A STUDY OF THE EXTENDED PERCEPTUALLY WEIGHTED PEAK SIGNAL-TO-NOISE RATIO (XPSNR) FOR VIDEO COMPRESSION WITH DIFFERENT RESOLUTIONS AND BIT DEPTHS

Christian R. Helmrich[1], Sebastian Bosse[1], Heiko Schwarz[1, 2], Detlev Marpe[1], and Thomas Wiegand[1, 3]

[1] Video Coding and Analytics Department, Fraunhofer Heinrich Hertz Institute (HHI), Berlin, Germany
[2] Institute for Computer Science, Free University of Berlin, Germany
[3] Image Communication Group, Technical University of Berlin, Germany

Abstract *– Fast and accurate estimation of the visual quality of compressed video content, particularly for quality-of-experience (QoE) monitoring in video broadcasting and streaming, has become important. Given the relatively poor performance of the well-known peak signal-to-noise ratio (PSNR) for such tasks, several video quality assessment (VQA) methods have been developed. In this study, the authors' own recent work on an extension of the perceptually weighted PSNR, termed XPSNR, is analyzed in terms of its suitability for objectively predicting the subjective quality of videos with different resolutions (up to UHD) and bit depths (up to 10 bits/sample). Performance evaluations on various subjective-MOS annotated video databases and investigations of the computational complexity in comparison with state-of-the-art VQA solutions like VMAF and (MS-)SSIM confirm the merit of the XPSNR approach. The use of XPSNR as a reference model for visually motivated control of the bit allocation in modern video encoders for, e.g., HEVC and VVC is outlined as well.*

Keywords – Data compression, HD, HEVC, PSNR, QoE, SSIM, UHD, video coding, VMAF, VQA, VVC, WPSNR

1. INTRODUCTION

The consumption of compressed digital video content via over-the-air broadcasting or Internet Protocol (IP) based streaming services is steadily increasing. This, in turn, leads to a rapid increase in the amount of content distributed using these services. Thus, it is desirable to make use of schemes for automated monitoring of the instantaneous fidelity of the distributed audio-visual signals in order to maintain a certain degree of quality of service (QoS) or, as pursued more recently, quality of experience (QoE) [1]. With regard to the video signal part, such monitoring is realized by way of automated video quality assessment (VQA) algorithms which analyze each distributed moving-picture sequence frame-by-frame with the objective of providing a frame-wise or scene-wise estimate of the subjective visual quality of the tested video, as it would be perceived by a group of human observers. Full-reference VQA methods are generally employed, which means that the *distributed* video—here, the coded and decoded signal—is evaluated in relation to the spatio-temporally synchronized, uncoded *reference* video. In other words, the reference video represents the *input* sequence to the video encoder while the distributed video is the *output* sequence of the video decoder, as illustrated in Fig. 1.

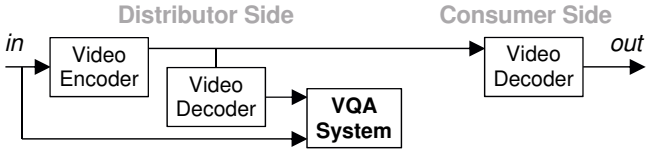

Fig. 1 – Location of automatic VQA on the video distribution side.

Given the well-known inaccuracy of the peak signal-to-noise ratio (PSNR) in predicting an average subjective judgment of perceptual coding quality [2] for a specific codec (coder-decoder) *c* and image or video stimulus (or simply, signal) *s*, various better performing models have been devised over the last years. The most widely employed are the structural similarity measure (SSIM) [3] and its multiscale extension, MS-SSIM [4], as well as a more recently proposed video multimethod assessment fusion (VMAF) approach combining several other measures using a support vector machine [5]. Further VQA metrics worth noting are [6]–[9], which account for frequency dependence in the human visual system.

Although VMAF was found to be a feasible tool for the evaluation of video coding technology [10],[11], its use for direct encoder control is challenging since it is not differentiable [12]. Furthermore, VMAF currently does not allow for local quality prediction below frame level and, owing to its reliance on several other VQA calculations, is quite complex computationally. The aspect of relatively high complexity is shared by the approach of [6]–[8], utilizing block-wise 8×8 DCTs. However, low-complexity reliable metrics which avoid the use of DCT or multiscale algorithms and which can easily be integrated into video encoders as control model for visually optimized bit allocation purposes, as is the case with PSNR and SSIM based approaches, are highly desirable.

1.1 Prior work by the authors

In JVET-H0047 [13] the authors proposed a block-wise perceptually weighted distortion model as an improve-

ment of the PSNR measure, termed WPSNR, which was examined further in JVET-K0206 [14] and finalized in JVET-M0091 [15]. More recently, model-free weighting was also studied [16]. The WPSNR output values were found to correlate with subjective mean opinion score (MOS) data at least as well as (MS-)SSIM; see [17],[18]. One particular advantage of the WPSNR is its backward compatibility with the conventional PSNR. Specifically, by defining the exponent $0 \leq \beta \leq 1$ [13] controling the impact of the local visual activity measure on the block-wise distortion weighting parameter w (see Sec. 2) as

$$\beta = 0, \qquad (1)$$

all weights w reduce to 1 and, as a result, the WPSNR becomes equivalent to the PSNR [13],[17]. It is shown in [13],[19] that a block-wise perceptual weighting of the local distortion, i. e., the sum of squared errors *SSE* between the decoded and original picture block signal,

$$WSSE = w \cdot SSE = w \cdot \sum_{x,y}(s_{\mathrm{dec}}[x,y] - s[x,y])^2, \qquad (2)$$

can readily be utilized to govern the quantization step-size in an image or video encoder's bit allocation unit. In this way, an encoder can optimize its compression result for maximum performance (i. e., minimum mean weighted block *SSE* and, thus, maximum visual reconstruction quality) according to the WPSNR.

Although, as noted above, the WPSNR proved useful in the context of still-image coding and achieved similar, or even better, subjective performance than MS-SSIM-based visually motivated bit allocation in video coding [19], its use as a general-purpose VQA metric for video material of varying resolution, bit depth, and dynamic range is limited. This is evident from the relatively low correlation between the WPSNR output values and the corresponding MOS data available, e. g., from the study published in [10],[11] or the results of JVET's 2017 Call for Proposals (CfP) on video compression technologies with capability beyond HEVC [20]. In fact, this correlation was found to be worse than that of (MS-)SSIM and VMAF, particularly for ultra-high-definition (UHD) and mixed 8-bit/10-bit video content with a resolution of more than, say, 2048×1280 luma samples.

1.2 Outline of this paper

Given the necessity for an improvement of the WPSNR metric as indicated in Sec. 1.1 above, this paper focuses on and proposes modifications to several details of the WPSNR algorithm. After summarizing the block-wise operation of the WPSNR in Section 2, the paper follows up with descriptions of low-complexity extensions for motion picture processing (Section 3), improved performance in case of varying video quality (Section 4) or input/output bit depth (Section 5), and the handling of videos with very high and low resolutions (Section 6). Section 7 then summarizes the results of experimental evaluation of the respectively extended WPSNR, called XPSNR in this paper, on various MOS annotated video databases and Section 8 concludes the paper. Note that parts of this paper were previously published in [18].

2. REVIEW OF BLOCK-BASED WPSNR

The WPSNR_s output for a codec and a video frame (or still image) stimulus s is defined, similarly to PSNR, as

$$\mathrm{WPSNR}_s = 10 \cdot \log_{10}\left(\frac{W \cdot H \cdot (2^{BD}-1)^2}{\sum_k WSSE_k}\right), \qquad (3)$$

where W and H are the luma-channel width and height, respectively, of s, BD is the coding bit depth per sample,

$$WSSE_k = w_k \cdot \sum_{[x,y]\in B_k}(s_{\mathrm{dec}}[x,y] - s[x,y])^2 \qquad (4)$$

is the equivalent of *(2)* for block B_k at index k, with x, y as the horizontal and vertical sample coordinates, and

$$w_k = \left(\frac{a_{\mathrm{pic}}}{a_k}\right)^{\beta} \text{ with exponent } \beta = \frac{1}{2} \qquad (5)$$

represents the *visual sensitivity* weight (a scale factor) associated with the $N \times N$ sized B_k and calculated from the block's *spatial activity* measure a_k and an average overall activity a_{pic}. Details can be found in [17]–[19].

$$N = \mathrm{round}\left(128 \cdot \sqrt{\frac{W \cdot H}{3840 \cdot 2160}}\right) \qquad (6)$$

was chosen since, for the commonly used HD and UHD resolutions of 1920×1080 and 3840×2160 pixels, this choice conveniently aligns with the largest block size in modern video codecs. a_{pic} is defined empirically such that, on average, $w_k \approx 1$ over a large set of test images and video frames with a specified resolution $W \cdot H$ and bit depth BD [14]; see also Sec. 5. Hence, as indicated in Sec. 1.1, the WPSNR is a *generalization* of the PSNR by means of a block-wise weighting of the assessed *SSE*. For video signals, the frame-wise logarithmic WPSNR_s values are averaged arithmetically to get a single result:

$$\mathrm{WPSNR} = \frac{1}{F} \cdot \sum_{i=1}^{F} \mathrm{WPSNR}_{s_i}, \qquad (7)$$

with F denoting the evaluated number of video frames.

3. EXTENSION FOR MOVING PICTURES

The *spatially adaptive* WPSNR method of [17],[19] and Sec. 2 can easily be extended to motion picture signals s_i, where i is the frame index in the video sequence, by introducing a *temporal adaptation* into the calculation of the visual activity a_k. Given that in our prior studies,

$$a_k = \max\left(a_{\min}^2, \left(\frac{1}{4N^2}\sum_{[x,y]\in B_k}\left|h_{s_i}[x,y]\right|\right)^2\right), \qquad (8)$$

where $h_\mathrm{s} = s * H_\mathrm{s}$ is the signal resulting from filtering s with the *spatial highpass* filter H_s, the temporal adaptation can be integrated by adding to h_s the weighted result $h_\mathrm{t} = s * H_\mathrm{t}$ of a *temporal highpass* filtering step:

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

Note that the lower limit remains at $a_{\min} = 2^{BD-6}$ as in [17]. Two simple highpass filters H_t were found useful. The first one, a first-order FIR applied for frame rates of 30 Hz or lower, is $h_{\mathrm{t}_i}[x,y] = s_i[x,y] - s_{i-1}[x,y]$ and the second one, a second-order FIR used, accordingly, for frame rates above 30 and up to 60 Hz, is defined as $h_{\mathrm{t}_i}[x,y] = s_i[x,y] - 2s_{i-1}[x,y] + s_{i-2}[x,y]$. In other words, one or two previous frames are used to obtain a simple estimate of the *temporal activity* in each block B_k of each signal s over time. Naturally, for frame rates higher than 60 Hz, a third-order FIR could be specified, but due to a lack of correspondingly recorded content, such operating points have not been examined yet. The dependency of the filter order of H_t on the frame rate is based upon psychovisual considerations: the limited temporal (highpass-like) integration of visual stimuli in human perception [21] implies that a shorter filter impulse response should be employed at relatively low frame rates than at higher ones. Naturally, filters which more accurately model the nonlinear temporal contrast sensitivity of the human visual system could be used, but for the sake of simplicity and low complexity, such an option is not considered here. Note, also, that taking the absolute values of the first-order highpass outputs as above is identical to the "absolute value of temporal information" (ATI) filter described in [22].

The relative weight γ is an experimentally determined constant for which $\gamma = 2$ was selected. To compensate for the increased variance in $\hat{a}_k$ relative to a_k after the introduction of term h_t (the sum may increase while $a_{\min}$ remains unchanged), a_pic is modified accordingly [18], resulting in $\hat{a}_\mathrm{pic}$ which, in turn, yields the weight

$$\hat{w}_k = \left(\frac{\hat{a}_\mathrm{pic}}{\hat{a}_k}\right)^\beta \text{ with, again, } \beta = \frac{1}{2} \qquad (9)$$

as a *spatio-temporal* visual sensitivity measure. It must be noted that the temporal activity component of $\hat{a}_k$ is a quite crude, but very low-complexity, approximation of a B_k-wise motion estimation operation, as typically performed in all modern video codecs. Evidently, more elaborate, but computationally more complex, activity metrics accounting for block-internal motion between frames i, i–1 and, for high frame rates, i–2 before deriving h_{t_i} may be devised [23],[24]. Such designs, which may use neural networks [25] or estimations of multi-scale statistical models [26], are not considered here since one objective is to maintain very low complexity.

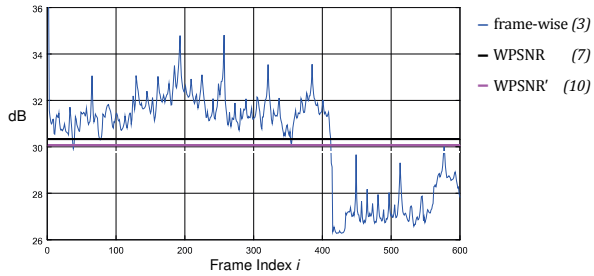

Fig. 2 – Effect of different temporal WPSNR averaging methods on coded video with visual quality drop (*MarketPlace*, HD, 10s [20]).

4. TEMPORALLY VARYING VIDEO QUALITY

It was described in Sec. 2 that, for video sequences, the traditional approach is to average the individual frame (W)PSNR values so as to obtain a single measurement value for a complete video. It was observed [18] that, for compressed video content which strongly varies in visual quality over time, such averaging of frame-wise model output may not always correlate well with MOS values provided by human observers. The averaging of logarithmic (W)PSNR values appears to be especially suboptimal on decoded video material of high overall visual quality in which, however, brief passages exhibit relatively low quality. With the growing popularity of rate adaptive video streaming, particularly on mobile devices, such situations actually occur quite frequently. It was discovered experimentally that, under such circumstances, non-expert viewers assign relatively low scores to the tested video (compared with a video with balanced visual quality) during subjective VQA tasks, even if the majority of frames of the compressed video are of excellent quality to their eyes. This observation, which is confirmed by QoE-related feedback of consumers as reported by, e.g., Netflix [12], indicates that the log-domain average WPSNR values of *(7)* tend to *overestimate* the subjective quality in such cases.

A simple solution to this problem is to apply a square-mean-root (SMR) approach [27] which takes the arithmetic average of the square roots of the linear-domain frame-wise WPSNR_s data (before taking their base-10 logarithms) and, only thereafter, applies the logarithm:

$$\mathrm{WPSNR}' = 20 \cdot \log_{10}\left(\frac{F\cdot\sqrt{W\cdot H}\cdot\left(2^{BD}-1\right)}{\sum_{i=1}^{F}\sqrt{\sum_k WSSE_k}}\right). \qquad (10)$$

Note the use of constant 20 in *(10)* instead of 10 in *(3)*, representing the linear-domain squaring operation. A comparison of WPSNR and WPSNR′ is shown in Fig. 2, with a 0.35 dB lower value of the latter in this example.

5. VARYING INPUT OR OUTPUT BIT DEPTH

Typically, the input and output bit depths of the color planes of a video presentation are held constant for a specific distribution path. Sometimes, however, it may

be beneficial to perform automated VQA across video content of varying bit depth, e. g., when, as in [10],[11], part of the source material is not available in high-bit-depth (at least 10 bits per sample) format. The WPSNR measure was designed so as to "favor" high bit depths over lower ones by returning somewhat higher output values on otherwise identical input signals, assuming that 10 or 12-bit playback would exhibit slightly higher quality than 8-bit playback. This was done by defining

$$a_{\mathrm{pic}} = 2^{BD} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}}, \qquad (11)$$

i. e., dependent on *BD before* exponentiation by β in *(5)*. Experiments conducted by the authors after study [18], however, suggest that the gain in visual quality due to bit depths higher than 8 bits per sample is much lower than the gain anticipated by the WPSNR metric. Hence, the correlation of WPSNR with MOS data on annotated databases of mixed-bit-depth video content is reduced by an undesirable and unnecessary amount.

In order to address this issue, a modified definition of the a_{pic} constant used in [17]–[19] is suggested. Given the change to the visual activity a_k described in Sec. 3,

$$\hat{a}_{\mathrm{pic}} = 2^{BD/\beta - 9} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}} \qquad (12)$$

is proposed to render the visual sensitivity weight $\hat{w}_k$ independent of *BD after* the exponentiation by β in *(9)* and, at the same time, compensate for the modified $\hat{a}_k$.

6. HIGH AND LOW-RESOLUTION VIDEOS

It was noted that particularly for UHD image and video material, the initial WPSNR assessments of [13]–[15] or [17],[19] still correlate quite poorly with subjective judgments of visual coding quality. In fact, on this kind of content the WPSNR was found to perform only marginally better than the conventional PSNR metric. One possible explanation is that, on consumer devices like TVs or laptops, UHD stimuli are generally being viewed on the same screen area, and from the same viewing distance, as HD stimuli with at most 1920×1080 luma samples even though, due to the fixed average angular resolution of the human eye of about 60 pixels per degree, such a form of presentation is discouraged [28]. As a result, samples of an UHD image are being displayed smaller than those of (upscaled) HD images, a fact that should be taken into account during the calculation of the visual activity in the algorithm of Sec. 3.

A logical solution here is to extend the spatial support of the highpass filter H_{s} so that it extends across more neighboring samples of $s[x,y]$. Since in [14], the filter is

$$h_{\mathrm{s}_i}[x,y] = s_i[x,y] * \begin{bmatrix} -1 & -2 & -1 \\ -2 & 12 & -2 \\ -1 & -2 & -1 \end{bmatrix} \qquad (13)$$

and in [17], a scaled version thereof (multiplied by ¼)[1], 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 7×7. This, however, would significantly increase the computational complexity of the calculation of $\hat{a}_k$.

Therefore, an alternative approach is pursued in which $\hat{a}_k$ is determined from a 4× (2× horizontal, 2× vertical) *downsampled* version of frame sequence s_{i-2}, s_{i-1}, s_i when the image or video is larger than, say, 2048×1280 luma samples (something between HD and UHD). Thus, only a single value of $h_{\mathrm{s}_i}[x,y]$ and, for videos, $h_{\mathrm{t}_i}[x,y]$ is computed for each 2×2 quadruple of the pixels of s_i.

It is worth mentioning in this respect that the downsampling and highpass operations can be unified into a single processing step by designing the highpass filters appropriately, thereby resulting in minimum computational overhead. The following filter was devised:

$$\check{h}_{\mathrm{s}_i}[x,y] = s_i[x,y] * \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & -2 & -3 & -3 & -2 & -1 \\ -1 & -3 & 12 & 12 & -3 & -1 \\ -1 & -3 & 12 & 12 & -3 & -1 \\ -1 & -2 & -3 & -3 & -2 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix}, \qquad (14)$$

$$\check{h}_{\mathrm{t}_i}[x,y] = \check{s}_i[x,y] - \check{s}_{i-1}[x,y] \text{ or } \check{s}_i[x,y] - 2\check{s}_{i-1}[x,y] + \check{s}_{i-2}[x,y], \qquad (15)$$

where accent ˇ denotes the downsampling process and

$$\check{s}_i[x,y] = s_i[x,y] + s_i[x+1,y] + s_i[x,y+1] + s_i[x+1,y+1].$$

Naturally, higher downsampling factors like 16× could be used for very high resolutions such as 8192×4320 (8K), but as with very high frame rates in Sec. 3, such a configuration could not be studied intensively due to a lack of correspondingly recorded material.

By way of $\check{s}_i[x,y]$, the spatio-temporal highpass values required for the calculation of $\hat{a}_k$ need to be obtained only for the *even* values of x and y, i. e., for every fourth value of s. This particular benefit of the downsampled highpass operation, resulting in a *per-pixel* UHD filter complexity similar to that for HD input, is illustrated in Fig. 3 for an exemplary case of a WPSNR analysis block of 12×12 samples (B_k with $N = 12$). Other than restricting x and y to be incremented only in steps of two in the downsampling case, the computation of $\hat{a}_k$ (or a_k), as in Sec. 3, can stay unchanged, including the division by $4N^2$ in *(8)* which is compensated for by $\check{h}_{\mathrm{s}_i}$ and $\check{h}_{\mathrm{t}_i}$.

It must be emphasized that the downsampling of s_i is only done implicitly during the derivation of the block-wise $\hat{a}_k$ (or a_k for still-image signals). The *SSE* values accumulated by the WPSNR methods in *(3)*, *(7)*, or *(10)*

[1] Divisions by powers of 2 can be implemented via right-shifts.

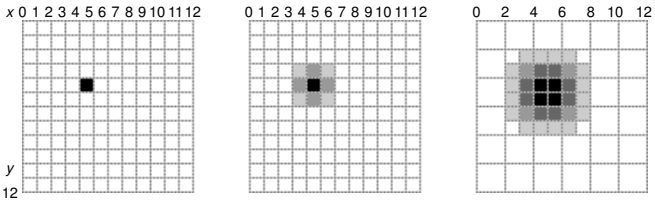

Fig. 3 – Sample-wise spatial highpass filtering, using H_S, of signal $s_i[4, 4]$ (left) without (center) and with (right) downsampling of s_i during filtering. When downsampling, 4 inputs yield 1 output.

are still obtained at the input resolution as in the PSNR, without downsampling even for UHD signals.

Having addressed high (UHD) resolutions, a particular shortcoming in the WPSNR model was also discovered at very low resolutions where, due to *(6)*, the block size becomes relatively small. Specifically, it was observed that, for $N \leq 24$ or so, large outliers may exist in the w matrix, causing undesirably strong weighting of some block *SSE* values in *(4)* and, as a result, a relatively low WPSNR figure which does not coincide with subjective impression. In order to minimize the likelihood of such occurrences, in-place *smoothing* of the block weights w or $\hat{w}$ prior to their use in *(4)* is proposed. Let w_k be the previous block's weight (i.e., w_{k+1} has already been calculated) and assume that k increases first from left to right and then from top to bottom across the picture (i. e., in line-scan fashion), starting at $k = 0$ which is located at the top-left corner of the picture. Then, given

- the *left* neighbor, L, of w_k as $L = 0$ if k is in the first picture block-column, otherwise $L = w_{k-1}$,
- the *top* neighbor, T, of w_k as $T = 0$ if k is in the first picture block-row, otherwise $T = w_{k-WK}$,
- the *right* neighbor, R, of w_k as $R = 0$ if k is in the last picture block-column, otherwise $R = w_{k+1}$,

with WK being the picture width in units of k (stride), for all $k = 1, 2, \ldots, N-1$ (excluding $k = 0$) change w_k to

$$w'_k = \min\big(w_k, \max(L, T, R)\big) \qquad (16)$$

and use w'_k instead of w_k or $\hat{w}_k$ whenever it is available and the picture is of size 640×480 samples or smaller. This post-processing of the w_k is easy to implement in hardware and software and effectively removes strong "peaks" in the block weighting for some low-resolution input without sacrificing the benefit of perceptual *SSE* weighting over PSNR-like unweighted *SSE* assessment. An example is shown in Fig. 4 for a set of MPEG-4 encoded videos of resolution 352×288 luma samples. As can be observed, the smoothing moves the outlier data of video *hall* closer to the data of the remaining videos.

7. EXPERIMENTAL EVALUATION

The WPSNR method extended by the algorithms presented in the previous sections, called "XPSNR" for the sake of differentiability, was evaluated on a selection of

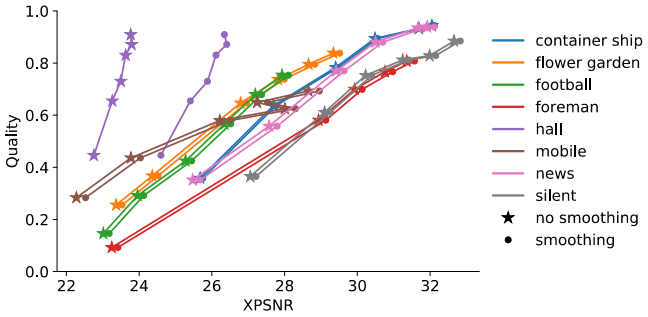

Fig. 4 – Effect of in-place smoothing of $\hat{w}$ on stability of estimating MOS values (quality, normalized to the range 0–1) by the WPSNR with the extensions of Secs. 3–5 (XPSNR, in dB units). Videos are MPEG-4 compressed 352×288 sized stimuli of ECVQ dataset [36].

MOS annotated databases of compressed and decoded video content of different resolutions and bit depths, up to and including UHD and 10 bits per sample. Specifically, two types of mean MOS-vs-XPSNR correlation coefficients were calculated to determine the average accuracy by which the objective XPSNR values predict the subjective MOS assessments for a specific class of videos. Pearson's linear correlation coefficient (PLCC) quantifies the degree of linear-model fit [29] whereas Spearman's rank order correlation coefficient (SROCC) assesses how well the relationship between MOS and XPSNR pairs of values can be described using a monotonic—but not necessarily linear—function [30]. The correlation statistics for the widely used PSNR, SSIM, MS-SSIM, and VMAF measures as well as the original block-based WPSNR [17] serve as comparative figures. Note that the PLCC values are only provided for comparability with the authors' previous publications and that the SROCC data are statistically more suitable for evaluation. Furthermore, since the statistical analysis does not follow more rigorous recommendations such as, e. g., those of ITU-T P.1401 [31], the results are only provided for informative purposes, and no conclusions on any cross-model statistical significances are drawn.

7.1 Correlation with subjective MOS data

In order to simplify comparisons, the annotated video databases already utilized in [32] are adopted for this evaluation. Specifically, all *video compression* distorted (MPEG-2, MPEG-4, H.264, H.265, and Dirac) subsets of the Yonsei [33], LIVE [34], IVP [35], ECVQ, EVVQ [36], as well as SJTU 4K Video Subjective Quality [37] databases, all of which offer per-video MOS data, are used. To add more content compressed with state-of-the-art codecs, we further included the sequences coded with H.265/HEVC [38] and Fraunhofer HHI's response [39], created for evaluation in JVET's recent CfP [20], as well as those coded by HM [40] and VTM [41] for the HEVC-VVC visual evaluation published in [10],[11], for which B-Com kindly agreed to provide per-video VQA values. The ECVQ and EVVQ subsets exhibit the lowest and the Yonsei, SJTU, CfP, and HEVC-vs-VVC subsets the highest

(UHD) video resolutions. The others, e. g. IVP and parts of [10],[11] and [20] (HD), end up in between. Further information on the sets is planned to be hosted at [42]. The VMAF data for the aforementioned datasets were obtained with software version 1.3.15 (Sep. 2019) and VQA model version 0.6.1 (4K subvariant for UHD content). The VMAF software [5] also provided the PSNR and SSIM statistics. The MS-SSIM, WPSNR, and XPSNR values were calculated with proprietary C++ software, using only the luma component of the video sequences to derive the model output. Further metrics like those of [22],[26] as well as the handling of the chroma components are planned to be added in a follow-up study and will also be provided at [42] once available.

Tables 1 and 2 contain the dataset-wise (in rows) PLCC and SROCC results, respectively, for the comparisons of the subjective MOS annotations and the results of each objective VQA measure (in columns) examined herein. The closer the correlation value for a VQA metric is to one, the better the metric succeeds in predicting average (across frames and videos) quality as judged by a group of human viewers. For each dataset, the best result is written in **bold** type. The LIVE and IVP sets also contain distortion types resembling error concealment techniques applied, e. g., during IP packet loss. As the tested VQA methods were not explicitly developed for such scenarios, the correlation values here are somewhat lower than those for the other sets.

Overall, it can be observed that the original WPSNR of [14],[17] reaches considerably higher correlation with the MOS data than the PSNR metric and that the XPSNR algorithm further increases this advantage. Moreover, the performance of the XPSNR is, for both PLCC as well as SROCC, very similar to that of the best of the other evaluated VQA methods (VMAF for PLCC and SSIM for SROCC) on average, as tabulated in the "mean" rows, a result which indicates the clear merit of this approach.

7.2 Comparison of computational complexity

To evaluate the computational complexity of XPSNR in relation to that of the other VQA methods tested, software runtime analysis (using virtual in-RAM I/O) with single-threaded execution was carried out. The results, albeit not very accurate, indicate that the WPSNR and SSIM algorithms are about 2× as complex as the PSNR design, XPSNR is about 3× as complex as PSNR, and the MS-SSIM and VMAF methods are about 9× as complex. Note that, due to independent calculability of the per-block w_k and $WSSE_k$ data, the WPSNR and XPSNR can easily be optimized for fast, multi-threaded execution, possibly with "single instruction multiple data" (SIMD) operations on modern CPUs thanks to the inter-block independence during the spatio-temporal activity calculation. However, unlike in VMAF, such optimizations have not been implemented in XPSNR as of this writing.

Table 1 – Evaluation results for Pearson linear correlation. Higher values indicate higher correlation with associated MOS values.

Set	PSNR	SSIM	MS-SSIM	VMAF	WPSNR	XPSNR
Yonsei	0.822	0.789	0.765	**0.942**	0.916	0.919
LIVE	0.539	0.626	0.675	**0.729**	0.637	0.702
IVP	0.632	0.570	0.546	0.591	0.686	**0.707**
ECVQ	0.733	**0.879**	0.853	0.830	0.848	0.840
EVVQ	0.727	0.881	0.874	**0.937**	0.880	0.898
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**
[10]	0.722	0.826	0.799	**0.855**	0.759	0.817
mean	0.702	0.766	0.758	**0.822**	0.775	**0.822**

Table 2 – Evaluation results for Spearman rank order correlation. As in Table 1, smoothing is used only on the ECVQ and EVVQ sets.

Set	PSNR	SSIM	MS-SSIM	VMAF	WPSNR	XPSNR
Yonsei	0.860	**0.949**	0.925	0.915	0.939	0.935
LIVE	0.523	0.694	0.732	**0.752**	0.605	0.675
IVP	0.647	0.635	0.574	0.580	0.690	**0.709**
ECVQ	0.762	**0.916**	0.881	0.736	0.859	0.847
EVVQ	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
[10]	0.703	0.848	0.832	**0.850**	0.730	0.813
mean	0.717	0.821	0.817	0.796	0.783	**0.831**

8. DISCUSSION AND CONCLUSION

This paper reviewed, and proposed extensions to, the authors' previous work on perceptually weighted peak signal-to-noise (WPSNR) assessments of compressed video material. More precisely, a low-complexity temporal visual activity model (Sec. 3), an alternative frame averaging method (Sec. 4), bit depth independent output value scaling (Sec. 5), and spatial 4× downsampling (for very high resolutions) and in-place smoothing (for very low resolutions) while calculating the block-wise visual sensitivity values w (Sec. 6) were incorporated into the WPSNR design. The resulting *extended WPSNR* algorithm, called XPSNR, was demonstrated to perform at least as well as the best competing state-of-the-art VQA solutions in predicting the subjective judgments of a number of MOS annotated coded-video databases. This was achieved with a notably lower computational complexity than that required for obtaining, e.g., VMAF or MS-SSIM results, and without optimizing the XPSNR parameters (a_{min}, a_{pic}, β, and γ) for the particular sets

of compressed video sequences employed in this study. Moreover, as noted in Sec. 1.1, the per-block sensitivity weight w_k of *(5)* or $\hat{w}_k$ of *(9)* can be used to easily adapt the quantization parameter (QP) in traditional coders to the instantaneous input characteristics, without having to rely on s_{dec} [13]–[15],[17]–[19],[39]. Specifically,

$$\mathrm{QP}_k = \mathrm{QP}' - \mathrm{round}\big(3 \cdot \log_2(\hat{w}_k)\big) \qquad (17)$$

can be used to XPSNR-optimize the quantization stepsize inside an HEVC or VVC encoder, initialized using a per-frame constant QP′, on a coding block basis. Aside from further statistical evaluation using more datasets, this beneficial aspect, along with the incorporation of chroma-component and/or high dynamic range (HDR) statistics (see, e.g., [43]) and multi-threaded operation, will be the focus of future work on this VQA algorithm.

9. ACKNOWLEDGMENT

The authors thank Pierrick Philippe (formerly B-Com) for helping to calculate the VQA values on the VTM and HM coded videos of the comparative test published in [10],[11] and Sören Becker for assistance in the collection of the correlation values and the creation of Fig. 4.

REFERENCES

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

[2] B. Girod, "What's Wrong With Mean-squared Error?" *Digital Images and Human Vision*, A. B. Watson Ed., Cambridge, MA, US: MIT Press, pp. 207–220, 1993.

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

[4] 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, CA, US, Nov. 2003.

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

[6] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Worksh. Vid. Process. & Quality Metrics*, Scottsdale, AZ, US, Jan. 2006.

[7] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proc. 3rd Int. Worksh. Vid. Process. & Quality Metrics*, US, Jan. 2007.

[8] N. Ponomarenko, O. Eremeev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in Proc. CADSM, Polyana-Svalyava, pp. 305–311, 2011.

[9] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja, "A modified PSNR metric based on HVS for quality assessment of color images," in *Proc. IEEE Int. Conf. on Commun. & Industr. Applic.*, Kolkata, IN, Dec. 2011.

[10] 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, SE, July 2019.

[11] N. Sidaty, W. Hamidouche, P. Philippe, J. Fournier, and O. Deforges, "Compression Performance of the Versatile Video Coding: HD and UHD Visual Quality Monitoring," in *Proc. IEEE Picture Coding Symposium*, Ningbo, CN, Nov. 2019.

[12] Z. Li, "VMAF: The Journey Continues," in *Proc. Mile High Video Workshop*, Denver, 2019, link: http://milehigh.video/files/mhv2019/pdf/day1/1_08_Li.pdf.

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

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

[15] 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, MA, Dec. 2018.

[16] S. Bosse, S. Becker, K.-R. Müller, W. Samek, and T. Wiegand, "Estimation of distortion sensitivity for visual quality prediction using a convolutional neural network," *Digital Sig. Process.*, vol. 91, pp. 54–65, 2019.

[17] 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. Image Process. (ICIP)*, Taipei, pp. 2339–2343, Sep. 2019.

[18] C. R. Helmrich, M. Siekmann, S. Becker, S. Bosse, D. Marpe, and T. Wiegand, "XPSNR: A Low-Complexity Extension of the Perceptually Weighted Peak Signal-to-Noise Ratio for High-Resolution Video Quality Assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Sig. Process. (ICASSP)*, virtual/online, May 2020.

[19] 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 Compression Conf. (DCC)*, Snowbird, UT, US, pp. 172–181, Mar. 2019.

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

[21] A. Valberg, *Light Vision Color*, 1st e., Wiley, Mar. 2005.

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

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

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

[25] 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, DE, pp. 219–234, Sep. 2018.

[26] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment," *IEEE Trans. Circuits Systems f. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.

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

[28] ITU-R, Recommendation BT.500-13, "Method for the subjective assessment of the quality of television pictures," Geneva, CH, Jan. 2012.

[29] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosoph. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[30] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*, p. 508, 2nd e., Lawrence Erlbaum, 2003.

[31] ITU-T, Recommendation P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," Geneva, CH, Jan. 2020.

[32] 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 Machine Learning for Sig. Process. (MLSP)*, Pittsburg, PA, US, pp. 1–6, Oct. 2019.

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

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

[35] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP subjective quality video database," 2011–2012, link: http://ivp.ee.cuhk.edu.hk/research/database/subjective.

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

[37] 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 BMSB*, Nara, JP, June 2016.

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

[39] 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 Systems f. Video Technol.*, vol. 30, no. 5, pp. 1281–1295, May 2020.

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

[41] JVET and Fraunhofer HHI, "VVCSoftware_VTM," link: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM.

[42] C. R. Helmrich, "ecodis – XPSNR – Information Page", Mar. 2020, link: http://www.ecodis.de/xpsnr.htm.

[43] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "ESPL-Live HDR Image Quality Database," May 2016, link: http://live.ece.utexas.edu/research/HDRDB/hdr_index.html.