https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-307
© 2019, Society for Imaging Science and Technology

Combining Quality Metrics using Machine Learning for improved and robust HDR Image Quality Assessment

Anustup Choudhury, Scott Daly; Dolby Laboratories Inc.; Sunnyvale, CA, USA

Abstract

We improve High Dynamic Range (HDR) Image Quality Assessment (IQA) using a full reference approach that combines results from various quality metrics (HDR-CQM). We combine metrics designed for different applications such as HDR, SDR and color difference measures in a single unifying framework using simple linear regression techniques and other non-linear machine learning (ML) based approaches. We find that using a non-linear combination of scores from different quality metrics using support vector machine is better at prediction than the other techniques such as random forest, random trees, multilayer perceptron or a radial basis function network. To improve performance and reduce complexity of the proposed approach, we use the Sequential Floating Selection technique to select a subset of metrics from a list of quality metrics. We evaluate the performance on two publicly available calibrated databases with different types of distortion and demonstrate improved performance using HDR-CQM as compared to several existing IQA metrics. We also show the generality and robustness of our approach using cross-database evaluation.

Introduction

High dynamic range (HDR) and wide color gamut (WCG) capability have now become mainstream in consumer TV displays and is making headway into desktop monitors, laptops and mobile device products. In the consumer industry, the term HDR generally means the combination of HDR and WGC, and we will use that shorthand terminology here. Since HDR systems provide a more complete representation of information that the human visual system can perceive, which makes evaluating content shown on these HDR systems is essential. Since subjective evaluations can be time-consuming and expensive, there is a need for objective quality assessment tools. Various full reference HDR quality metrics such as HDR-VDP-2 (HDR visual difference predictor) [1, 2], DRIM (Dynamic range independent metric) [3], HDR-VQM (HDR video quality measure) [4] have been proposed for image and video quality assessment (IQA/VQA). HDR-VDP2 and HDR-VQM require modeling of both the human visual system (HVS) and the display, whereas DRIM, in addition to HVS modeling, results in three distortion output maps making it more difficult for interpretation. Alternatively, due to lack of HDR objective metrics, LDR/SDR (low/standard dynamic range) metrics were also used to evaluate HDR quality. Examples of full reference LDR metrics that have been used in literature for HDR quality evaluation are MS-SSIM (Multi-scale structural similarity index) [5], IFC (Information fidelity criterion) [6], VIFp (pixel-based visual information fidelity) [7], FSIM (Feature similarity index) [8], VIF (visual information fidelity) [7] and so on.

Recent studies [9, 10, 11, 12], have evaluated both HDR and LDR quality metrics for HDR quality assessment. In particular, [11] evaluated the performance of 35 objective metrics on a publicly available database. [12] evaluated the performance of 12 metrics on five different HDR databases. Although the HDR based metrics, HDR-VDP-2 and HDR-VQM outperform existing LDR metrics, modifying some LDR metrics such as MS-SSIM by applying calibrated non-linearities can result in performance close to the HDR based metrics in terms of correlation [11].

Since visual content and its corresponding distortion have varying degrees of diversification, it is quite challenging to rely on a single metric. Many of the SDR metrics are applied to the video signals in the code value domain and do not consider the effects of the display they are being viewed. HDR metrics generally consider the specific code-value-to-luminance relationship of the display. In our previous work [13], we proposed an HDR IQA technique that combined various HDR and LDR quality metrics. This was inspired by the Video Multi-method Assessment Fusion (VMAF) approach [14, 15], which in turn is based on [16, 17]. In this work, we combine various metrics designed for different applications (HDR, LDR/SDR and color difference measures) in one framework for improved performance of HDR IQA. We use a greedy approach based on Sequential Forward Selection (SFS) [18] to select metrics to be included in the final model. Next, we combine these different quality metrics using machine learning (ML) approaches. To find the best technique for combining these metrics, apart from linear regression, we also evaluate Support Vector Machine (SVM) regression [19, 20], Random Trees (RT) regression [21], Random Forests (RF) regression [22], Multilayer perceptron [23] (MLP) regression, and Radial Basis Function (RBF) [24, 25] network regression. The resulting model (HDR-CQM) can be used to predict overall quality for HDR images. We confirm our choice of the ML technique being used by assessing its performance on two databases with different distortions. We also perform cross-database evaluation to show the generality and robustness of our proposed metric. To the best of our knowledge, this is the first approach that aims at combining various metrics, including HDR, LDR and color difference metrics, in a single unifying framework for HDR IQA in an efficient manner using machine learning techniques.

Combining Quality Metrics for HDR IQA (HDR-CQM)

A brief overview of our method to combine various quality metrics for HDR IQA is shown in Figure 1. As seen in Figure 1, the proposed metric has a training and testing component. During training, we collect pairs of reference and corresponding distorted HDR images. We first compare the reference and distorted images using various IQA metrics. We then use the SFS method for selecting a subset of IQA metrics in an efficient manner, whose

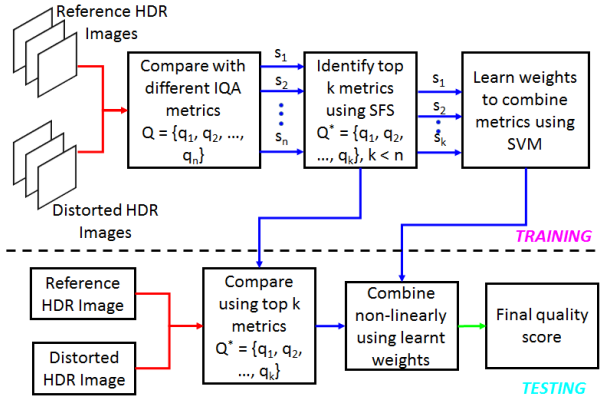


Figure 1. *Block diagram of proposed HDR-CQM metric*

combination gives the best performance, from the pool of various IQA metrics. Finally we combine the scores of the selected metrics in a non-linear manner using support vector machine since that gives the best result.

Let's assume that we want to combine the scores of k IQA metrics and we have T training images. The quality scores of the t-th training image using the various metrics can be denoted as $\mathbf{x}_t = \{x_{t1}, \ldots, x_{tk}\}$ where $t = \{1, 2, \ldots, T\}$ are the image indices. The final quality score can be denoted as $q(\mathbf{x}_t)$, which can either be a linear combination of individual scores if we use LR, or a non-linear combination of k scores if we use ML techniques. During training, we would like to determine the weights in case of (say) LR or SVM, or the decision boundaries in case of RT or RF, such that it minimizes the difference between $q(\mathbf{x}_t)$ and the Mean Opinion Score (MOS) score (obtained from subjective studies) and can be represented as

$$\text{argmin} \, ||q(\mathbf{x}_t) - MOS_t||, \tag{1}$$

where $t = \{1, 2, \ldots, T\}$ are the training image indices, $||.||$ denotes a certain norm such as Euclidean norm or l_1 norm and MOS_t is the mean opinion score for image t. The difference metric will depend on the type of ML techniques being used. Please note that the training is an offline process.

During testing, we take as input a reference image and a distorted image. We then compare those two images using the pool of k IQA metrics identified by the SFS method during training. Then, we combine the scores from each of the k metrics using weights derived during training and combine them to get the final quality score.

IQA metrics

Various IQA metrics have been proposed in literature to evaluate human visual quality experience. Recent studies [11, 12] performed extensive analysis of objective quality metrics for HDR IQA. Different metrics were used for evaluation in both work [11, 12]. In addition to the metrics that were presented in our previous work [13] viz., HDR-VDP-2 [1, 2], HDR-VQM [4], MS-SSIM [5], IFC [6], VIFp [7], FSIM [8] and FSITM [26], we considered three additional metrics – UQI [27] and two color difference measures – CIE ΔE2000 [28] and ΔIC_TC_P [29].

HDR-VDP-2 and HDR-VQM were developed for HDR quality assessment. HDR-VDP-2 is a calibrated metric and takes into account models regarding point spread function of the eye, the light-adaptive CSF, and masking within an oriented multi-scale decomposition. HDR-VQM is a video quality metric computed in PU [30] space and also relies on multi-scale and multi-orientation analysis, as well as simple temporal differences which are pooled. In this setup, we compute HDR-VQM on still images, thus having zero temporal error. Both HDR-VDP-2 and HDR-VQM perform spatial pooling to compute overall quality score.

We considered six metrics that were all designed for LDR content. MS-SSIM is a multi-scale technique that models the quality based on the assumption that the HVS uses structural information from a scene. VIFp analyzes the natural scene statistics and is based on Shannon information. FSIM analyzes high phase congruency, extracting the gradient magnitude to encode contrast information. IFC uses natural scene statistics to model natural scenes and its distortion to quantify statistical information shared between reference and distorted images, and uses that information about fidelity as a measure of quality. FSITM compares the locally weighted mean phase angle map of the reference image to that of its associated distorted image. UQI does not use any HVS modeling but models the image distortion as a combination of loss of correlation, luminance and contrast distortions.

Finally, we use a couple of color difference metrics. CIE ΔE2000 is a color difference measure that includes weighting factors for lightness, chroma and hue along with the ability to handle the relationship between chroma and hue. It was designed for the CIELAB color space, which is limited to SDR. ΔIC_TC_P is a newer metric based on the IC_TC_P [31] color representation, which was designed to be more perceptually uniform with HDR signals.

LDR metrics are designed for gamma encoded images with small range of luminance values whereas the HDR images in these datasets have linear values to account for wider luminance ranges. Since the databases that we considered are comprised of HDR images in the linear range, we applied HDR-VDP-2 and HDR-VQM directly since these are calibrated metrics and require absolute luminance. The LDR metrics were computed in the PQ domain [32]. For the LDR metrics, we first convert the gamma domain code values to linear luminance and then convert luminance to the PQ domain and denote that using _PQ suffix. This is based on the results from [11] that found calibrating the LDR metrics via either the PQ or PU non-linearities always improved their performance compared to applying them directly on the code values of the signal space. Further, such processing focuses the quality on the achromatic channel of human vision, which is known to have the better spatial performance. Any purely color differences (i.e., iso-luminant) are ignored in the LDR analysis due to the models limitations, since there may be some chromatic distortion if 422 and 420 profiles were used, or in the tone-mapping distortions. Then we normalize the RGB color components to [0, 1] range and transform the RGB color space to YC_bC_r color space. The quality score was computed on the luminance (Y) channel since [11] found that using the Y channel alone instead of using the mean of the Y, C_b and C_r color channels resulted in the best performance. We thus consider only the Y channel for the LDR metrics and denote that using _Y suffix. The color difference measures were not computed in the transformed spaces. CIE ΔE2000 require a conversion from RGB to CIELAB color space considering a D65 100nits reference white point whereas ΔIC_TC_P require a conversion from RGB to IC_TC_P color space. We assume that

the databases are either in sRGB or BT. R. 709 formats, both of which use the same primaries. They use a normalized luminance, but a calibrated chromaticity. Please note that the normalization was not applied to HDR and color difference metrics.

A list of all IQA metrics considered for HDR-CQM are shown in Table 1 along with the average computational cost of those metrics for one 1920 X 1080 image using Matlab on a computer with Intel Xeon processor with 16GB RAM.

Table 1: List of IQA metrics with their computational cost

Index	IQA Metrics	Time (sec./image)
q1	HDR-VDP-2	22
q2	HDR-VQM	5.8
q3	MS-SSIM_PQ_Y	0.8
q4	IFC_PQ_Y	8.6
q5	VIFp_PQ_Y	0.8
q6	FSIM_PQ_Y	0.8
q7	FSITM_PQ_Y	3.6
q8	UQI_PQ_Y	0.6
q9	CIE ΔE2000	2.8
q10	ΔIC_TC_P	1.4

Databases

We consider two different publicly available databases to compare the performance of the different metrics. The first database [33] (referred to as Database 1) contains 20 HDR images with a resolution of 1920 X 1080 pixels. The images are adjusted for a SIM2 HDR monitor and compressed with JPEG XT with various profiles and quality levels. 240 compressed HDR images were created using two different tone mapping operations [34, 35] for the base layer, four different bit rates were chosen for each original image using 3 profiles of JPEG XT. The images were presented in a side-by-side manner, one of which was reference and the other a distorted version, and the subjective scores were collected from 24 participants.

The second database [12] that we considered is a combination of two different databases proposed in [36] and [12]. [36] is composed of 5 original HDR images which were first tone-mapped using [37], following which 50 compressed images were obtained using three different coding schemes – JPEG, JPEG2000 and JPEG XT. These images were presented one after the other on a SIM2 HDR47E display and scores were collected from 15 participants. [12] use a similar experimental paradigm as [36] and once again use 5 original HDR images from which 50 compressed images were obtained. They used JPEG and JPEG2000 (with different bit rates) and the LDR images were obtained using two different mapping operations [37, 32]. Thus our second database (referred to as Database 2) has 100 images.

Metric Selection

We consider 10 different quality metrics shown in Table 1 and to reduce complexity, we need to find which combination of these metrics is best for HDR-CQM. Including a poor IQA metric may negatively affect the performance of HDR-CQM. In such cases, we would not like to include that metric in the final model. Similar to [16, 17], we use a greedy approach based on SFS [18]. SFS starts with an empty target set and repeatedly adds the most significant metric. It has better performance than branch-and-bound techniques and lower cost than exhaustive approaches.

Algorithm 1 Sequential Forward Selection (SFS)

Input: $Q = \{q_1, q_2 \ldots, q_n\}$ ▷ Set of available metrics
Output: $Q^* = \{q_1, q_2 \ldots, q_k\}, k < n$ ▷ Set of selected metrics

1: **procedure** SFS
2: Start with the empty set $Q^* = \{\}$
3: Select the best metric

$$q^+ = \underset{q \in Q - Q_k^*}{\operatorname{argmax}} O(Q_k^* + q)$$

4: $Q_{k+1}^* \leftarrow Q_k^* + q^+$
5: $k \leftarrow k + 1$
6: **end procedure**

Given a set of quality metrics $Q = \{q_1, q_2 \ldots, q_n\}$, we would like to determine a subset $Q^* = \{q_1, q_2 \ldots, q_k\}$ where $k \leq n$ to maximize an objective function $O(Q^*)$ and can be represented as

$$\operatorname{argmax} O(Q^*) = \operatorname{argmax} PLCC(q(Q^*), MOS), \quad (2)$$

where PLCC is the Pearson linear correlation coefficient. Details are summarized in Algorithm 1.

The metrics selected using SFS are shown in Tables 2 and 3. The best individual performing metric is listed in the first row, and the best performing n metrics are listed in the following rows that incrementally increase the number of combined (fused) metrics. Metrics that contribute negatively are italicized. Please refer to Table 1 to find correspondence between index and IQA metric. SVM regression is used to combine the metrics. We use the databases presented in [33] and [12] for evaluation and use 10-fold cross-validation.

Table 2: Performance measure of combined metrics on Database 1 [33] using SFS. Best result is shown in bold.

#	Indices of selected combined FR-IQA metrics	PLCC
1	q2	0.9582
2	q2, q1	0.9662
3	q2, q1, q7	0.9692
4	q2, q1, q7, q4	0.9723
5	q2, q1, q7, q4, q3	0.9727
6	**q2, q1, q7, q4, q3, q6**	**0.9729**
7	q2, q1, q7, q4, q3, q6, *q5*	0.9713
8	q2, q1, q7, q4, q3, q6, *q5*, *q10*	0.9723
9	q2, q1, q7, q4, q3, q6, *q5*, *q10*, *q8*	0.9725
10	q2, q1, q7, q4, q3, q6, *q5*, *q10*, *q8*, *q9*	0.9702

From Tables 2 and 3, we can see that using a combination of as few as two metrics, we can achieve better performance than using a single metric. Adding more metrics results in further improvement. An interesting observation is that if we use all 10 metrics, then the performance drops. This could be because of increased data requirement for better estimates and increased sources of error when more metrics are present.

Exploring learning algorithms

In this section, we compare the performance of different ML approaches to learn the relationship between the models and the

Table 3: Performance measure of combined metrics on Database 2 [12, 36] using SFS. Best result is shown in bold.

#	Indices of selected combined FR-IQA metrics	PLCC
1	q1	0.9360
2	q1, q7	0.9480
3	q1, q7, q5	0.9581
4	q1, q7, q5, q3	0.9606
5	q1, q7, q5, q3, q4	0.9615
6	**q1, q7, q5, q3, q4, q8**	**0.9623**
7	q1, q7, q5, q3, q4, q8, *q10*	0.9615
8	q1, q7, q5, q3, q4, q8, *q10*, *q9*	0.9611
9	q1, q7, q5, q3, q4, q8, *q10*, *q9*, *q2*	0.9547
10	q1, q7, q5, q3, q4, q8, *q10*, *q9*, *q2*, *q6*	0.9489

Table 4: Quantitative comparison of ML techniques on Database 1 [33]. Best result is shown in bold.

	PLCC	SROCC	RMSE
LR	0.9663	0.9632	0.3197
RF	0.9640	0.9618	0.33
RT	0.9253	0.9152	0.4706
SVM	**0.9702**	**0.9678**	**0.3006**
MLP	0.9684	0.9652	0.3097
RBF	0.9686	0.9663	0.3089

subjective scores (MOS). We compared linear regression (LR) with machine learning (ML) techniques such as Support Vector Machine (SVM) regression [19, 20], Random Trees (RT) regression [21], Random Forests (RF) regression [22], Multilayer perceptron [23] (MLP) regression, and Radial Basis Function (RBF) [24, 25] network regression. For the SVM regression, we used ν-SVM regression with a RBF kernel. Its kernel parameter was set to the inverse of the number of metrics. RT use recursive partitioning to split the data into different segments. RF uses an ensemble of decision trees from randomly sampled sub-spaces of the input features and the final results are obtained by combining results from different trees via voting. MLP is a feed-forward neural network and we use three layers. RBF network is a three layer network with one hidden layer that has a non-linear RBF activation function.

To benchmark the performance of various ML techniques, we combine the results using all 10 IQA metrics shown in Table 1. We randomly divide each database into disjoint 40% training set and 60% testing set. The best parameters of ML techniques were chosen on the basis of 10-fold cross validation. The performance of the different learning methods is shown in Tables 4 and 5, and we can see that SVM regression outperforms other techniques, similar to another study involving a different class of HDR distortions (color, bit-depth, local contrast and tonescale) generally being lower in frequency than compression distortions [38]. Thus, we use SVM regression for the combination of various metrics.

Experimental Results and Discussion

To evaluate the performance of various metrics, we compare the MOS obtained from the subjective study with the MOS values predicted from the different metrics. We used the technique described in [39] to fit a monotonic logistic function to fit the objective prediction to the subjective scores as follows -

$$f = \alpha + \frac{\beta}{1 + e^{-\gamma.(x-\delta)}}, \tag{3}$$

where f is the fitted objective score, x is the predicted score using different techniques and $\alpha, \beta, \gamma, \delta$ are the parameters that define the shape of the logistic fitting function and are determined by minimizing the least squares error between the subjective and the fitted objective scores.

Table 5: Quantitative comparison of ML techniques on Database 2 [12, 36]. Best result is shown in bold.

	PLCC	SROCC	RMSE
LR	0.9459	0.9354	9.3202
RF	0.9455	0.9378	9.3365
RT	0.9214	0.8987	11.1732
SVM	**0.9489**	**0.9497**	**9.0543**
MLP	0.9348	0.9366	10.2511
RBF	0.9417	0.9359	9.6492

To quantify the performance, we use Root mean square error (RMSE), Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SROCC) [39]. RMSE is used for measuring prediction consistency, PLCC for prediction accuracy and SROCC for prediction monotonicity respectively. Lower values of RMSE indicates better performance and higher values of PLCC and SROCC imply better accuracy and prediction monotonicity.

We compare the individual performance of several state-of-the-art IQA metrics with a few variations of HDR-CQM in Tables 6 and 7. From Table 6, we can see that while HDR-VQM (q2) performs the best with database 1, it degrades the metric performance when used in database 2 (Table 3). On the other hand, q1 (HDR-VDP-2) has the best performance with the database 2, and is the 2nd most important contributing metric in the first data base (Table 2). Thus, it is the best overall contributor. Both metrics are the most computationally expensive, with HDR-VDP-2 being about 4x that of HDR-VQM. HDR-VQM is intended as a temporal video metric, while all the others including HDR-VDP-2 were intended to be used with still images. That aspect may help explain the peculiar behavior of HDR-VQM, whereas the high computational cost of HDR-VDP-2 helps its performance, and the video capability of HDR-VQM may cause it to perform negatively with the second database.

Amongst LDR metrics, MS-SSIM calibrated in PQ domain performs the best for database 1 and VIFp in PQ domain works the best for database 2. In general, HDR metrics perform better than LDR metrics. This is because their calibration and HVS front-end non-linearities evenly distribute the perception of distortions across the image's full tone-scale. While applying such front-end non-linearities to the LDR metrics does improve their performance (PLCC of the best performing LDR metric on database 1 (MS-SSIM) increases from 0.8635 to 0.9323 and PLCC of the best performing LDR metric on database 1 (VIFp) increases from 0.7213 to 0.9231), the HDR metrics are in general more advanced, such as having orientation channels.

Also, ΔIC_TC_P is better than CIE $\Delta E2000$, for both databases. This is expected because the HDR achromatic non-linearity of ΔIC_TC_P (i.e., PQ) is known to better match visibility for the HDR

Table 6: Performance comparison of FR-IQA metrics on Database 1. Best method is highlighted in bold, the 2^{nd} best method is italicized and the 3^{rd} best method is underlined.

Method	PLCC	SROCC	RMSE
HDR-VDP-2	0.9559	0.9552	0.3648
HDR-VQM	0.9596	0.9594	0.3490
MS-SSIM_PQ_Y	0.9323	0.9264	0.4488
FSITM_PQ_Y	0.9161	0.9147	0.5
VIFp_PQ_Y	0.9251	0.9226	0.4730
FSIM_PQ_Y	0.9173	0.9164	0.4941
IFC_PQ_Y	0.9010	0.8963	0.5461
UQI_PQ_Y	0.8608	0.8536	0.6550
CIE ΔE2000	0.7806	0.7717	0.7777
ΔIC_TC_P	0.8166	0.8244	0.7172
MS-SSIM_Y	0.8635	0.8624	0.6260
HDR-CQM (2 Metrics)	0.9666	0.9645	0.3182
HDR-CQM (4 Metrics)	*0.9726*	*0.9703*	*0.2888*
HDR-CQM (6 Metrics)	**0.9730**	**0.9708**	**0.2866**

Table 7: Performance comparison of FR-IQA metrics on Database 2. Best method is highlighted in bold, the 2^{nd} best method is italicized and the 3^{rd} best method is underlined.

Method	PLCC	SROCC	RMSE
HDR-VDP-2	0.9360	0.9305	10.1266
HDR-VQM	0.9212	0.9073	11.1652
MS-SSIM_PQ_Y	0.8885	0.8769	13.1792
FSITM_PQ_Y	0.7640	0.7489	19.4431
VIFp_PQ_Y	0.9231	0.9088	11.0306
FSIM_PQ_Y	0.85	0.8355	15.8197
IFC_PQ_Y	0.8297	0.7987	16.1657
UQI_PQ_Y	0.7645	0.7542	18.5437
CIE ΔE2000	0.5938	0.7717	22.1409
ΔIC_TC_P	0.6881	0.7036	21.0638
VIFp_Y	0.7213	0.7543	21.0401
HDR-CQM (2 Metrics)	0.9555	0.9568	8.5374
HDR-CQM (4 Metrics)	*0.9648*	*0.9625*	*7.5707*
HDR-CQM (6 Metrics)	**0.9653**	**0.9655**	**7.4883**

luminance range, and in particular where the L* achromatic nonlinearity of CIELAB is known to fail for luminance less than 1nits. There are only two metrics that contribute negatively to both databases, CIE ΔE2000 and ΔIC_TC_P respectively, which are color models with substantiated accuracy in certain applications. Both are solely pixel-wise comparisons, utilizing no spatial processing. Their overall poor performance shows that for the distortions tested in the two data sets, that the spatial and luminance aspects (that is the achromatic performance), dominates over any color or advantages that the color models may have. This is likely due to the types of distortions in the two databases, which did not explicitly probe common chromatic distortions such as color saturation & desaturation, or hue shifts.

Please note that our reported numbers are slightly different from [11] because we show results on randomly divided sample of the database containing 60% of the images. Also note that the results reported in Tables 6 and 7 are slightly different than the results reported in Tables 2 and 3 because we fit the logistic function, Equation 3 to the objective scores resulting in better fit to the MOS. We can see that our combination of quality metrics (HDR-CQM) has better performance than the individual metrics. The top three ranked metrics in Tables 6 and 7 are variations of the proposed method.

To test the generality of the proposed approach, we perform cross-database evaluation. We use all images from one database for training and test on the images from the other database. We choose four IQA metrics (best performing ones for the training database) for combination and the results are summarized in Table 8. We observe that the PLCC is high for both cases (also higher than any individual metric) which verifies the generality and robustness of the proposed combined metrics.

Table 8: Cross-Database PLCC evaluation (in terms of whole database) using four metrics

	Database 1	Database 2
Database 1	-	0.9417
Database 2	0.9675	-

Conclusion & Future Work

In this paper, we introduce a new HDR quality measure (HDR-CQM) that is one of the first attempts towards combining different HDR, LDR and color difference IQA metrics to improve the prediction of HDR image quality. In order to reduce complexity and to identify which metrics to combine, we use a greedy method based on SFS. We find that naively combining various metrics might lead to worse results and combining the right number of metrics is important. In order to combine the metrics, we find that SVM regression is better at prediction than linear regression and other ML techniques such as multi-layer perceptron, random forest, random trees and RBF network regressor. We use two different databases for evaluation and show that the proposed metric outperforms state-of-the-art quality metrics by a significant margin. We also perform the tests across databases to show the generality and robustness of the proposed metric.

For future work, we would like to further test the performance on more databases. We would like to explore the performance of other FR-IQA metrics and other color spaces in addition to YC_bC_r. We would also like to explore how well this technique can scale across other distortions.

References

[1] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W., "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.* **30**, 40:1–40:14 (July 2011).

[2] Narwaria, M., Mantiuk, R., Silva, M. P. D., and Callet, P. L., "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging* **24**, 24 – 24 – 3 (2015).

[3] Aydin, T., Mantiuk, R., Myszkowski, K., and Seidel, H. P., "Dynamic range independent image quality assessment," *ACM Trans. Graph.* **27**, 69:1–69:10 (Aug. 2008).

[4] Narwaria, M., Silva, M. P. D., and Callet, P. L., "HDR-VQM: An Objective Quality Measure for High Dynamic Range Video," *Signal Processing: Image Communication* **35**, 46–60 (July 2015).

[5] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multiscale structural similarity for image quality assessment," in [*37th*

Asilomar Conference on Signals, Systems and Computers], **2**, 1398–1402, IEEE (Nov. 2003).

[6] Sheikh, H. R., Bovik, A. C., and de Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing* **14**, 2117–2128 (Dec 2005).

[7] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE TIP* **15**, 430–444 (Feb 2006).

[8] Zhang, L., Zhang, L., Mou, X., and Zhang, D., "FSIM: A feature similarity index for image quality assessment," *IEEE TIP* **20**, 2378–2386 (Aug 2011).

[9] Azimi, M., Banitalebi-Dehkordi, A., Dong., Y., Pourazad, M., and Nasiopoulos, P., "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content," in [*International Conference on Multimedia Signal Processing*], (November 2014).

[10] Hanhart, P., Rerbek, M., and Ebrahimi, T., "Subjective and objective evaluation of hdr video coding technologies," in [*QoMEX*], 1–6 (June 2016).

[11] Hanhart, P., Bernardo, M., Pereira, M., Pinheiro, A. M. G., and Ebrahimi, T., "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP Journal on Image and Video Processing* **2015**, 39 (Dec 2015).

[12] Zerman, E., Valenzise, G., and Dufaux, F., "An extensive performance evaluation of full-reference HDR image quality metrics," *Quality and User Experience* **2**, 5 (Apr 2017).

[13] Choudhury, A. and Daly, S., "Hdr image quality assessment using machine-learning based combination of quality metrics," in [*IEEE GlobalSIP*], (November 2018).

[14] Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., and Manohara, M., "Toward a practical perceptual video quality metric," (2016).

[15] Li, Z., Norkin, A., and Aaron, A., "VMAF - video quality metric alternative to PSNR," *Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11* (October 2016).

[16] Liu, T. J., Lin, W., and Kuo, C. C. J., "Image quality assessment using multi-method fusion," *IEEE Transactions on Image Processing* **22**, 1793–1807 (May 2013).

[17] Lin, J. Y., Liu, T. J., Wu, E. C. H., and Kuo, C. C. J., "A fusion-based video quality assessment (fvqa) index," in [*Signal and Information Processing Association Annual Summit and Conference (APSIPA)*], 1–5 (Dec 2014).

[18] Whitney, A. W., "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers* **C-20**, 1100–1103 (Sept 1971).

[19] Cortes, C. and Vapnik, V., "Support-vector networks," *Machine Learning* **20**, 273–297 (Sept. 1995).

[20] Basak, D., Pal, S., and Patranabis, D., "Support vector regression," in [*Neural Information Processing Letters and Reviews*], 203–224 (October 2007).

[21] Breiman, L., Friedman, J., Olshen, R., and Stone, C., [*Classification and Regression Trees*], Wadsworth and Brooks, Monterey, CA (1984).

[22] Breiman, L., "Random forests," *Machine Learning* **45**, 5–32 (Oct 2001).

[23] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. in [*Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*], ch. Learning Internal Representations by Error Propagation, 318–362, MIT Press, Cambridge, MA, USA (1986).

[24] Schwenker, F., Kestler, H. A., and Palm, G., "Three learning phases for radial-basis-function networks," *Neural Networks* **14**(4), 439 – 458 (2001).

[25] Broomhead, D. and Lowe, D., "Multivariable functional interpolation and adaptive networks," *Complex Systems* **2**, 321–355 (1988).

[26] Nafchi, H. Z., Shahkolaei, A., Moghaddam, R. F., and Cheriet, M., "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Processing Letters* **22**, 1026–1029 (Aug 2015).

[27] Wang, Z. and Bovik, A. C., "A universal image quality index," *IEEE Signal Processing Letters* **9**, 81–84 (March 2002).

[28] Luo, M. R., Cui, G., and Rigg, B., "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application* **26**(5), 340–350 (2001).

[29] Pieri, E. and Pytlarz, J., "Hitting the mark - a new color difference metric for hdr and wcg imagery," in [*SMPTE 2017 Annual Technical Conference and Exhibition*], 1–13 (Oct 2017).

[30] T. Aydin, R. Mantiuk, H. S., "Extending quality metrics to full luminance range images," (2008).

[31] "BT2100: Image parameter values for high dynamic range television for use in production and international programme exchange," *International Telecommunication Union* (July 2016).

[32] Miller, S., Nezamabadi, M., and Daly, S., "Perceptual signal coding for more efficient usage of bit codes," in [*The 2012 Annual Technical Conference Exhibition*], 1–9 (Oct 2012).

[33] Korshunov, P., Hanhart, P., Richter, T., Artusi, A., Mantiuk, R., and Ebrahimi, T., "Subjective quality assessment database of HDR images compressed with jpeg xt," in [*QoMEX*], 1–6 (May 2015).

[34] Mantiuk, R., Myszkowski, K., and Seidel, H.-P., "A perceptual framework for contrast processing of high dynamic range images," *ACM Trans. Appl. Percept.* **3**, 286–308 (July 2006).

[35] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J., "Photographic tone reproduction for digital images," *ACM Trans. Graph.* **21**, 267–276 (July 2002).

[36] Valenzise, G., Simone, F. D., Lauga, P., and Dufaux, F., "Performance evaluation of objective quality metrics for hdr image compression," in [*SPIE optical engineering + applications, International Society for Optics and Photonics*], (2014).

[37] Mai, Z., Mansour, H., Mantiuk, R., Nasiopoulos, P., Ward, R., and Heidrich, W., "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Transactions on Image Processing* **20**, 1558–1571 (June 2011).

[38] Choudhury, A., Farrell, S., Atkins, R., and Daly, S., "Prediction of HDR quality by combining perceptually transformed display measurements with machine learning," in [*Proc. SPIE*], **10396**, 10396 – 10396 – 16 (2017).

[39] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," (2003).

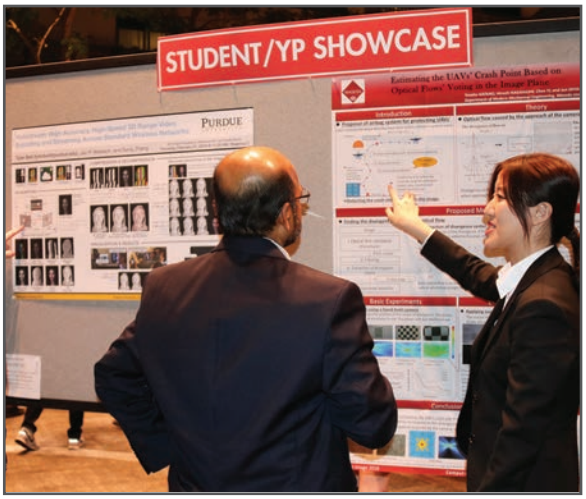
JOIN US AT THE NEXT EI!

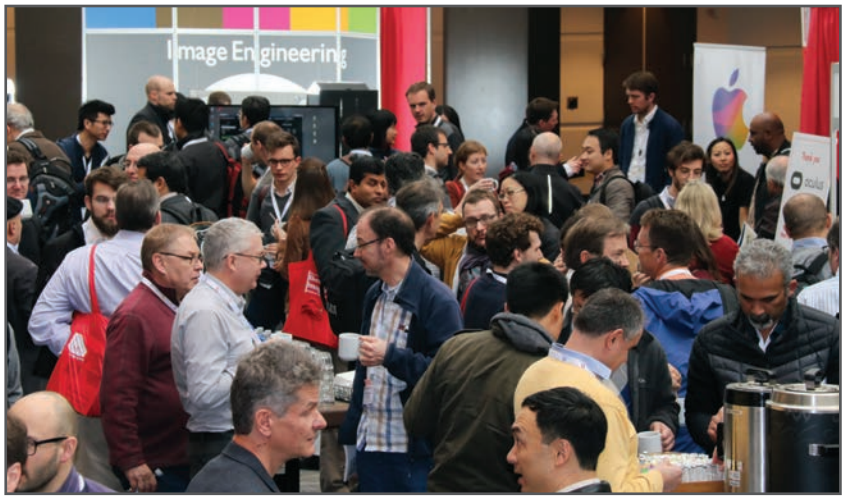
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org