100% - 55% = 45%. As it can be observed on Figure 3, most metrics barely achieve *Correct Decision* frequency about 60%, which means that they cannot distinguish between quality levels when comparing a pair of video sequences. In particular, HDR-VQM never exceeds 55% *Correct Decision*, so simply saying that the two video sequences have similar quality would lead to the same performance as this metric. On the other hand, some metrics can achieve about 70% *Correct Decision* rate.

Table 4 reports the classification errors for the ∆OM value that maximizes the *Correct Decision* frequency. Results show that PSNR computed in the linear domain, PSNR-DEx, PSNR-Lx, HDR-VDP-2, and avColorPSNR can achieve the highest *Correct Decision* rates. However, at the maximal *Correct Decision* value, avColorPSNR, PSNR-DEx, and PSNR-B have non-zero *False Ranking* errors, which is the most offensive error. PSNR-Lx has a faster decaying *False Ranking* error rate than PSNR-DEx, but PSNR-DEx makes less *False Ranking* errors at low ∆OM values, which means that PSNR-DEx is better than PSNR-Lx at detecting small differences.

**Table 4 – Classification errors for the ∆OM value that maximizes the *Correct Decision* frequency.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | | *Correct*  *Decision* (%) | | *False*  *Tie* (%) | *False*  *Differentiation* (%) | | *False*  *Ranking* (%) |
| **PSNR-R** | | **67.0** | | **30.7** | **2.3** | | **0.0** |
| **PSNR-G** | | **70.5** | | **23.3** | **6.2** | | **0.0** |
| **PSNR-B** | | **71.0** | | **6.8** | **21.6** | | **0.6** |
| mPSNR | | 55.1 | | 38.1 | 2.3 | | 4.5 |
| tPSNR-Yyuv | | 59.1 | | 33.5 | 2.8 | | 4.5 |
| tPSNR-U | | 60.8 | | 29.5 | 2.8 | | 6.8 |
| tPSNR-V | | 64.8 | | 22.2 | 9.1 | | 4.0 |
| tPSNR-YUV | | 59.7 | | 32.4 | 1.7 | | 6.2 |
| tPSNR-R | | 60.8 | | 33.5 | 1.7 | | 4.0 |
| tPSNR-G | | 60.2 | | 31.8 | 2.3 | | 5.7 |
| tPSNR-B | | 58.0 | | 34.1 | 1.7 | | 6.2 |
| tPSNR-RGB | | 59.1 | | 33.0 | 1.7 | | 6.2 |
| tPSNR-X | | 60.2 | | 35.2 | 1.1 | | 3.4 |
| tPSNR-Yxyz | | 60.2 | | 34.1 | 1.7 | | 4.0 |
| tPSNR-Z | | 57.4 | | 34.7 | 1.7 | | 6.2 |
| tPSNR-XYZ | | 58.5 | | 32.4 | 1.7 | | 7.4 |
| tPSNR-Yyupvp | | 60.2 | | 34.1 | 1.7 | | 4.0 |
| tPSNR-up | | 55.1 | | 42.0 | 0.0 | | 2.8 |
| tPSNR-vp | | 57.4 | | 25.0 | 8.0 | | 9.7 |
| tPSNR-Yupvp | | 55.7 | | 39.8 | 0.0 | | 4.5 |
| **PSNR-DE100** | | **67.0** | | **22.7** | **8.5** | | **1.7** |
| **PSNR-L100** | | **68.8** | | **29.5** | **1.7** | | **0.0** |
| **PSNR-DE1000** | | **70.5** | | **17.6** | **9.7** | | **2.3** |
| **PSNR-L1000** | | **68.8** | | **30.1** | **1.1** | | **0.0** |
| avLumaPSNR | | 61.9 | | 33.0 | 1.1 | | 4.0 |
| **avColorPSNR** | | **65.9** | | **21.0** | **10.8** | | **2.3** |
| avLumaErr | | 55.1 | | 43.2 | 0.0 | | 1.7 |
| avColorErr | | 55.7 | | 36.9 | 0.0 | | 7.4 |
| PQ2SSIM | | 60.8 | | 30.1 | 7.4 | | 1.7 |
| PQ2MS-SSIM | | 59.1 | | 30.1 | 8.0 | | 2.8 |
| PQ2VIFP | | 56.8 | | 42.0 | 1.1 | | 0.0 |
| **HDR-VDP-2** | | **69.9** | | **27.8** | **2.3** | | **0.0** |
| HDR-VQM | | 55.1 | | 43.8 | 0.0 | | 1.1 |
|  | |  | | |  | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
| **Figure 4 – Different/Similar (green curve), Better/Worse (blue curve), and Better/Equal-Worse (red curve) ROC analysis. The dashed lines indicate the TPR and FPR values for ∆OM=0.** | | | | | | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
| **Figure 4 – Different/Similar (green curve), Better/Worse (blue curve), and Better/Equal-Worse (red curve) ROC analysis. The dashed lines indicate the TPR and FPR values for ∆OM=0. *(Continued)*** | | | | | | | |
|  | |  | | |  | | |
|  | |  | | |  | | |
|  | |  | | |  | | |

**Figure 4 – Different/Similar (green curve), Better/Worse (blue curve), and Better/Equal-Worse (red curve) ROC analysis. The dashed lines indicate the TPR and FPR values for ∆OM=0. *(Continued)***

Figure 4 reports ROC analysis: Different/Similar (green curve), Better/Worse (blue curve), and Better/Equal-Worse (red curve). As it can be observed, the Different/Similar ROC curve in green is close to the diagonal for most metrics. Since there are 55% of the pairs where no statistically significant difference was observed in the subjective evaluation, this observation shows that most metrics cannot discriminate quality differences. On the other hand, some metrics such as PSNR-B show pretty good performance at determining which video in a pair has the best visual quality for pairs with significant visible differences. The three ROC curves can also be used to determine for which classification the metric is best at. For example, avColorErr has reasonable performance at determining when there is significant difference in a pair, but it has very poor performance at telling which video has the best visual quality when there are differences. However, PQ2MS-SSIM performs about the same in any case. The Better/Equal-Worse ROC analysis relies on both abilities to discriminate between significant and not significant visual quality differences in a pair of video sequences and to determine which video sequence in a pair has the best visual quality for pairs with significant visual quality differences. Therefore, the Better/Equal-Worse ROC curve typically lies between the Different/Similar and Better/Worse ROC curves.

Table 5 reports the AUC computed for the different ROC curves for each metric. Results are similar to those of classification errors, i.e., PSNR computed in the linear domain, HDR-VDP-2, PSNR-DEx, PSNR-Lx, and avColorPSNR are among the best metrics with one or more AUC value above 0.8. These metrics are better at determining which video sequence in a pair has the best visual quality for pairs with significant visual quality differences than at discriminating between significant and not significant visual quality differences in a pair, which also means that these metrics produce fewer *False Ranking* errors than *False Differentiation* or *False Tie* errors.

**Table 5 – AUC values for the different ROC analyses.**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Different/Similar | Better/Worse | Better/Equal-Worse |
| **PSNR-R** | **0.7297** | **0.9285** | **0.8271** |
| **PSNR-G** | **0.7035** | **0.9466** | **0.8404** |
| **PSNR-B** | **0.7723** | **0.9832** | **0.8995** |
| mPSNR | 0.7110 | 0.3203 | 0.3705 |
| tPSNR-Yyuv | 0.6362 | 0.3698 | 0.3853 |
| tPSNR-U | 0.6308 | 0.5376 | 0.5393 |
| tPSNR-V | 0.6512 | 0.7438 | 0.6787 |
| tPSNR-YUV | 0.7335 | 0.4595 | 0.4597 |
| tPSNR-R | 0.6845 | 0.5078 | 0.4921 |
| tPSNR-G | 0.6340 | 0.3780 | 0.4034 |
| tPSNR-B | 0.6661 | 0.5349 | 0.5261 |
| tPSNR-RGB | 0.7364 | 0.4762 | 0.4679 |
| tPSNR-X | 0.7325 | 0.4406 | 0.4408 |
| tPSNR-Yxyz | 0.6760 | 0.3857 | 0.3913 |
| tPSNR-Z | 0.6704 | 0.5587 | 0.5446 |
| tPSNR-XYZ | 0.7298 | 0.4653 | 0.4552 |
| tPSNR-Yyupvp | 0.6760 | 0.3857 | 0.3913 |
| tPSNR-up | 0.6535 | 0.4304 | 0.4530 |
| tPSNR-vp | 0.6556 | 0.4650 | 0.4813 |
| tPSNR-Yupvp | 0.6039 | 0.4046 | 0.4360 |
| **PSNR-DE100** | **0.7470** | **0.8973** | **0.8128** |
| **PSNR-L100** | **0.6278** | **0.8584** | **0.7539** |
| **PSNR-DE1000** | **0.7642** | **0.9104** | **0.8291** |
| **PSNR-L1000** | **0.6383** | **0.9119** | **0.7990** |
| avLumaPSNR | 0.6448 | 0.4033 | 0.4113 |
| **avColorPSNR** | **0.7394** | **0.8929** | **0.8164** |
| avLumaErr | 0.5645 | 0.4456 | 0.4676 |
| avColorErr | 0.7323 | 0.1008 | 0.1828 |
| PQ2SSIM | 0.6779 | 0.6815 | 0.6121 |
| PQ2MS-SSIM | 0.5941 | 0.6449 | 0.5934 |
| PQ2VIFP | 0.6410 | 0.3858 | 0.4042 |
| **HDR-VDP-2** | **0.6864** | **0.9338** | **0.8282** |
| HDR-VQM | 0.6161 | 0.2842 | 0.3499 |

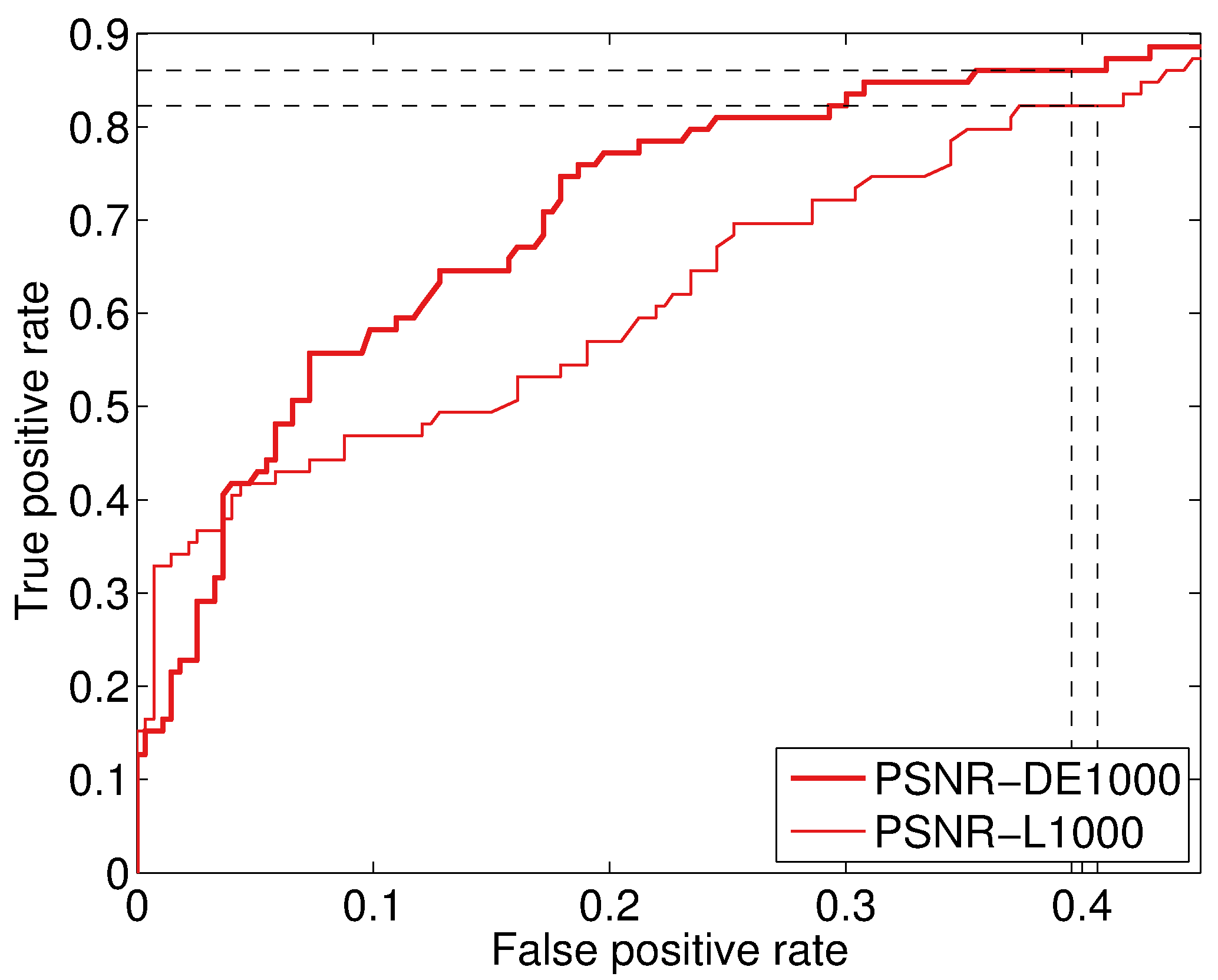
In the classical benchmarking based on MOS and CI results, PSNR-Lx performs better than PSNR-DEx, whereas it seems that the two metrics have quite similar performance, with a slight advantage for PSNR-DEx, based on the paired comparison data. To further investigate the performance of these metrics, we will focus our analysis on their performance at detecting small differences, i.e., for low ∆*OM* values. The first question is related to which reference luminance value leads better performance. Previous results, in particular Table 5, seem to indicate that the 1000 cd/m2 luminance reference leads better performance. Figure 5 shows a zoom of the Better/Equal-Worse ROC curve with results for both reference luminance levels plotted on the same graph for each metric (note that for the Better/Equal-Worse ROC analysis, TPR=FRP=0 for , whereas TPR=FRP=1 for ). For the same *True positive rate*, a lower *False positive rate* can be observed most of the time for the 1000 cd/m2 luminance reference, especially for PSNR-Lx. The difference is also more important at lower ∆*OM* values, which means that the metric can better predict small visual quality difference with the 1000 cd/m2 luminance reference level.

|  |  |
| --- | --- |
| PhD:Projects:MPEG HDR CfE:objective:PSNR-DE1000_cmp.png | PhD:Projects:MPEG HDR CfE:objective:PSNR-L1000_cmp.png |

**Figure 5 –Better/Equal-Worse ROC analysis for 100 cd/m2 and 1000 cd/m2 luminance reference. The dashed lines indicate the TPR and FPR values for ∆OM=0.**

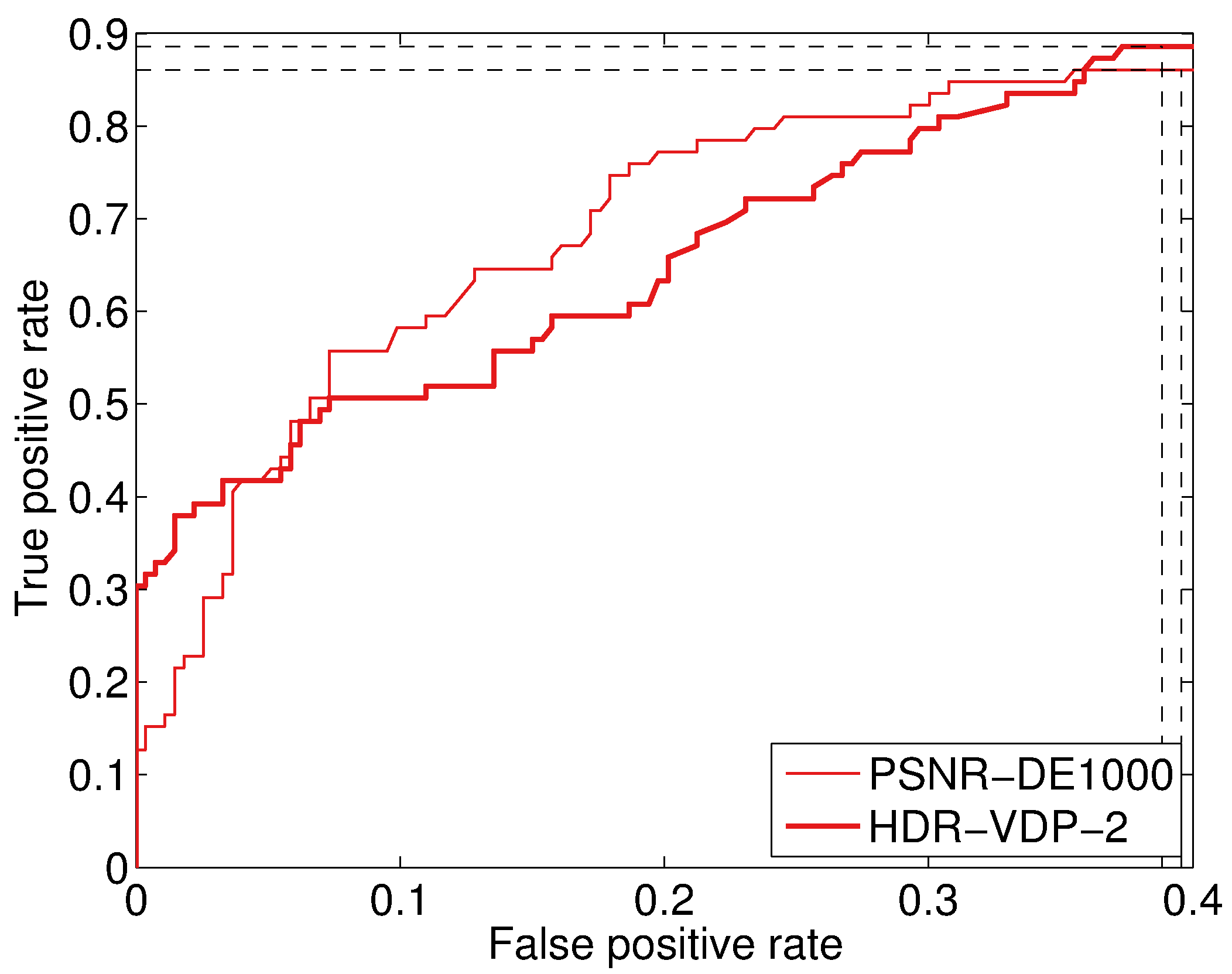
Figure 6 shows a zoom of the Better/Equal-Worse ROC curve with results for both PSNR-L1000 and PSNR-DE1000 plotted on the same graph. At low ∆*OM* values, a lower *False positive rate* can be observed for the same *True positive rate* for PSNR-DE1000. However, above a certain threshold (around *∆OM=*0.3), PSNR-L1000 shows better performance than PSNR-DE1000. We believe that for small visual quality differences, such as color artifacts, the PSNR-DE1000 metric performs better as it takes into account color differences. However, for large visual quality differences, the weighting between the lightness and color-opponent dimensions does not give enough importance to the lightness difference, as the PSNR-L1000 metric, which only considers the lightness component, performs better in this case. Thus, we believe that for small visual quality differences, e.g., color artifacts, the PSNR-DE1000 metric is better, whereas PSNR-L1000 is better when there are large visual quality differences. Therefore, for tuning the performance of an HDR video compression algorithm, PSNR-DE1000 would be better suited.

We believe that the performance difference for PSNR-DE1000 between the two datasets is due to the fact that color impairments were less visible on the Sim2 than on the Pulsar monitor because of the viewing angle dependency of the Sim2 monitor, as reported in [2]. We believe that these impairments were captured by PSNR-DE1000 but not perceived on the Sim2 monitor, which would explain the lower correlation and higher prediction error for PSNR-DEx when compared to PSNR-Lx reported in Table 3. However, when performing the side-by-side pair comparison on the Pulsar, the color impairments can be better discriminated by the subjects thanks to the higher resolving power of the test method and more accurate rendering of the Pulsar monitor.



**Figure 6 –Better/Equal-Worse ROC analysis for PSNR-L1000 (thin curves) and PSNR-DE1000 (thick curves). The dashed lines indicate the TPR and FPR values for ∆OM=0.**

The metrics showing the best performance across all three analyses are HDR-VDP-2 and PSNR computed in the linear domain. The problem with PSNR-R/G/B is that is takes only one color channel into account. An average of the three indexes could be used, but the performance of PSNR computed on the three components can be quite different. For example, PSNR-B has non-zero *False Ranking* error and higher *False Differentiation* error at the optimum ∆*OM* value. Moreover, the optimum ∆*OM* value is quite different between PSNR-R, PSNR-G, and PSNR-B. HDR-VDP-2 can achieve the best performance in all analyses, especially as it shows less content dependency and thus can also be used to predict the quality level of any content. However, this metric has a rather high computational complexity. A good alternative to HDR-VDP-2, with similar performance across the different analysis but higher content dependency, is PSNR-DE1000. Additionally, PSNR-DE1000 considers color differences, which are very important for WCG. Figure 7 shows a zoom of the Better/Equal-Worse ROC curve with results for both HDR-VDP-2 and PSNR-DE1000 plotted on the same graph. As it can be observed, HDR-VDP-2 can achieve the highest *True positive rate*, but there are cases where PSNR-DE1000 can achieve better performance. Thus, based on this extensive benchmarking and considering other studies on HDR quality assessment [3,12], we recommend using either HDR-VDP-2 or PSNR-DE1000 as quality metrics for HDR video quality assessment. Regarding other metrics considered, i.e., HDR-VQM and VIF, results show that these two metrics cannot determine reliably which video sequence in a pair has the best visual quality, even though they have good correlation with MOS. Therefore, we do not recommend using HDR-VQM or VIF.



**Figure 7 –Better/Equal-Worse ROC analysis for HDR-VDP-2 and PSNR-DE1000. The dashed lines indicate the TPR and FPR values for ∆OM=0.**

# Conclusion

This contribution reported the results of an extensive benchmarking of several objective metrics for HDR video quality assessment. Based on the results, we recommend using HDR-VDP-2, which shows the best performance across all analyses, or PSNR-DE1000, which shows similar performance for a lower computational complexity and considers color differences. We do not recommend using HDR-VQM or VIF, as these two metrics cannot determine reliably which video sequence in a pair has the best visual quality.

# Patent rights declaration(s)

**EPFL does not have any current or pending patent rights relating to the technology described in this contribution.**

**References**

1. ISO/IEC JTC1/SC29/WG11, “Results of HDR CfE subjective evaluations conducted at EPFL,” Doc. M36728, Warsaw, Poland, July 2015.
2. JCT-VC, " HDR CE3: Results of subjective evaluations conducted with the DSIS method," JCTVC-W0090, JCT-VC Meeting, San Diego, USA, February 2016.
3. ISO/IEC JTC1/SC29/WG11, “Correlation of subjective scores and objective metrics for HDR video quality assessment,” Doc. M 35273, Strasbourg, France, October 2014.
4. ISO/IEC JTC1/SC29/WG11, “Propose HDR Lightness and Color Error Metrics,” Doc. M35866, Geneva, Switzerland, February 2015.
5. ISO/IEC JTC1/SC29/WG11, “An Objective Metric for HDR,” Doc. M36464, Warsaw, Poland, June 2015.
6. 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,” ACM Transactions on Graphics, vol. 30, no. 4, July 2011.
7. M. Narwaria, M. Perreira Da Silva, and P. Le Callet, “HDR-VQM: An objective quality measure for high dynamic range video,” Signal Processing: Image Communication 35(0), 46–60 (2015).
8. S. Miller, M. Nezamabadi, and S. Daly, “Perceptual Signal Coding for More Efficient Usage of Bit Codes,” SMPTE Motion Imaging Journal, vol. 122, no 4, p. 52-59, May 2013.
9. J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, “High dynamic range texture compression for graphics hardware,” ACM Transactions on Graphics, vol. 25, no. 3, p. 698–706, July 2006.
10. ITU-T P.1401, Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models, International Telecommunication Union, July 2012.
11. ITU-T J.149, “Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM).” International Telecommunication Union, March 2004.
12. P. Hanhart, M. Bernardo, M. Pereira, A. Pinheiro, and T. Ebrahimi, “Benchmarking of objective quality metrics for HDR image quality assessment,” in EURASIP Journal on Image and Video Processing, 2015.

1. MeTriX MuX v1.1: <http://foulard.ece.cornell.edu/gaubatz/metrix_mux/> [↑](#footnote-ref-1)
2. HDR-VDP-2 v2.2.1: <http://hdrvdp.sourceforge.net/> [↑](#footnote-ref-2)
3. HDR-VQM v1: <http://sites.google.com/site/narwariam/hdr-vqm/> [↑](#footnote-ref-3)