XPSNR: A LOW-COMPLEXITY EXTENSION OF THE PERCEPTUALLY WEIGHTED PEAK SIGNAL-TO-NOISE RATIO FOR HIGH-RESOLUTION VIDEO QUALITY ASSESSMENT

Christian R. Helmrich[1], *Mischa Siekmann*[1], *Sören Becker*[1], *Sebastian Bosse*[1], *Detlev Marpe*[1], *and Thomas Wiegand*[1, 2]

[1] Video Coding and Analytics Department, Fraunhofer Heinrich Hertz Institute (HHI), Berlin, Germany
[2] Image Communications Group, Technical University of Berlin, Germany

ABSTRACT

The objective PSNR metric is known to correlate quite poorly with subjective assessments of video coding quality. Thus, a number of alternative VQA measures such as (MS-)SSIM and VMAF have been proposed. These, however, are often algorithmically complex and difficult to use for visually motivated encoder optimization tasks, especially subjectively optimized bit allocation. In this paper we show that, by way of low-complexity enhancements of our previous work on a perceptually weighted PSNR (WPSNR) metric, addressing shortcomings with video and ultra high-definition content, the prediction of human judgments of video coding quality by the WPSNR can be improved. In fact, the resulting XPSNR seems to match the performance of the aforementioned state-of-the-art methods.

Index Terms— PSNR, SSIM, UHD, video coding, VQA

1. INTRODUCTION

With the introduction of high-quality video streaming services about ten years ago, the demand for real-time automated video quality assessment (VQA) for so-called quality of experience (QoE) purposes [1] increased. The basic goal of VQA here is to estimate the subjective visual quality of a coded / decoded video presentation, usually in relation to the uncoded original input video sequence (full-reference VQA) frame-by-frame, scene-by-scene, or file-by-file. Given the well-known inaccuracy of the peak signal-to-noise ratio (PSNR) in predicting an average subjective judgment of visual coding quality [2]–[4] for a given codec *c* and image or video stimulus *s*, numerous better performing measures have been developed over the last two decades. The most commonly applied are the structural similarity measure (SSIM) [2] and its multi-scale extension, the MS-SSIM [3] as well as a recently proposed video multi-method assessment fusion (VMAF) design combining several other metrics by means of machine learning [4]. The VMAF approach was found to be especially useful for the assessment of video coding quality [5], but determining objective VMAF scores is, algorithmically, quite complex. More importantly, however, the VMAF algorithm is not differentiable [6]. Thus, it cannot be adopted as a perceptual control model for visually optimized bit-allocation strategies during image or video encoding, as it is possible with PSNR or SSIM-based measures.

In JVET-H0047 [7], we proposed a block-wise perceptually weighted distortion measure as an improvement of the PSNR metric, called WPSNR, which was enhanced in JVET-K0206 [8] and finalized in JVET-M0091 [9]. Recently, this WPSNR measure was found to correlate with subjective mean opinion score (MOS) data at least as well as (MS-)SSIM across a set of MOS annotated still image databases [10], as seen in Table **1**. On video data, however, the correlation with MOS scores, e. g., those provided in [5] or the results of JVET's recent Call for Proposals (CfP) on video compression solutions [11], was found to be worse than that of (MS-)SSIM or VMAF. This is especially true for ultra high-definition (UHD) content with a resolution of more than, say, 2048×1280 luminance (or luma) samples and indicates a necessity for improvement.

In the following four sections, a summary of the block-wise WPSNR measure (Sec. 2) and descriptions of low-complexity extensions for motion picture processing (Sec. 3), improved performance in case of temporally varying video quality (Sec. 4), and the handling of UHD image and video material (Sec. 5), are provided to address the mentioned shortcomings. Sec. 6 presents the outcome of empirical evaluations of the extended perceptually weighted PSNR (XPSNR) on various MOS annotated video databases. Lastly, Sec. 7 concludes the paper.

2. REVIEW OF BLOCK-BASED WPSNR MEASURE

The $\mathrm{WPSNR}_{c,s}$ value for a codec *c* and frame of a video sequence (or still image stimulus) *s* is given, similarly to PSNR, by

$$\mathrm{WPSNR}_{c,s} = 10\log_{10}\left(\frac{W \cdot H \cdot (2^{BD}-1)^2}{\sum_k \left(w_k \cdot \sum_{[x,y]\in B_k} (s_c[x,y]-s[x,y])^2\right)}\right),$$

where *W* and *H* are the luma width and height, respectively, of *s*, *BD* is the coding bit-depth per pixel, and *sensitivity weight*

$$w_k = \left(\frac{a_{\mathrm{pic}}}{a_k}\right)^{\beta} \quad \text{with} \quad a_{\mathrm{pic}} = 2^{BD} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}}, \quad \beta = 0.5$$

is a scale factor associated with each $N \times N$ sized block B_k and derived from the block's *spatial activity* a_k [10], [12]. We set

$$N = \mathrm{round}\left(128 \cdot \sqrt{\frac{W \cdot H}{3840 \cdot 2160}}\right)$$

Correlation	PSNR	SSIM	MS-SSIM	WPSNR
SROCC	0.8861	0.9509	0.9569	**0.9604**
PLCC	0.8730	0.9231	0.9103	**0.9408**

Table 1. *Mean correlation between subjective MOS and objective values across JPEG and JPEG 2000 compressed still images of four databases. SROCC: Spearman rank-order, PLCC: Pearson linear correlation coefficients. Data taken from [10].*

since, for the commonly utilized HD and UHD resolutions of 1920×1080 and 3840×2160 luma samples, respectively, this nicely aligns with the largest block sizes used in modern video codecs. a_{pic} was specified such that, on average, $w_k \approx 1$ over a large set of images. If $w_k = 1$ for all k, the PSNR is obtained. In other words, the WPSNR represents a *generalization* of the PSNR by way of a block-wise weighting (via w_k) of the mean squared error (MSE, also called distortion) between the input signal s and distorted (here, coded by c) output signal s_c.

For video sequences, the frame-wise logarithmic $\text{WPSNR}_{c,s}$ values can be averaged arithmetically to obtain a single result:

$$\text{WPSNR}_c = \frac{1}{F} \cdot \sum_{i=1}^{F} \text{WPSNR}_{c,s_i},$$

where F indicates the total number of frames in the sequence.

3. EXTENSION OF WPSNR FOR MOVING PICTURES

The spatially adaptive WPSNR method of [10], [12] and Sec. 2 can easily be extended to motion picture signals s_i, where i represents the frame index in the video, by adding a temporal adaptation into the calculation of the visual activity a_k. Given that, in our prior work, a_k was determined from a filtered s_i as

$$a_k = \max\left(a_{\text{min}}^2, \left(\frac{1}{4N^2} \sum_{[x,y]\epsilon B_k} \left|h_{\text{s}_i}[x,y]\right|\right)^2\right),$$

where x and y are the horizontal and vertical indices of input s_i and h_{s} is a high-pass signal obtained using the convolution $h_{\text{s}} = s * H_{\text{s}}$ with *spatial* filter H_{s}, the temporal adaptation can be incorporated by adding to h_{s} a temporal high-pass signal h_{t}:

$$\hat{a}_k = \max\left(a_{\text{min}}^2, \left(\frac{1}{4N^2} \sum_{[x,y]\epsilon B_k} \left|h_{\text{s}_i}[x,y]\right| + \gamma\left|h_{\text{t}_i}[x,y]\right|\right)^2\right)$$

To obtain $h_{\text{t}} = s * H_{\text{t}}$, two simple *temporal* filters H_{t} were found to work well. The first one, a first-order FIR applied for frame rates of 30 Hz or less (e. g., 24, 25, and 30 frames per second), is given by $h_{\text{t}_i}[x,y] = s_i[x,y] - s_{i-1}[x,y]$ whereas the second one, a second-order FIR employed for frame rates higher than 30 Hz (e. g., 48, 50, and 60 frames per second), is defined as $h_{\text{t}_i}[x,y] = s_i[x,y] - 2s_{i-1}[x,y] + s_{i-2}[x,y]$. Put differently, one or two past frame inputs are used to determine an estimate of the *temporal activity* in each block B_k of every frame s over time. The dependency of the filter order of H_{t} on the frame rate is founded on psychovisual considerations: the limited temporal (high-pass-like) integration of visual stimuli in human perception [13] implies that a shorter filter must be used for relatively low frame rates than for higher ones. Note that taking the absolute value of the outputs of our first-order high-pass is equivalent to the ATI filter utilized in [14].

The relative weighting parameter γ is an empirically determined constant for which we chose $\gamma = 2$. To compensate for the increased intensity of a_k after introducing $|h_{\text{t}}|$, we readjust w_k:

$$\hat{w}_k = \left(\frac{\hat{a}_{\text{pic}}}{\hat{a}_k}\right)^{\beta} \text{ with } \hat{a}_{\text{pic}} = 2^{(BD+1)} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}}, \quad \beta = 0.5.$$

It is worth noting that the temporal activity component of $\hat{a}_k$ introduced in this work is a relatively crude, but very low-complexity, approximation of a block-wise motion estimation process, as it is applied in all modern video codecs. Naturally, more sophisticated, but computationally more complex, temporal activity measures that account for block-internal motion between frames i, $i-1$ and if applicable, $i-2$ before subjecting s_i to the temporal filter h_{t} in i may be devised [15], [16]. Such higher-complexity designs, which may apply neural networks [17] or estimations of multi-scale statistical models [18], are not considered here.

4. TEMPORALLY VARYING VIDEO QUALITY

In Sec. 2 it was noted that, for video sequences, the traditional approach is to average the individual frame (W)PSNR values to obtain a single measurement value for an entire sequence. We observed that, for compressed video material that strongly varies in perceptual quality over time, this form of averaging frame-wise model outputs may not correlate well with MOS values given by human, especially non-expert, observers. The averaging of the logarithmic (W)PSNR values appears to be particularly suboptimal on some video content of high overall visual quality in which, however, brief temporal regions exhibit relatively low quality. With the growing popularity of rate adaptive video streaming, such scenarios actually occur quite often. We discovered experimentally that non-expert viewers, under such circumstances, assign relatively low scores during video quality assessment tasks even if the majority of frames of the compressed video are of excellent quality to their eyes. The consequence is that the log-domain averaged WPSNRs tend to *overestimate* the subjective quality in such cases.

A simple solution to this problem is to average the frame-wise w-weighted distortions $d_i[x,y] = (s_{c,i}[x,y] - s_i[x,y])^2$ derived during the $\text{WPSNR}_{c,s}$ calculations (i. e., the denominator in 1st equation in Sec. 2) instead of the $\text{WPSNR}_{c,s}$ values themselves:

$$\text{WPSNR}'_c = 10\log_{10}\left(\frac{F \cdot W \cdot H \cdot (2^{BD}-1)^2}{\sum_{i=1}^{F}\left(\sum_k\left(w_k \cdot \sum_{[x,y]\epsilon B_k} d_i[x,y]\right)\right)}\right).$$

This root-mean-square (RMS) solution, however, sometimes results in an *underestimation* of the visual quality, so we apply

$$\mathrm{WPSNR}''_c = 20\log_{10}\left(\frac{F\cdot\sqrt{W\cdot H}\cdot(2^{BD}-1)}{\sum_{i=1}^{F}\sqrt{\sum_k\left(w_k\cdot\sum_{[x,y]\in B_k} d_i[x,y]\right)}}\right),$$

i. e., a square-mean-root (SMR) approach [19] yielding output values lying about midway between the log-domain WPSNR_c and the linear-domain WPSNR'_c results, as shown in Figure **1**.

5. VERY HIGH-RESOLUTION IMAGES AND VIDEOS

It was noted that, particularly for UHD images and video sequences, the initial WPSNR assessments of [7]–[10] and [12] still correlate quite poorly with subjective judgments of visual coding quality. In fact, on such content the WPSNR performs only marginally better than the traditional PSNR metric. One possible explanation is that UHD videos are typically viewed on similar screen sizes as lower-resolution content with, e. g., only 1920×1080 luma samples. As a result, the samples of a UHD image are displayed smaller than those of an (upscaled) HD picture, a fact which should be taken into account during the visual activity calculation in the WPSNR algorithm.

Here, a logical countermeasure is to extend the support of the spatial high-pass filter H_s such that it extends across more neighboring samples of $s[x, y]$. Given that in [8], [10] we used

$$H_s = \begin{bmatrix} -1 & -2 & -1 \\ -2 & 12 & -2 \\ -1 & -2 & -1 \end{bmatrix}$$

or a scaled version thereof (multiplied by ¼ in [10]), a simple approach would be to *upsample* H_s by a factor of two, i. e., to increase its size from 3×3 to 6×6 or even 7×7. This, however, would cause a considerable increase of the algorithmic complexity of the spatio-temporal visual activity calculation.

For this reason, we employ an alternative solution in which we acquire the visual activity $\hat{a}_k$ from a *downsampled* version of the input frame sequence s_{i-2}, s_{i-1}, s_i when the input picture or video is larger than 2048×1280 luma samples. Hence, only a single value of $h_{s_i}[x, y]$ and, in case of videos, a single value of $h_{t_i}[x, y]$ is calculated for each 2×2 quadruple of samples of s_i. This approach is not new and has been applied in a number of VQA methods, most prominently the MS-SSIM [3].

It is worth noting in this regard that the downsampling and high-pass operations can be unified into one processing step by designing the high-pass filters appropriately, thus resulting in minimal computational overhead. We utilize the following:

$$\check{h}_{s_i}[x,y] = s_i[x,y] * \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & -2 & -3 & -3 & -2 & -1 \\ -1 & -3 & 12 & 12 & -3 & -1 \\ -1 & -3 & 12 & 12 & -3 & -1 \\ -1 & -2 & -3 & -3 & -2 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix},$$

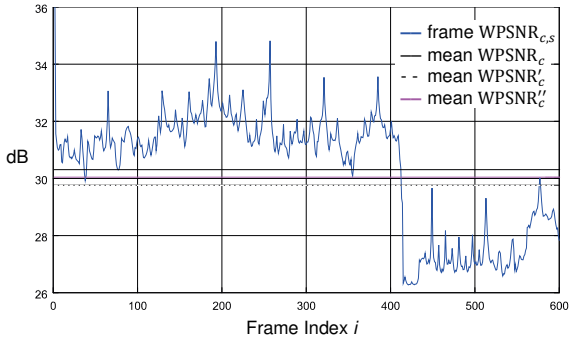

Figure 1. *Results of three temporal (W)PSNR averaging methods on coded video with visual quality drop (MarketPlace, 10s [11]).*

$$\check{h}_{t_i}[x,y] = \check{s}_i[x,y] - \check{s}_{i-1}[x,y] \quad \text{or}$$
$$\check{s}_i[x,y] - 2\check{s}_{i-1}[x,y] + \check{s}_{i-2}[x,y],$$

where the ˇ denotes the downsampling process and $\check{s}_i[x,y] = s_i[x,y] + s_i[x+1,y] + s_i[x,y+1] + s_i[x+1,y+1]$.

Using $\check{s}_i[x,y]$, the sample-wise activity values required for the computation of $\hat{a}_k$ (or a_k for still-image input) need to be determined only for the *even* values of x and y, i. e., for every fourth value of the input sample set s. This particular benefit of the downsampled high-pass operation is illustrated in Figure **2** for the exemplary case of a WPSNR analysis block B_k of size 12×12 samples (i. e., $N = 12$). Other than restricting x and y to be incremented only in steps of two in the downsampling case, the calculation of $\hat{a}_k$ (or a_k), as described in Sec. 3 (or 2), remains unchanged, including the division by $4N^2$.

It must be emphasized that the downsampling of s_i is only applied temporarily during the calculation of the block-wise activity value $\hat{a}_k$ (or a_k for single images). The distortion sum accumulated by the WPSNR metric, i. e., $\sum_{[x,y]\in B_k} d_i[x,y]$ in the above equations, is still determined at the input resolution without downsampling, even for UHD input, as in the PSNR.

6. EXPERIMENTAL EVALUATION

The WPSNR version extended by the techniques described in the previous sections, which we call XPSNR for the sake of differentiability, was evaluated on a selection of MOS annotated databases of compressed and decoded videos of different resolutions and bit-depths, up to UHD and 10 bit per sample. Specifically, two types of mean MOS-vs-XPSNR correlation coefficients were determined to quantify the overall accuracy with which the objective XPSNR values predict the subjective MOS assessments for a given set of videos. Pearson's linear correlation coefficient (PLCC) indicates the degree of linear-model fit while Spearman's rank-order correlation coefficient (SROCC) shows how well the relationship between MOS and XPSNR pairs of values can be described using a monotonic function. The correlation statistics for the widely used PSNR, SSIM, MS-SSIM, and VMAF metrics as well as the original block-based WPSNR method [10] serve as comparative data.

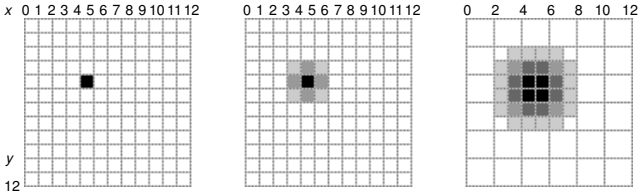

Figure 2. *Sample-wise high-pass filtering by H_s of input signal $\mathbf{s}$ (left) without (center) and with (right) spatial downsampling of $\mathbf{s}$ during the filtering. When downsampling, 4 inputs yield 1 output.*

6.1. Selection of MOS Annotated Databases

For easy comparison, we adopt the annotated video databases already used in [20] for this evaluation, namely, subsets of the Yonsei [21], Live [22], and IVP [23] sets, as well as the ECVQ and EVVQ databases introduced in [24]. To add more videos compressed with state-of-the-art codecs, we further included the SJTU 4K Video Subjective Quality dataset of [25] and the sequences coded with HEVC [26] and Fraunhofer HHI's proposal [27], created for evaluation in JVET's recent CfP [11], for both of which sequence-wise MOS data are available. In addition, B-Com kindly agreed to calculate VQA statistics for the HM [28] and VTM [29] coded sequences of the HEVC-vs-VVC subjective comparison published in [5] for this study.

6.2. Resulting Metric-vs-MOS Correlations

Tables 2 and 3 contain the database-wise (in rows) PLCC and SROCC results, respectively, for the comparison between the corresponding MOS annotations and the individual objective VQA metrics (in columns) assessed in this study. The closer the value for a VQA measure is to one, the better the measure succeeds in predicting subjective video quality. Note that the Live [22] and IVP [23] datasets contain not only visual coding distortion subsets but also, e. g., error concealment distortion created when using such techniques in case of packet loss and other types of transmission errors. As the tested VQA methods are not explicitly designed for such scenarios, the correlation values here are somewhat lower than on the other datasets.

Overall, it can be observed that the original WPSNR design [10] achieves significantly higher correlation with the MOS data than the PSNR model and that, except for the low-video-resolution ECVQ set, the extensions resulting in the XPSNR further increase this advantage. Moreover, the performance of the XPSNR, averaging at a satisfactory 0.82 for PLCC and 0.83 for SROCC, matches that of the other VQA methods.

7. DISCUSSION AND CONCLUSION

In this paper we introduced extensions to our previously proposed VQA algorithm, called WPSNR, to address identified shortcomings when used with motion picture and UHD image or video input. By incorporating a low-complexity temporal visual activity model (Sec. 3), modified frame averaging (Sec. 4), and spatial downsampling in the visual activity calculation

DB	PSNR	SSIM	MS-SSIM	VMAF	WPSNR	XPSNR
[21]	0.822	0.789	0.765	**0.942**	0.916	0.919
[22]	0.539	0.626	0.675	**0.729**	0.637	0.702
[23]	0.632	0.570	0.546	0.591	0.686	**0.707**
ECV	0.733	**0.879**	0.853	0.830	0.848	0.784
EVV	0.727	0.881	0.874	**0.937**	0.880	0.897
SJTU	0.721	0.765	0.810	0.827	0.783	**0.829**
CfP	0.717	0.794	0.743	0.862	0.692	**0.863**
[5]	0.722	0.826	0.799	**0.855**	0.759	0.818
mean	0.702	0.766	0.758	**0.822**	0.775	0.815

Table 2. *Evaluation results for Pearson linear correlation. Higher values mean higher correlation with associated MOS values.*

DB	PSNR	SSIM	MS-SSIM	VMAF	WPSNR	XPSNR
[21]	0.860	**0.949**	0.925	0.915	0.939	0.935
[22]	0.523	0.694	0.732	**0.752**	0.605	0.675
[23]	0.647	0.635	0.574	0.580	0.690	**0.709**
ECV	0.762	**0.916**	0.881	0.736	0.859	0.816
EVV	0.764	0.908	0.911	0.874	0.905	**0.926**
SJTU	0.739	0.807	0.799	0.791	0.814	**0.877**
CfP	0.739	0.810	**0.881**	0.867	0.724	0.866
[5]	0.703	0.848	0.832	**0.850**	0.730	0.812
mean	0.717	0.821	0.817	0.796	0.783	**0.827**

Table 3. *Evaluation results for Spearman rank-order correlation.*

(Sec. 5), the MOS prediction performance of the resulting extended WPSNR (XPSNR), measured on numerous annotated video databases (Sec. 6), was shown to improve to the point where it matches that of other commonly used VQA methods. Given that, unlike VMAF or some other designs, the XPSNR is equally usable for perceptual image/video encoder control purposes (e. g., bit-allocation) than the WPSNR approach [7], [8], [12] – only the block-wise sensitivity weight w_k needs to be changed to $\widehat{w}_k$ – while maintaining the benefit of very low computational complexity, we consider the XPSNR a valuable addition to the list of coding quality specific VQA solutions. Note that, in this study, only the luma components were considered. Therefore, we will focus on incorporating the chroma channels into the visual activity model in our future work.

8. ACKNOWLEDGMENT

The authors thank Pierrick Philippe (formerly B-Com) for his support in calculating the measurement values on the HM and VTM coded videos of the subjective test presented in [5].

9. REFERENCES

[1] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale Structural Similarity for Image Quality assessment," in *Proc. IEEE 37th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, Nov. 2003.

[4] Netflix Inc., "VMAF – Video Multimethod Assessment Fusion," 2019, online: https://github.com/Netflix/vmaf, https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[5] P. Philippe, W. Hamidouche, J. Fournier, and J. Y. Aubié, "AHG4: Subjective comparison of VVC and HEVC," Joint Video Experts Team, doc. JVET-O0451, Gothenburg, July 2019.

[6] Z. Li, "VMAF: the Journey Continues," in *Proc. Mile High Video workshop*, Denver, 2019, online: http://mile-high.video/files/mhv2019/pdf/day1/1_08_Li.pdf.

[7] S. Bosse, C. R. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, "Perceptually optimized QP adaptation and associated distortion measure," doc. JVET-H0047, Macau, Oct./Dec. 2017.

[8] C. R. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, "AHG10: Improved perceptually optimized QP adaptation and associated distortion measure," doc. JVET-K0206, Ljubljana, July 2018.

[9] C. R. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, "AHG10: Clean-up and finalization of perceptually optimized QP adaptation method in VTM," doc. JVET-M0091, Marrakech, Dec. 2018.

[10] J. Erfurt, C.R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, "A Study of the Perceptually Weighted Peak Signal-to-Noise Ratio (WPSNR) for Image Compression," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Taipei, pp. 2339–2343, Sep. 2019.

[11] V. Baroncini, "Results of Subjective Testing of Responses to the Joint CfP on Video Compression Technology with Capability beyond HEVC," doc. JVET-J0080, San Diego, Apr. 2018.

[12] C. R. Helmrich, S. Bosse, M. Siekmann, H. Schwarz, D. Marpe, and T. Wiegand, "Perceptually Optimized Bit-Allocation and associated Distortion Measure for Block-Based Image or Video Coding," in *Proc. IEEE Data Compress. Conf. (DCC)*, Snowbird, pp. 172–181, Mar. 2019.

[13] A. Valberg, *Light Vision Color*, 1st ed., Wiley, Mar. 2005.

[14] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[15] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal Trajectory Aware Video Quality Measure," *IEEE J. Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, Apr. 2009.

[16] K. Seshadrinatan and A. C. Bovik, "Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos," *IEEE Trans. Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[17] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. 15th Europ. Conf. on Computer Vision (ECCV)*, Munich, pp. 219–234, Sep. 2018.

[18] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatio-temporal Feature Integration and Model Fusion for Full-Reference Video Quality Assessment," *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 29, no. 8, Aug. 2019.

[19] D. McK. Kerslake, *The Stress of Hot Environments*, p. 37, 1st ed., Cambridge University Press, July 1972, online: https://books.google.de/books?id=FQo9AAAAIAAJ&pg=PA37&f=false#v=snippet&q=%22square%20mean%20root%22&f=false.

[20] S. Becker, K.-R. Müller, T. Wiegand, and S. Bosse, "A Neural Network Model of Spatial Distortion Sensitivity for Video Quality Estimation," in *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, Pittsburg, pp. 1–6, Oct. 2019.

[21] M. Cheon and J. Lee, "Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience," *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 28, no. 7, pp. 1467–1480, July 2018.

[22] K. Seshadrinatan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Trans. Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[23] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP Subjective Quality Video Database," 2011–2012, online: http://ivp.ee.cuhk.edu.hk/research/database/subjective.

[24] M. Vranješ, S. Rimac-Drlje, and D. Vranješ, "ECVQ and EVVQ Video Quality Databases," in *Proc. 54th Int. Symposium ELMAR-2012*, Zadar, pp. 13–17, Sep. 2012.

[25] Y. Zhu, L. Song, R. Xie, and W. Zhang, "SJTU 4K Video Subjective Quality Dataset for Content Adaptive Bitrate Estimation without Encoding," in *Proc. IEEE Int. Symposium on BMSB*, Nara, June 2016.

[26] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[27] J. Pfaff *et al.*, "Video Compression Using Generalized Binary Partitioning, Trellis Coded Quantization, Perceptually Optimized Encoding and Advanced Prediction and Transform Coding," *IEEE Trans. Circuits and Systems for Video Technol.*, to appear in Dec. 2019.

[28] JCT-VC and Fraunhofer HHI, "High Efficiency Video Coding (HEVC)," online: https://hevc.hhi.fraunhofer.de.

[29] JVET, Fraunhofer HHI, "VVCSoftware_VTM," online: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM