IV-PSNR – the objective quality metric for immersive video applications

Adrian Dziembowski, Dawid Mieloch, Jakub Stankowski, and Adam Grzelka

***Abstract*—This paper presents a new objective quality metric that was adapted to the complex characteristics of immersive video (IV) which is prone to errors caused by processing and compression of multiple input views and virtual view synthesis. The proposed metric, IV-PSNR, contains two techniques that allow for the evaluation of quality loss for typical immersive video distortions: corresponding pixel shift and global component difference. The performed experiments compared the proposal with 31 state-of-the-art quality metrics, showing their performance in the assessment of quality in immersive video coding and processing, and in other applications, using commonly used image quality assessment databases – TID2013 and CVIQ. As presented, IV-PSNR outperforms other metrics in immersive video applications and still can be efficiently used in the evaluation of different images and videos. Moreover, basing the metric on the calculation of PSNR allowed the computational complexity to remain low. Publicly available, efficient implementation of IV-PSNR software was provided by the authors of this paper and is used by ISO/IEC MPEG for evaluation and research on the upcoming MPEG Immersive video (MIV) coding standard.**

***Index Terms*—image quality, immersive video, video compression, view synthesis**

I. Introduction

The subject of measuring the objective quality of videos is one of the widest in the area of image and video processing. In this paper, the focus is put on the performance of available metrics in evaluating the quality of immersive video. The purpose of immersive video is to allow the viewer to freely navigate in the entire scene by changing her/his position and direction of viewing, e.g., in a whole room in 6DoF applications [1], or its part, allowing free navigation in a limited range in 3DoF+ systems [2].

Regardless of the type of the immersive video system, the virtual navigation of the viewer is provided by synthesizing virtual views between views captured by the cameras [3], [4], using the information about the three-dimensional geometry of a scene represented usually in a form of depth maps. Although this process is not required in systems based on the use of a single omnidirectional 360 video, unfortunately, these systems do not provide a fully immersive experience, as the navigation is limited to the change of the direction of viewing the scene by a user. Therefore, 360 videos, as well as stereopairs used for stereovision, are not in the scope of this paper. Relevant image quality assessment techniques for 360 videos can be found, e.g., in [108], [110], [111], [114]. Fig. 1 provides an overview of the visual systems described above. In further considerations, we focus entirely on the 6DoF (immersive video [1]) systems.

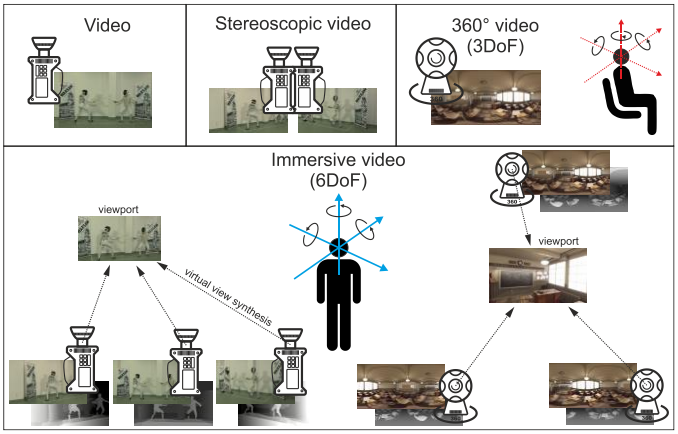


Fig. 1. An overview of the visual systems. Immersive video (bottom) utilizes virtual view synthesis performed using several input views and corresponding depth maps to provide a possibility of six-degrees-of-freedom movement for a viewer. 360º video (top right) enables only the change of orientation of viewing.

Previous reviews and benchmarks of metrics that measure the synthesized view quality indicate the need for new objective metrics that will provide a better correlation with subjective scores [5], as the characteristics of synthesis-induced errors in videos are very specific. The available models of such distortions (e.g., [6], [7], [8]) list possible sources of errors, which include not only typical errors in the texture of scene objects (present also in traditional single-camera videos, e.g., blur, noise, color distortions) but also errors in their position, caused by faulty three-dimensional reprojection [117].

One of the main sources of texture errors in immersive video is color inconsistency [9], as, in order to allow a user to virtually navigate within a scene, the scene has to be captured by a

Manuscript received January 21, 2022. The research was supported by the Ministry of Science and Higher Education of Republic of Poland.

Authors are with the Institute of Multimedia Telecommunications, Poznań University of Technology, 60-965 Poznań, Poland. Corresponding author: Adrian Dziembowski (e-mail: adrian.dziembowski@put.poznan.pl).

multicamera system. In general, each camera can acquire data differently due to non-Lambertian reflections, different characteristics of camera sensors, and camera automatics (e.g., automatic exposure time or white balance).

As it was presented in the review of related works, included in Section III, many metrics that measure the quality of synthesized view do not assume that the synthesis process in practical systems can be performed after lossy compression of input views [10]. A quality metric which is to be used in immersive video applications should also take into account distortions caused by the encoding, which is especially important as the emergence of relevant video codecs can be seen [1] (a summary of state-of-the-art techniques in immersive video compression is presented in Section VII-C).

In this paper, we focus on immersive video, which is a subset of visual immersive media. Relevant quality metrics that measure the quality of other media closely related to immersive video have been proposed e.g., for point-cloud-based systems [11], [12]. Their authors also indicate that, though the use case of these systems is different, compression-induced errors and distortion of geometry of encoded three-dimensional points are among the main problems which relevant quality metrics have to focus on. Moreover, in immersive video, the view presented to the final user cannot be directly compared to any reference view as free navigation is not limited to input views. It makes image quality assessment much more complex and unpredictable.

Given the significance of immersive media and the need for objective evaluation of immersive videos, this paper proposes a new metric called IV-PSNR. This metric introduces two novel techniques that allow for dealing with typical immersive video distortions: corresponding pixel shift (described in detail in Section IV-A), which adapts the traditional PSNR metric to be less sensitive to small, unnoticeable for the viewer, synthesis-induced errors, and global component difference (Section IV-B), which measures color inconsistencies in virtual views. The proposal was shown to be adapted to the complex characteristics of immersive video described above (as indicated by correlation with an independent subjective viewing experiment described in Section VII), while its computational complexity remained low (see Section X) due to being based on the calculation of PSNR.

II. Typical Distortions in Immersive Videos

As mentioned in Section I, an objective quality metric adapted for immersive video applications should properly mimic HVS for typical distortions introduced simultaneously by virtual view synthesis, immersive video processing, and compression.

While other factors also highly influence the quality of the final image presented to a viewer, e.g., the accuracy of camera parameters, lens distortions, or insufficient synchronization of cameras in the multiview systems, these factors influence most of all the depth estimation process, not the quality of the acquired video. Errors in estimated depth, as it is presented in the following Section II-A, directly influence the synthesis process, therefore, the factors listed above can be discussed jointly as errors in the synthesized virtual view.

A. Corresponding Pixel Shift Error

Virtual view synthesis is performed by reprojecting pixels from input views to the virtual views. For perspective views, the reprojection of a pixel can be defined as [57]:

$$\begin{bmatrix} x_v \\ y_v \\ z_v \\ 1 \end{bmatrix} = \boldsymbol{H}_{i,v} \cdot \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix}, \tag{1}$$

where (x_i, y_i) is the position of the pixel in the input view, z_i – its depth, $\mathbf{H}_{i,v}$ is the homography matrix acquired from the multiplication of the projection matrix of the virtual view and the inverted projection matrix of the input view, (x_v, y_v) is the position of the reprojected pixel in the virtual view and z_v – depth of the reprojected pixel.

While x_i and y_i are aligned with the pixel grid of the input view, x_v and y_v have to be rounded in order to fit the finite resolution of the virtual view. Therefore, each pixel of the virtual view may be slightly shifted. For omnidirectional video, the reprojection equation is different, but the problem remains.

Moreover, also the depth resolution is finite, as in practical systems, where video information is being compressed, the depth maps are stored as integers (e.g., 10 or 16-bps videos). In such a case, the z_i value is rounded which implies worse projection accuracy and thus an even bigger shift of the reprojected pixel.

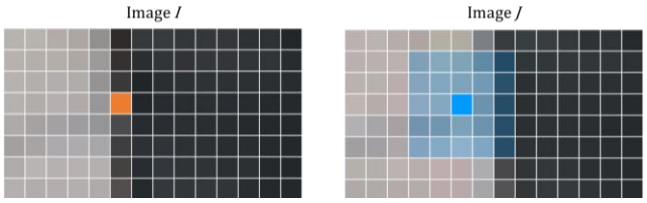


Fig. 2. Calculation of IV-MSE, Eq. (5) vs. typical MSE, Eq. (3). MSE: orange pixel in image ***I*** is compared to the colocated opaque blue pixel in image ***J***; IV-MSE: orange pixel in image ***I*** is compared to all blue pixels in image ***J*** (5×5 neighborhood of the colocated pixel), the difference is calculated between the value of the orange pixel and the most similar pixel within the blue block.

This problem can be noticed especially on the edges of objects. For example, in the left part of Fig. 2, a small, significantly magnified fragment of the input view (*Museum* sequence [58]) is presented. The fragment contains pixels representing the background (floor, brighter fragment) and a person in the foreground (dark part). The right part of Fig. 2 contains a co-located fragment of the virtual view synthesized at the same position. Despite the *Museum* being a CG sequence, so its depth maps do not contain any artifacts, the reprojection caused the edge between the floor and the person to be shifted by more than one pixel. However, although edge shifting decreases the quality of pixel-wise quality metrics (e.g., PSNR),

it is practically unnoticeable by the viewer, because the viewer is not aware of the exact position of each object. Therefore, the quality metric for immersive video should be not sensitive to these small, unnoticeable for the viewer, reprojection-induced errors [117]. Naturally, a large shift in the position of an object in the virtual view can be easily recognized as an error (e.g., an object significantly shifted because of wrong depth, Fig. 3).

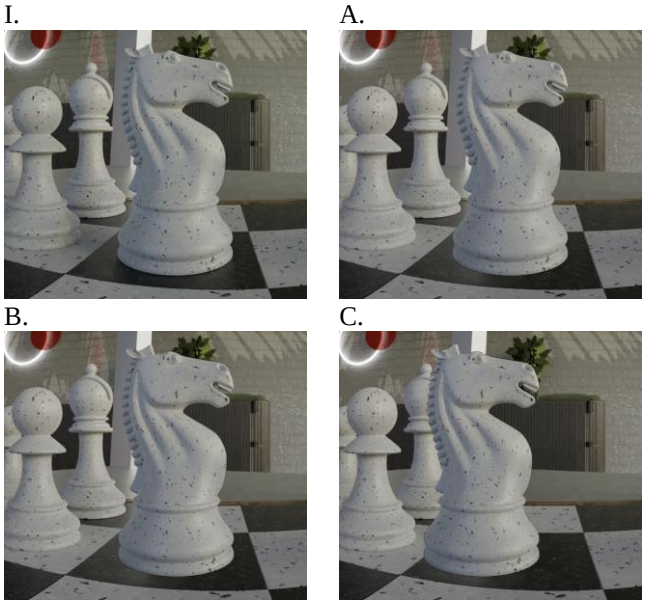


Fig. 3. Noticeability of pixel shift; (I) input view, (A – C) synthesized view: (A) correct position of the knight, (B) knight shifted by 2 pixels to the right, (C) knight shifted significantly, *Chess* sequence [115].

B. Global Color Offset Error

In order to provide a better experience for the viewer, inter-view inconsistencies should be reduced using color correction methods (e.g., [9], [61]). Using the color correction causes the virtual views to contain fewer color artifacts (Fig. 4), having better subjective quality.

Fig. 4. Virtual view synthesized using views captured by cameras with different color characteristics (A) and color-corrected views (B), vs. input view (I), *SoccerArc* sequence.

Color correction may cause the global color characteristics of the view (e.g., its overall brightness) to be wrong. On the other hand, a slight change of colors of the entire view is barely noticeable by the viewer, but significantly changes the values of some metrics, which simply calculate the difference between values of pixels with two images.

III. Related Works

This section provides an overview of related works on image quality assessment. A particular emphasis is put on the usefulness of presented methods in immersive video applications, expressed as their adaptation to the typical distortions described in the previous section.

A. PSNR-based Methods

The widest group of available image quality assessment methods are based on the calculation of PSNR. Such methods have a very great advantage over other metrics, as they are commonly used in a broad spectrum of image-related research (e.g., compression), making them very intuitive to use for researchers. Some PSNR-based methods were enhanced to be more useful in immersive video, e.g., WS-PSNR [13], CPP-PSNR [14], and OV-PSNR [15]. These methods take into account the possibility of tested images being omnidirectional, enabling a better direct comparison of ERP videos in immersive media coding. However, such a comparison is not very practical, as such final video generated to the viewer after decoding is a regular, perspective video.

PSNR-HVS [16] is based on the assumption that the human visual system (HVS) is more sensitive to low-frequency distortions. Calculations are performed using DCT coefficients that were modified using quantization tables from the JPEG encoder. However, the correlation with MOS for compression-induced errors is the same as in PSNR. Adding the contrast sensitivity function into account in PSNR-HVS-M [17] increased the correlation with MOS when Gaussian noise and spatially correlated additive Gaussian noise were present in tested images. Further modifications of the abovementioned methods, that were implemented in PSNR-HA [18] and PSNR-HMA [18], also include mean level shift compensation and contrast stretching. CS-PSNR [19] uses new weighting coefficients derived from extensive subjective tests in order to measure the final quality considering the characteristics of color sensitivity of HVS.

Similar methods are often based on SNR calculation, e.g., SRE (signal to reconstruction error ratio [20]), which measures the error relative to the power of the signal, not the constant peak intensity, which better mimics the subjective quality assessment for images with different brightness. The usability of SNR-based methods is dependent on the desired application. In the case of immersive media, large differences in brightness should be considered as an error that influences the subjective quality of the image. Other SNR-based methods are, e.g., WSNR [21] (SNR calculated in the frequency domain is additionally weighted, reducing the influence of frequencies less important for HVS), or VSNR [22] – wavelet-based SNR.

B. SSIM-based and Related Methods

Another important group of metrics, based on SSIM (structural similarity image assessment [23]), focuses on the extraction of structural information from compared images, which follows the hypothesis that the human visual system is highly adapted to changes in this domain. These methods, which expand the previous UIQ method [24], provide relatively good performance when compared to PSNR.

During the calculation of SSIM, 3 types of information are considered, i.e., the luminance of compared images, their contrast, and structure. MS-SSIM [25] additionally performs a comparison of contrast and structure on multiple scales of tested images. Unfortunately, the usefulness of these methods is limited in the evaluation of synthesized videos. In such videos, possible small shifts can be introduced by the synthesis process. While such small distortions are not significant for the viewer of the video, they can still strongly influence the SSIM score (as well as other methods based on the comparison of structure, e.g., the method [26] based on gradient similarity). Other relevant SSIM-based metrics are, e.g., FSIM [27], which focuses on low-level features of images [28], [29], or SSRM [30], which performs the comparison in the frequency domain.

Frequency-domain-based assessment is also performed in SAM [31]. This metric is calculated from the angle between the spectra of two images. While being independent of different brightness of compared images, even unnoticeable spatial shifts between them highly decrease the final estimated score.

C. Methods Based on Trained Models

Examples of full-reference metrics focused on deriving models from natural scene statistics are VIF [32], in which images are decomposed in different subbands to separate the information of the source image from its distortions, and SFF [33], which is based on a comparison of sparse features derived from detectors trained on samples of natural images. While such an approach is designed to simulate the properties of HVS [34], the evaluation of the quality of CGI-based sequences, commonly used in immersive video research, can be disturbed because of their different visual characteristics.

VIF, together with DLM [35] (also a wavelet-based metric), was combined in the SVM-regression-based [36] VMAF method [37] (and its expansion for omnidirectional images [38]). It also introduces a simple temporal difference analysis that increases the correlation with the subjective quality of videos. Similarly to other machine-learning-based methods (e.g., LPIPS [39]), the performance of this metric is directly dependent on the pre-trained model, which should be different for various applications.

The abovementioned metrics were intended to be versatile, and thus were not designed to measure distortions caused only by specific types of errors. However, the characteristics of distortions present in virtual views generated through the synthesis process, as described in Sections II-A and II-B, differ from distortions induced by other processing.

The method presented in [113] had been shown to be not sensitive to offset present in a tested image, also for compressed images. However, especially for highly compressed images, the large offsets do not change the measured distortion (what should not be a case in immersive video, as discussed in Section II-A). Moreover, the offset is added to a whole image, in contrast to the reprojection-related offset (error), which changes the position of all points of the image independently.

D. Metrics for Synthesized Video

Many methods focus mainly on the correct evaluation of the quality of synthesized video. For example, the method described in [40] assesses the quality by penalizing pixels that belong to non-consistent contours. It allows the edges to not be exactly in the same places in both compared images, so small shifts of contours induced by depth-image-based rendering (DIBR) do not influence the final score.

Method [41] performs an additional exhaustive search between blocks of evaluated images, so the further DWT-based comparison is resilient to DIBR-induced shifts. However, the search heavily increases the computational complexity of the metric. The relatively high complexity can also be seen in the LOGS metric [42], which uses SIFT to find disoccluded areas in the synthesized view. PSNR-based methods are much more efficient in terms of complexity, e.g., MP-PSNR and MW-PSNR [43]. These methods introduce a multiscale comparison that uses morphological filters to focus on the geometrical distortions of the synthesized video.

E. No-reference Quality Metrics

A very interesting class of metrics does not require any reference image to assess the quality. Such no-reference metrics seem to be very useful for evaluating the quality of the virtual view, as such view does not have to be placed in the same position as any available input view. Unfortunately, such methods often focus on the quality of stereoscopic video only (e.g., [44], [45], [116]), or monoscopic and stereoscopic omnidirectional systems (e.g., [46], [47], [108]), which are very specific types of non-fully immersive video. Other very application-specific metrics are described in [48], [49], as they are intended to be used for light-field images only.

Other methods do not state any assumption on the type of tested video [50], [51], but do not take into account compression-induced distortions. Method [52] measures errors caused by both compression and the virtual view, so it is well adjusted to the characteristics of immersive video. Unfortunately, evaluating the quality of video also requires a depth map to be present. It highly decreases the versatility of such methods, as the evaluation has to be performed before the virtual view synthesis is performed. A similar requirement is also stated in other methods, e.g., [53], [54]. A no-reference method that is versatile and adapted to the characteristics of immersive video is NR-MWT [55], however, it performs best when the input data for view synthesis is uncompressed. Similarly, an interesting no-reference method [56] cannot be efficiently used when input views are compressed, as it would highly influence the results of its internal JPEG compressor which is used to estimate the complexity of the input frame. The recent method DoC-DoG-GRNN [118] is based on a series of defined morphological operators which are used to extract features that are fed into a trained neural network. It shows

a high correlation with subjective quality for datasets which include both synthesis and compression-induced errors. However, re-training of a GRNN is required for each dataset, as the proposed model can fail without re-training, especially in the evaluation of high-quality synthesized views.

IV. IV-PSNR

The proposed IV-PSNR metric contains two techniques that allow for dealing with typical immersive video distortions: the corresponding pixel shift (Section IV-A), and global component difference (Section IV-B). Finally, the full description of IV-PSNR calculation, which implements these two techniques, is included in Section IV-C.

A. Proposed Solution for Corresponding Pixel Shift

As IV-PSNR tries to simulate the perceived quality of immersive video, it is insensitive to slight shifts of the edges due to the modified mean square error calculation, Eq. (5). For each analyzed pixel of image I, the error is estimated as the difference between that pixel and the most similar pixel within a colocated block in image J.

For example, the error for the orange pixel in Fig. 2 will be estimated as a difference between the orange pixel (representing the person in the foreground) and the most similar of the semi-transparent blue pixels on the right side of the 5×5 block. When calculating PSNR or another pixel-wise metric it would be compared to the opaque blue pixel, which represents the background in this example.

As presented in Table I, IV-PSNR allows to properly assess that the subjective similarity between images I and C is much lower than between I and A. On the other hand, a slight shift of the knight in image B, which is unnoticeable to the viewer, does not negatively impact the IV-PSNR value (compare IV-PSNR(I, A) and IV-PSNR(I, B) in Table I).

TABLE I OBJECTIVE QUALITY OF FRAGMENTS PRESENTED IN FIG. 3.

Compared images, Fig. 3	I, A	I, B	I, C
$PSNR_Y$	41.78 dB	39.41 dB	29.90 dB
IV-PSNR	47.23 dB	47.22 dB	37.61 dB

The size of the block was set to 5×5 to handle shifts smaller than 2 pixels, which is less than 0.1% of the image width in HD video. As presented in Table II, such a size provides a high correlation with MOS while preserving reasonable computational time.

TABLE II CORRELATION BETWEEN IV-PSNR AND MOS, AND COMPUTATIONAL TIME FOR DIFFERENT BLOCK SIZES; RESULTS OBTAINED USING METHODOLOGIES DESCRIBED IN SECTIONS VII AND VIII; COMPUTATIONAL TIME ESTIMATED USING THE METHODOLOGY DESCRIBED IN SECTION X, SINGLE-THREADED IMPLEMENTATION.

Max shift		0	1	2	3	4	5	6	7
Block size		1×1	3×3	5×5	7×7	9×9	11×11	13×13	15×15
SROCC	VII	0.585	0.718	0.729	0.730	**0.732**	0.728	0.727	0.722
	VIII	0.502	0.558	**0.565**	0.495	0.469	0.458	0.456	0.448
Time [s]		**0.19**	0.65	1.40	2.56	4.01	5.65	7.68	9.84

Moreover, the described pixel shift compensation is also beneficial in the assessment of compression-induced errors. Compression can lead to blurring or even an edge shift in the image, especially in highly compressed videos because of the high share of temporally predicted parts of videos [59]. As it was proven in [60], the proposed modification of MSE is useful in the assessment of similarity between compressed views.

B. Proposed Solution for Global Component Difference

IV-PSNR takes the abovementioned phenomenon into account by excluding the influence of slight changes in the global characteristics of the image. For each component, the average difference between colocated pixels of two images is calculated – Eq. (6). Then, this difference is considered when calculating the square error for every pixel of the image, Eq. (5).

In Fig. 4, the comparison of two synthesized images (A and B) is presented. Image A was synthesized using inconsistent input views, image B – using views corrected by [62]. Subjectively, image B is better because of fewer color artifacts. However, as the image is slightly brighter than the input view, the degradation of the $PSNR_Y$ value can be observed. On the other hand, sharp-edged fragments of the pitch visible in image A have a color more similar to the fragments of the input view I, causing a higher value of $PSNR_Y$. As presented in Table III, IV-PSNR mimics the subjective perception of the quality much better than $PSNR_Y$.

TABLE III OBJECTIVE QUALITY OF FRAGMENTS PRESENTED IN FIG. 4.

Compared images, Fig. 4	I, A	I, B	difference
$PSNR_Y$	26.76 dB	26.62 dB	–0.14 dB
IV-PSNR	32.22 dB	32.93 dB	+0.71 dB

C. Calculation of IV-PSNR

The proposed IV-PSNR metric is based on the Peak Signal-to-Noise Ratio (PSNR), which is the most widespread and commonly used metric in all image or video processing applications and can be described as fast, robust, easily implementable, and interpretable. PSNR for component c of an image is calculated as:

$$PSNR_c = 10 \cdot log_{10}\left(\frac{(2^b-1)^2}{MSE_c}\right), \quad (2)$$

where b is the bit-depth of the image and MSE_c is the mean square error between component c of images I and J:

$$\mathrm{MSE}_c = \frac{1}{W \cdot H}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left(I_c^{x,y} - J_c^{x,y}\right)^2, \quad (3)$$

where H is the height of both images being compared, and W is their width (in pixels).

When compared to PSNR, the proposal includes two major modifications to adapt to two major distortions typical for immersive video: corresponding pixel shift (Section IV-A), and global component difference (Section IV-B).

The calculation of quality (or, more precisely, similarity) is performed independently for each component c, using the equation analogous to (2):

$$IV\text{-}PSNR_c^{I\to J} = 10 \cdot log_{10}\left(\frac{(2^b-1)^2}{IV\text{-}MSE_c^{I\to J}}\right). \quad (4)$$

While the idea of calculation of quality as a logarithm of mean square error is the same, the mean square error evaluation itself is different, and the value of IV-MSE$_c^{I\to J}$ is calculated in a pixel-to-most-similar-pixel-within-a-block manner, taking into account the corresponding pixel shift (cf. Section IV-A):

$$IV\text{-}MSE_c^{I\to J} = \frac{1}{W\cdot H}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\min_{\substack{w\in[x-B,\,x+B]\\ h\in[y-B,\,y+B]}}\left(I_c^{x,y} - J_c^{w,h} + GCD_c^{I\to J}\right)^2, \quad (5)$$

where B is the maximum considered shift of the corresponding pixel, by default set to 2 (thus the most similar pixel within a 5×5 block is being searched, cf. Fig. 2), and GCD$_c^{I\to J}$ is the global component difference (cf. Section IV-B) between component c of images I and J (averaged over entire image):

$$GCD_c^{I\to J} = \frac{1}{W\cdot H}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left(I_c^{x,y} - J_c^{x,y}\right), \quad (6)$$

As stated in [13], for omnidirectional images, where a 3D sphere is projected onto a 2D image (e.g., by ERP projection [63]), the latitude of each pixel should be considered when evaluating the quality in order to simulate the quality perceived by a viewer watching the scene, e.g., using a head-mounted display [64]. Therefore, for ERP video, the square error calculated for each pixel is additionally weighted similarly as in [13]:

$$IV\text{-}MSE_c^{I\to J} = \frac{\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\min_{\substack{w\in[x-B,\,x+B]\\ h\in[y-B,\,y+B]}}\left(I_c^{x,y} - J_c^{w,h} + GCD_c^{I\to J}\right)^2\cdot w_{x,y}}{\sum_{y=0}^{H-1}\sum_{x=0}^{W-1} w_{x,y}}, \quad (7)$$

where $w_{x,y}$ is the weight (different for each row of the images):

$$w_{x,y} = cos\frac{\left(y+0.5-\frac{H}{2}\right)\cdot\pi}{H}\cdot\frac{AOV_V}{180°}, \quad (8)$$

where AOV$_V$ is the vertical angle of view of the ERP camera (for a fully spherical camera, it is equal to 180°).

In order to produce one value, comprising the quality of all the components, IV-PSNR values calculated for all components are combined using the weighted average:

$$IV\text{-}PSNR_{YUV}^{I\to J} = \frac{\text{IV-PSNR}_Y^{I\to J}\cdot w_Y + \text{IV-PSNR}_U^{I\to J}\cdot w_U + \text{IV-PSNR}_V^{I\to J}\cdot w_V}{w_Y + w_U + w_V}. \quad (9)$$

By default, the weight for the luma component (w_Y) is set to 4, while the weights for both chroma components (w_U and w_V) are set to 1, as in the most commonly used chroma subsampling format (4:2:0 [65]) there are four luma samples for a sample of each chroma.

D. Symmetricity of IV-PSNR

It should be emphasized, that the proposed definition of IV-MSE is asymmetrical, thus:

$$IV\text{-}MSE_c^{I\to J} \neq IV\text{-}MSE_c^{J\to I}. \quad (10)$$

However, a robust objective quality metric has to be symmetrical. Therefore, to provide symmetricity, the final value of the IV-PSNR between images I and J is calculated as:

$$IV\text{-}PSNR(I,J) = min\left(IV\text{-}PSNR_{YUV}^{I\to J}, IV\text{-}PSNR_{YUV}^{J\to I}\right). \quad (11)$$

Such an approach allows us to properly assess the quality, even if one of the compared images contains small but noticeable artifacts, e.g., heavy salt-and-pepper impulse noise [66] presented in Fig. 5.

For example, if image J has such noise, the noised pixels would be never used when calculating the IV-MSE$_c^{I\to J}$ (for each component c they would be skipped because more similar pixels are available within a 5×5 block) affecting the too high value of IV-PSNR$_{YUV}^{I\to J}$, but IV-MSE$_c^{J\to I}$ and thus IV-PSNR$_{YUV}^{J\to I}$ will be much lower, indicating a lower similarity between the compared images (cf. Fig. 5 and Table IV).

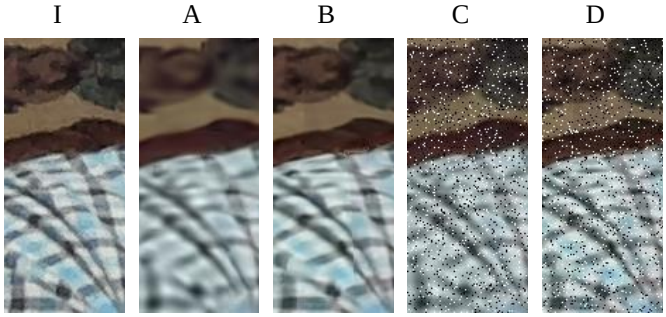


Fig. 5. I: Fragment of input video (sequence *Frog* [83]), A: fragment I compressed by VVC [112] with QP = 43, B: the same fragment of the synthesized view, C: fragment A with heavy salt-and-pepper impulse noise, D: fragment B with heavy salt-and-pepper impulse noise.

TABLE IV OBJECTIVE QUALITY OF FRAGMENTS PRESENTED IN FIG. 5.

	Distortion type	PSNR$_Y$	IV-PSNR$_{YUV}^{I\to J}$	IV-PSNR$_{YUV}^{J\to I}$	**IV-PSNR**
A	*Compression*	28.83	37.57	41.62	**37.57**
B	*View synthesis*	29.20	39.99	42.32	**39.99**
C	*Comp. + noise*	15.43	37.46	24.77	**24.77**
D	*Synthesis + noise*	15.45	39.82	25.40	**25.40**

As presented in Table IV, the IV-PSNR metric allows us to properly assess the quality of the synthesized or compressed video, even if the video contains a high amount of noticeable artifacts.

V. SOFTWARE IMPLEMENTATION

A. Overall

The reference implementation of the IV-PSNR algorithm is available as *free and open-source software* [67] (under the 3-clause BSD license) called "IV-PSNR software".

The software has been written in modern C++17 language [68] and is designed to be fast and reliable. We have put significant effort into algorithmic optimization and parallelization, including data-level parallelism (SIMD) and multithreading.

The IV-PSNR software was implemented by the authors of this paper and is publicly available in the public git repository

of the MPEG Immersive video group: https://gitlab.com/mpeg-i-visual/ivpsnr.

The software is designed to work with a simple lossless raw format called "yuv" commonly used in research on video compression [6]. The "yuv" format is mostly used to store video sequences in the YCbCr or RGB color space which allows for different chroma formats (4:4:4, 4:2:2, and 4:2:0) and bit depths up to 16 bit/sample. The software can process planar "yuv" files without any limitation to the size and number of frames. All common chroma formats (4:4:4, 4:2:2, and 4:2:0) are accepted and bit depth in the range 8-14 is supported.

The software processes each frame independently and calculates its IV-PSNR metric value. For video sequences, the IV-PSNR metric calculated for each frame is stored and the averaged IV-PSNR value for the entire sequence is calculated. In addition to the IV-PSNR metric, the software calculates PSNR and WS-PSNR [13] metrics.

B. Chroma Subsampling Issues

The IV-PSNR corresponding pixel shift (see Section IV-A) requires per-pixel access to all components. Unfortunately, some of the formats used in video compression (and video transmission) use decimated chroma components. This approach leads to chroma sampling schemes such as 4:4:4, 4:2:2, and 4:2:0. The 4:2:2 and 4:2:0 introduce chroma subsampling and result in a single chroma sample covering two or four luma samples respectively [69]. This makes the corresponding pixel shift calculation difficult.

To avoid the abovementioned issues, the 4:4:4 format is used internally to uniform the sizes of all components in all calculations. Therefore, input data in the 4:2:2 and 4:2:0 formats must be converted (interpolated) to the 4:4:4 format. This leads to a question if the calculation of PSNR-based quality metrics on an interpolated picture alters the metric value. We investigated this problem by performing a detailed analysis of the PSNR metric with the assumption that the 0-order (nearest neighbor) interpolation is used. The 0-order interpolation is the simplest approach for upscaling a picture by integer factor (applied for chroma upsampling). Moreover, this technique does not produce nonexistent pixel values and is very fast to compute.

As described in (2), the derivation of the PSNR metric requires the calculation of the mean square error (MSE). The equation describing the MSE metric (3) could be directly used to calculate luma MSE (MSE_Y).

In the case of MSE for chroma components (MSE_C) the width and/or height can be different depending on the selected chroma subsampling scheme [65]. Therefore, chroma (both U/Cb and V/Cr) can be calculated as follows:

$$MSE_C = \frac{1}{W_C \cdot H_C} \sum_{y=0}^{H_C-1} \sum_{x=0}^{W_C-1} \left(I_C^{x,y} - J_C^{x,y}\right)^2, \qquad (12)$$

where, H_C, W_C are chroma height and width, respectively.

Let us consider 4:2:0 chroma subsampling. In such a case both chroma components are subsampled by a factor of 2 leading to $W_C = W_Y/2$ and $H_C = H_Y/2$. As mentioned above this results in difficulties in the corresponding pixel shift step of IV-PSNR calculation. To make IV-PSNR calculation possible, we decided to apply 0-order interpolation to all chroma components and investigate if the PSNR-like metric can be calculated on interpolated components without result altering.

The interpolated chroma component has the same size as the luma component, therefore, the equation describing MSE_C calculation has to be modified in a way, that MSE is calculated for even rows and even samples (in a row) only. The following equation describes the calculation of chroma MSE using an interpolated image (MSE_{CI}):

$$\mathrm{MSE}_{CI} = \frac{1}{\frac{W_Y}{2} \cdot \frac{H_Y}{2}} \sum_{y=0}^{\frac{H_Y}{2}-1} \sum_{x=0}^{\frac{W_Y}{2}-1} \left(I_{CI}^{2\cdot x,2\cdot y} - J_{CI}^{2\cdot x,2\cdot y}\right)^2. \qquad (13)$$

Since the 0-order interpolation was used, the interpolated chroma component contains quads of pixels with the same value. Therefore, the following relation is true for every quad of pixels:

$$\left(I_{CI}^{2\cdot x,2\cdot y} - J_{CI}^{2\cdot x,2\cdot y}\right)^2 = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{\left(I_{CI}^{2\cdot x+j,2\cdot y+i} - J_{CI}^{2\cdot x+j,2\cdot y+i}\right)^2}{4}, \qquad (14)$$

Equation (13) can be modified by replacing the sum term by the formula from (14) and simplified:

$$\mathrm{MSE}_{CI} = \frac{1}{W_Y \cdot H_Y} \sum_{y=0}^{\frac{H_Y}{2}-1} \sum_{x=0}^{\frac{W_Y}{2}-1} \sum_{i=0}^{1} \sum_{j=0}^{1} \left(\mathrm{I}_{CI}^{2\cdot \mathrm{x}+\mathrm{j},2\cdot \mathrm{y}+\mathrm{i}} - \mathrm{J}_{CI}^{2\cdot \mathrm{x}+\mathrm{j},2\cdot \mathrm{y}+\mathrm{i}}\right)^2. \qquad (15)$$

Equation (15) contains two groups of nested sums. The outer group (over y and x) corresponds to the loop over all pixel quads. The inner group (over i and j) corresponds to a loop over all pixels within a quad. The nested sum in (15) describes the processing of all chroma pixels and can be further simplified:

$$\mathrm{MSE}_{CI} = \frac{1}{W_Y \cdot H_Y} \sum_{y=0}^{H_Y-1} \sum_{x=0}^{W_Y-1} \left(I_{CI}^{x,y} - J_{CI}^{x,y}\right)^2 \qquad (16)$$

The resulting equation (16) for the calculation of MSE_{CI} for 0-order interpolated chroma component corresponds to MSE_C (12). Therefore, chroma components can be interpolated (assuming 0-order interpolation) and PSNR for 4:2:0 (and 4:2:2) can be calculated in the 4:4:4 internal representation without any influence on the final result.

C. Rounding Error Reduction

The aim of IV-PSNR (both metric and software) is to measure image quality. Therefore, as a measurement tool, the software implementing the IV-PSNR metric has to be precise and produce reproducible results. Consequently, significant effort has been made to avoid (or at least reduce) any errors related to floating-point computations, while preserving high performance and low computational complexity.

Since the IV-PSNR software operates on input images with pixels represented as integer values, a significant part of computations is performed in the integer numbers domain, effectively eliminating rounding and accumulation errors.

The calculation (6) of the global component difference (GCD_c) can be separated into two steps. The first step is the calculation of the cumulative component difference ($\mathrm{CCD}_c^{I\to J}$). This step can be performed in the integer numbers domain to avoid accumulation errors:

$$CCD_c^{I\to J} = \sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left(I_c^{x,y} - J_c^{x,y}\right). \tag{17}$$

The second step is processed in the floating-point domain:

$$GCD_c^{I\to J} = \frac{CCD_c^{I\to J}}{W\cdot H}. \tag{18}$$

Moreover, the row-level error is also calculated in the integer domain. The IV-$\mathrm{MSE}_{I\to J}^{c}$ the calculation presented in (7) can be divided into three steps presented as follows:

$$RowIV\text{-}SSD_c^{I\to J}(y) = \sum_{x=0}^{W-1}\min_{\substack{w\in[x-B,\,x+B]\\ h\in[y-B,\,y+B]}}\left(I_c^{x,y} - J_c^{w,h} + GCD_c^{I\to J}\right)^2, \tag{19}$$

$$IV\text{-}SSD_c^{I\to J} = \sum_{y=0}^{H-1} RowIV\text{-}SSD_c^{I\to J}(y)\cdot w_{x,y}, \tag{20}$$

$$IV\text{-}MSE_c^{I\to J} = \frac{IV\text{-}SSD_c^{I\to J}}{W\cdot H\cdot\sum_{y=0}^{H-1}\sum_{x=0}^{W-1} w_{x,y}}. \tag{21}$$

$\mathrm{RowIV\text{-}SSD}_c^{I\to J}(y)$ (corresponding to the sum of squared differences calculated for image row) is calculated entirely in the integer numbers domain. The $w_{x,y}$ weight is not an integer value, therefore, the remaining calculations have to be performed in the floating-point domain. To reduce floating-point accumulation related errors, the calculation of $\mathrm{IV\text{-}SSD}_c^{I\to J}$ is performed in two steps. First, each intermediate $\mathrm{RowIV\text{-}SSD}_c^{I\to J}(y)\cdot w_{x,y}$ value is stored in a dedicated buffer. Second, after processing all rows, the buffered values are accumulated with the use of the Kahan-Babuška-Neumaier Summation (KBNS) [70] algorithm.

The same approach is used to calculate the average IV-PSNR for the entire sequence. IV-PSNR metrics for each picture are stored in the buffer and accumulated using KBNS.

The abovementioned approach results in IV-PSNR software increased robustness against accumulation errors in the case of very long sequences.

D. Implementation Details and Performance Optimization

Besides internal use of the 4:4:4 format (regardless of chroma format of input data), planar data read from "yuv" are converted to interleaved format, which improves memory locality and cache consistency. Some of the data processing routines are implemented in the standard C++ language and using SSE4.1 and AVX2 SIMD instructions. The SIMD optimized routines are automatically used if the software is built for compatible architecture (x86-64-v2 or higher level) [71].

IV-PSNR implementation includes thread-level parallelism with the use of dedicated, low overhead thread pool implementation. Computations are parallelized at the image row level, therefore a high level of parallelism is achievable.

VI. Overview of Experiments

The IV-PSNR metric was compared with 31 state-of-the-art quality metrics with publicly available implementations.

Among all tested metrics there were typical (2D) image and video quality metrics: 4 metrics implemented together with proposed IV-PSNR in [72]: $\mathrm{PSNR_Y}$ (PSNR for luma component), $\mathrm{PSNR_{YUV}}$ (weighted average of PSNR for 3 components with luma weight 6 times higher than the weights for both chroma components, as described in [73]), $\mathrm{WS\text{-}PSNR_Y}$ [13], and $\mathrm{WS\text{-}PSNR_{YUV}}$ (weighted in the same way as $\mathrm{PSNR_{YUV}}$); CS-PSNR [19]; VMAF [37]; SSIM [23] and a multiscale version of SSIM – MS-SSIM [25]; a pixel-based version of VIF (VIF-P) [32]; PSNR-HVS [16]; PSNR-HVS-M [17]; SFF [33]; PSNR-HA and PSNR-HMA [18]; VSNR [21]; WSNR [22]; 4 metrics implemented in [74]: SAM [31], SRE [20], FSIM [27], and UIQ [24].

Moreover, we have compared several metrics designed for 3D video: MP-PSNR in two variants, full – MP-PSNR-F [75] and reduced – MP-PSNR-R [76]; 2 variants of MW-PSNR: full – MW-PSNR-F [77] and reduced – MW-PSNR-R [43]; 3DSwIM [41]; CPP-PSNR [14]; LPIPS [39]; and 3 variants of OV-PSNR [15]: based on PSNR, CPP-PSNR, and WS-PSNR. Besides the above full-reference metrics, also the no-reference metric created for the assessment of synthesized video – NR-MWT [55] was compared.

In order to provide a valid comparison of all considered metrics, three experiments were performed. In the first one (Section VII), the influence of different artifacts induced by immersive video encoding was assessed, including both the artifacts introduced by conventional HEVC video compression and by compression based on inter-view redundancy removal.

In the second experiment (Section VIII), the metrics were evaluated on several aspects of immersive video processing, including color correction, depth map filtering, and the influence of different synthesizers.

In the third experiment (Section IX), all metrics were evaluated against the commonly-used IQA databases – TID2013 [78] and CVIQ [108]. In this experiment, the effectiveness of IV-PSNR in assessing the quality in non-immersive-video applications was evaluated.

In each experiment, the metrics were compared using two commonly used correlation coefficients, which allow for

assessing the monotonicity of the relationship between objective and subjective quality: Spearman and Kendall rank-order correlation coefficients: SROCC and KROCC [79]. In Section X, the computational time of all tested metrics was evaluated.

VII. Effectiveness for Different Immersive Video Encoding Techniques

A. Overview of the Experiment

In the first experiment, the effectiveness of the IV-PSNR metric was assessed using the results of the "MPEG Call for Proposals on 3DoF+ Visual" [80], purposed to evaluate proposed techniques of coding the immersive video content.

In general, N input views with corresponding depth maps and camera parameters are preprocessed by the immersive video coding technique (orange block in Fig. 6). It generates n videos with corresponding depth maps (where $n \leq N$) together with metadata, which allows for restoring the entire input information. Each of n texture videos and each of n depth videos is independently encoded using the HEVC encoder. In the end, all the data are packed into one bitstream.

At the decoder side, the bitstream is unpacked into n texture video streams and n depth video streams. After HEVC decoding, the immersive video decoding is performed as a postprocessing step in order to produce a final video presented to the viewer (e.g., any of the input views or views on a virtual trajectory of the viewer).

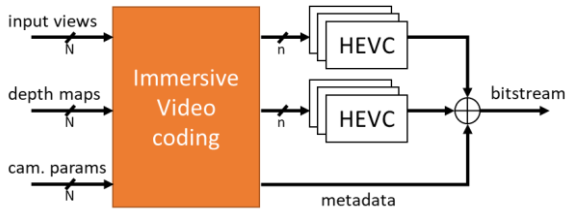


Fig. 6. Multiview video coding scheme used for "MPEG Call for Proposals on 3DoF+ Visual" [80]. ***N*** views with corresponding depth maps and camera parameters are preprocessed in the immersive video coding step (orange block).

B. Methodology

For each coding technique the same output (decoded) videos were generated:

- two "posetrace" videos – videos containing views synthesized on the preset virtual trajectory of the theoretical viewer of the immersive video system (Fig. 7),
- a subset of selected input views (4 for *ClassroomVideo* and *Painter*, 3 for *Hijack* and *Frog*, and 6 for *Museum*).

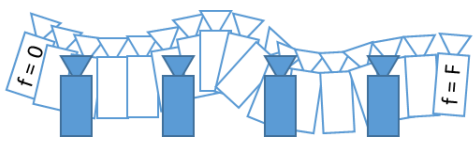


Fig. 7. The idea of the posetrace; blue – input views, white – the posetrace build of virtual views, containing F consecutive frames.

The duration of each video was 300 frames except for the *ClassroomVideo* sequence where only 120 frames are available.

The subjective quality evaluation was performed in EVATech and GBTech laboratories using the DSIS (Double Stimulus Impairment Scale) test method [90]. In total, 18 naïve viewers participated in the tests [91].

The participants were assessing the quality of the posetrace videos, simulating the case of virtual navigation within a scene. Each participant assessed the quality of 280 videos (4 rate points × 5 test sequences × 7 coding techniques × 2 posetraces).

The subjective quality evaluation was performed on a 2D screen by viewing posetraces instead of using VR headsets because, as described in [1], assessment with head-mounted displays has two flaws: it is more time consuming and each participant may arbitrarily change their viewpoint, making results from different participants incomparable. Moreover, as indicated in [92], the quality of a full omnidirectional view can be inferred from the quality of selected viewports

As presented in Fig. 7, the posetrace contains a set of virtual views, generated between the input ones. When using the full-reference quality metrics, it is not possible to evaluate the objective quality on the posetrace because of lack of the reference views. Therefore, for the calculation of the objective quality, the quality of the synthesized input views was calculated.

For each of the 280 test points, the Mean Opinion Score (MOS) was calculated. Then, MOS for two posetraces generated for the same rate point, sequence, and coding technique were averaged resulting in 140 test points. For all of them, the objective quality was evaluated. For each test point, the one value of each quality metric was obtained by averaging over all frames (300 or 120) and all synthesized input views (3, 4, or 6, depending on the sequence).

C. Dataset

In total, 7 compression techniques were compared. The videos encoded using each technique targeted in the same bitrates. For each test sequence, four target bitrates were chosen in order to compare the techniques in various conditions. The total bitrate should not exceed 6.5 Mbit/s for the first rate point (highest compression), and 10, 15, and 25 Mbit/s for lower compression rate points. All techniques were compared using a test set containing five multiview test sequences (Fig. 8), described in Table V, allowing efficiency assessment for the 3DoF+ scenario [109] (three omnidirectional ERP sequences) and also windowed-6DoF applications [109] (two sequences captured by multiview systems equipped with multiple perspective cameras).

Two of the techniques being compared were "anchors" [80]. The anchors were prepared using a previously existing video coding technique, i.e., HEVC simulcast. In anchor A, all the

input views and corresponding depth maps were independently encoded using HEVC. Of course, it implied very high compression required to fit within the bitrate limits. In this case, the orange block in Fig. 6 was just passing the input data and $n = N$. Anchor B was defined differently, and only the subset of input views and corresponding depth maps was encoding using HEVC simulcast. Due to the lower number of videos, the compression could be much lower, than for anchor A. On the other hand, when some views were skipped the disocclusions problem occurred (because the information from these views was not sent at all). The number of views being sent varied for different sequences: 9 for *ClassroomVideo*, 8 for *Museum*, 5 for *Hijack*, 8 for *Painter*, and 7 for *Frog*. The views were selected manually.

TABLE V TEST SEQUENCES USED FOR COMPARISON OF VARIOUS COMPRESSION TECHNIQUES. CG – COMPUTER GENERATED, NC – NATURAL CONTENT, ERP – EQUIRECTANGULAR PROJECTION, PERSP. – PERSPECTIVE VIEWS.

Sequence	Type	Source	Resolution	Input views
ClassroomVideo	CG/ERP	[81]	4096 × 2048	15
Museum	CG/ERP	[58]	2048 × 2048	24
Hijack	CG/ERP	[58]	4096 × 4096	10
Painter	NC/Persp.	[82]	2048 × 1088	16
Frog	NC/Persp.	[83]	1920 × 1080	13

Fig. 8. Test sequences. Top row (from left): Frog, Museum, and Hijack; bottom row (from left): Painter and ClassroomVideo.

Besides two anchors, 5 compression techniques proposed by different organizations were compared. List of proposals contained techniques proposed by (in alphabetical order): Nokia [84], Philips [85], PUT (Poznan University of Technology) and ETRI (Electronics and Telecommunications Research Institute) [86], Technicolor and Intel [87], and ZJU (Zhejiang University) [88].

The proposals were based on different approaches. One of the proposals was based on the processing of the point cloud containing projected pixels from all the views. In two of the proposals, several input views were chosen as base views and sent in their entirety, while other views were pruned in order to preserve only the non-redundant information. Another approach was to use the data from all the input views to synthesize the base view with the field of view high enough to contain information from the entire scene.

Different methods had their advantages and disadvantages. For example, if a subset of input views was chosen as the base ones, the quality of views presented to the viewer significantly differs depending on his/her position. If the viewer watches the scene from the position close to the base view, the quality is much higher than if the scene is watched from other viewpoints. On the other hand, when the base view is synthesized, the quality is more stable when virtually moving among the scene, but the peak quality is worse due to the reprojection artifacts.

The proposals have also different approaches for handling the preserved, non-redundant information from other (non-base) views. In some of them, various packing algorithms were used to organize the data from other views as a mosaic of patches, allowing for decreasing the pixel rate of the video [89]. In others, instead of packing, the non-redundant information from other views was stored as additional layers of the base view, additionally filled in the spatial and temporal domain in order to fit the non-empty areas into the CU-grid and GOP structure of the HEVC encoder.

Moreover, one of the proposals contained additional noise modeling, including a parametrization of the noise in input views, denoising of input views, and re-noising of the synthesized output views using synthetic noise with proper parameters. The synthetic noise is not correlated to the input noise, thus degrades the pixel-wise quality metrics like PSNR. On the other hand, the re-noised video seems to be subjectively better for the viewer.

Several proposals introduced also additional tools, which increased the efficiency of the immersive video coding or the subjective quality of the synthesized video. For example, one proposal contained additional filtering of the physical edges of the objects in the virtual views to improve the subjective quality; and two of the proposals included the refinement of depth maps for the natural content, as the input depth maps contained many artifacts and were inconsistent between views.

Each of 7 coding techniques (2 anchors + 5 proposals) was used for the generation of the bitstream (cf. Fig. 6).

D. Experimental Results

The correlation between subjective and objective quality (measured by 32 tested quality metrics) is presented in Fig. 9. The first conclusion that follows from the Fig. 9 is that the proposed IV-PSNR metric clearly outperforms other evaluated objective quality metrics. IV-PSNR achieved KROCC higher by 0.057, and SROCC higher by 0.065 than the second-best metric – the full variant of MP-PSNR [75] (0.578 vs. 0.521, and 0.728 vs. 0.663 for KROCC and SROCC, respectively). It should be noted that other metrics designed for multiview video (such as MP-PSNR and MW-PSNR) achieved relatively good results when compared to other state-of-the-art metrics, additionally showing the relevance of the experiment.

Among other metrics, the highest correlation was achieved by VIF-P and MS-SSIM, but the difference between them and IV-PSNR is even more significant than for MP-PSNR-F (0.081 for KROCC and 0.093 for SROCC).

It can be seen that despite IV-PSNR is not utilizing any temporal information (each frame of video is measured independently), it still provides a higher correlation than metrics that measure inter-frame similarity (e.g., VMAF). As it was shown in previous research on the quality of synthesized video [93], the temporal distortions in such video are mainly introduced not by timeline, but by changing the viewpoint, therefore, by the virtual view synthesis. It indicates that for measuring the quality of immersive video it is much more important to focus on the measurement of the synthesis-induced distortion than on the temporal instability of the content itself.

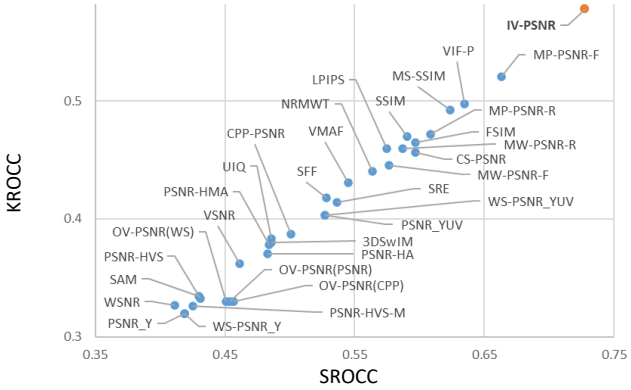


Fig. 9. SROCC and KROCC values for the considered metrics.

VIII. Effectiveness for Typical Processing of Immersive Video

A. Overview of the Experiment

The experiment described in the previous section assessed the correlation between the objective quality metrics and the subjective quality perceived by a viewer for different immersive video coding techniques. In the second experiment, the correlation between MOS and objective quality metrics for different types of immersive video processing was evaluated.

Three common types of processing of immersive video were analyzed: color correction of input views, filtration of reprojected views and depth maps, and the influence of using different view synthesis algorithms.

B. Methodology

For each test sequence, the quality of 4 videos was compared:

A. synthesized using View Synthesis Reference Software (VSRS) [100],
B. synthesized using MultiView Synthesizer (MVS) [101] with no color correction and filtration of reprojected views and depth maps,
C. synthesized using MVS with additional filtration [102],
D. synthesized using MVS with filtration and color correction of input views [62].

The subjective quality evaluation was performed using the PairComparison (PC) method [90], where a viewer compares the quality of two videos presented side-by-side using the scale [-3, 3], where -3 means, that the left video has significantly better quality, 0 – that the quality of both videos is equal, and 3 means that the right video is significantly better [103]. The PC method was chosen, as it performs better than the most popular Absolute or Degradation Category Rating (ACR, DCR) methods when the characteristics of errors/artifacts in videos being compared are significantly different [104], [105].

The viewers were comparing videos in 3 test types (Fig. 10):

1. different synthesizers: the quality difference between videos A and D,
2. color correction: the quality difference between D and B,
3. filtration: the quality difference between videos C and B.

In total, 44 naïve viewers participated in the viewing sessions. Each viewer made 72 assessments: 12 sequences × 3 test types × 2 presentation orders (L/R, R/L). The objective quality evaluation was performed by assessing the quality of videos A, B, C, and D compared to the reference input view.


Fig. 10. Fragments of virtual views compared in 3 test types (from left):
1. different synthesizers: VSRS (top) and MVS (bottom), *BBB Flowers Arc*,
2. color correction: disabled (top) and enabled (bottom), *SoccerArc*,
3. filtration: disabled (top) and enabled (bottom), *PoznanBlocks*.

Then, for each sequence and metric, three differences were calculated, in the same manner as was done for subjective quality assessment. For example, for PSNR metric and *Ballet* sequence, 3 differences were calculated: $\Delta\mathrm{PSNR}_{AD} = \mathrm{PSNR}_A - \mathrm{PSNR}_D$, $\Delta\mathrm{PSNR}_{DB} = \mathrm{PSNR}_D - \mathrm{PSNR}_B$ and $\Delta\mathrm{PSNR}_{CB} = \mathrm{PSNR}_C - \mathrm{PSNR}_B$. It should be highlighted, that the correlation between subjective and objective quality was calculated for these differences, not the absolute values of metrics.

C. Dataset

The experiment was performed on a test set containing 12 miscellaneous multiview sequences (Fig. 11): *BBB Butterfly Arc*, *BBB Flowers Arc* [94], *Carpark*, *Street* [95], *PoznanBlocks2*, *Fencing*, *PoznanService2* [96], *Breakdancers*, *Ballet* [97], *SoccerLinear*, *SoccerArc* [98], *PoznanBlocks* [99].

D. Experimental Results

The results of the second experiment are presented in Fig. 12. As previously, the metrics adapted for multiview video perform better, than other state-of-the-art methods.

When considering SROCC, IV-PSNR outperforms all other considered quality metrics, while the second-best metric is again MP-PSNR (however, as opposed to the previous experiment, in its reduced variant [76]). The difference in their correlations is higher than in the first experiment (0.088).

Fig. 11. Test sequences. 1st row (from left): BBB Butterfly Arc, BBB Flowers Arc, and PoznanService2; 2nd row (from left): Carpark, Street, and Fencing; 3rd row (from left): PoznanBlocks, PoznanBlocks2, and SoccerArc; 4th row (from left): Ballet, Breakdancers, and SoccerLinear.

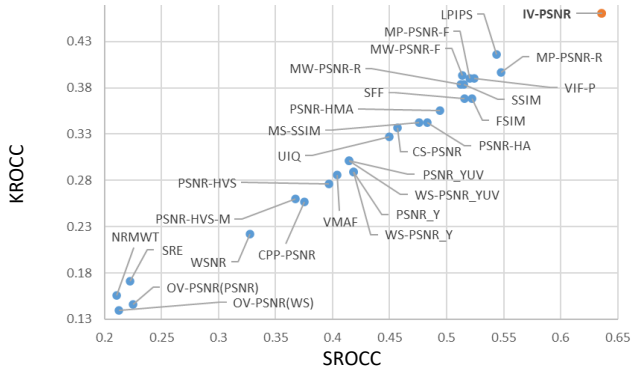


Fig. 12. SROCC and KROCC values for all considered metrics.

Also for KROCC, IV-PSNR is the best metric, slightly (0.044) above LPIPS [39]. However, it should be noted, that LPIPS is a machine-learning-based method, what makes it less robust for different applications (its efficiency highly depends on the pre-trained model). For example, when the quality of the immersive video compression was assessed (Section VII), the LPIPS performed much worse than other considered metrics.

IX. Effectiveness for Non-Immersive Video Applications

A. Overview of the Experiment

In the third experiment, the proposed quality metric IV-PSNR was compared to other objective quality metrics in non-immersive video including typical 2D video and simple 3DoF applications.

The rationale behind this experiment is to present, that the proposed IV-PSNR metric not only outperforms the state-of-the-art metrics for immersive video applications but also can be efficiently used in other typical scenarios, being competitive to other efficient quality metrics.

To perform a valid comparison, TID2013 [78] and CVIQ [108] databases were used.

The TID2013 database contains images distorted using 24 different types of distortions, such as various noise types, image blurring, compression, transmission errors, contrast and brightness change, etc. The complete list of the distortion types is presented in Table VI.

The CVIQ database contains several omnidirectional ERP videos, allowing to assess the efficiency of the metric for a simple virtual reality scenario (i.e., 3DoF [109]), where a user may look around the scene but cannot change his or her position.

Obviously, there are numerous image and video quality assessment databases, such as commonly used CSIQ [106], LIVE [107], and databases designed for omnidirectional video, e.g., VQA-ODV [110] and BIT360 [111]. However, we decided to show the performance of IV-PSNR on TID2013, as it contains the highest number of distortion types, what allows for comprehensively comparing IV-PSNR with state-of-the-art quality metrics. For omnidirectional video, we tested the metrics on CSIQ, which contains 528 images compressed with three coding techniques at several bitrates.

B. TID2013 Database

The correlation between subjective and objective quality was estimated separately for each distortion type and presented in Table VI. Then, for each distortion type, all considered quality metrics were ranked based on SROCC value. Fig. 13 contains SROCC and KROCC values averaged over all distortion types.

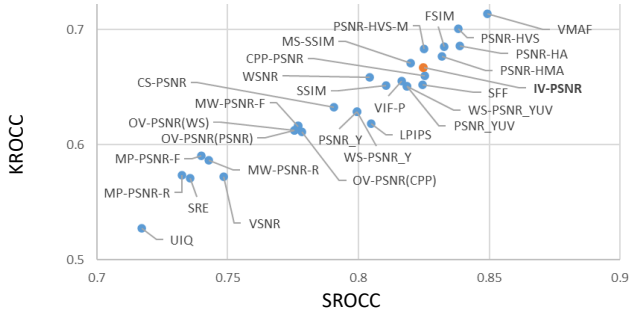


Fig. 13. TID2013: SROCC and KROCC values for the considered metrics.

As presented in Fig. 13, the IV-PSNR metric performs similarly to other commonly used state-of-the-art metrics, such as MS-SSIM, VIF, and VMAF. What was expected, among tested methods, the best are versatile methods designed for assessing the quality of non-immersive videos. Most of these methods are based on frequency-domain-based assessment (PSNR-HVS, PSNR-HA, PSNR-HMA, and, to some extent, FSIM), showing potential in this direction of research, or either trained to be sensitive to traditional distortions (VMAF). In general, IV-PSNR was the 7th best metric among 32 tested quality metrics in assessments it was not designed for. It implies, that it can be successfully used in any application, not only in immersive video systems.

C. *CVIQ database*

In contrast to the TID2013 database, which contains 24 types of distortions, in the CVIQ database, only the compression (using various encoders) is considered. Therefore, the correlation between objective and subjective quality was estimated for the entire database at once (not for each distortion separately, as for TID2013). The results are presented in Fig. 14.

The SROCC and KROCC results for the proposed IV-PSNR metric are very similar for both TID2013 and CVIQ databases, where the IV-PSNR performs as well as other commonly used objective quality metrics being at the 5th place among all tested metrics.

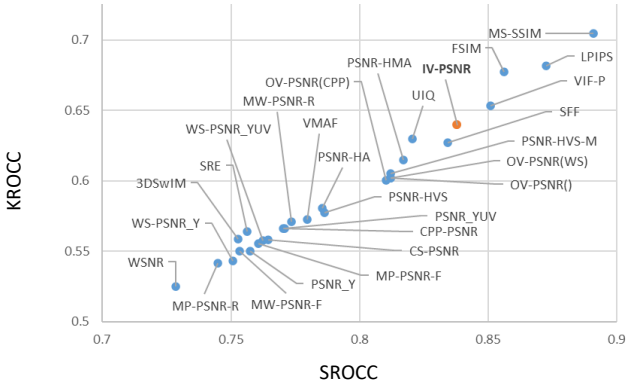


Fig. 14. CVIQ: SROCC and KROCC values for the considered metrics.

Presented results show that the proposed quality metric can be efficiently used not only for the sophisticated immersive video systems where a user may freely navigate in the scene (6DoF), but also for simplified systems, where a user can only look around the scene (3DoF).

X. Computational Time Estimation

In three previous sections, we presented the comparison between IV-PSNR and state-of-the-art metrics in terms of correlation with subjective quality. Of course, a practical objective quality should mimic the human visual system, but also it should provide the results as fast as possible.

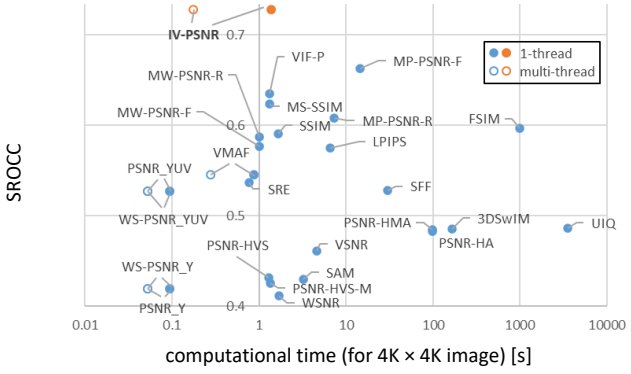


Fig. 15. SROCC and mean computational time for all considered metrics.

Fig. 15 presents the SROCC values for all considered metrics together with the time required for the calculation of the quality of one frame of high-resolution video (4096 × 4096 pixels, sequence *Hijack*). For the purposes of this comparison, SROCC values from Section VII were reused, as the immersive video coding is the main application of the IV-PSNR metric.

It should be noted that time values presented in Fig. 15 are approximate, as extremely different implementations of various metrics were used, including Matlab, C++, and Python with various libraries. We did not optimize implementations of any tested metric except for IV-PSNR, PSNR, and WS-PSNR (which are implemented within the IV-PSNR software), for all other metrics, the existing implementations were used.

All the calculations were performed on the PC equipped with AMD Ryzen 9 3900XT 12-core processor operating at 3.79 GHz, 32 GB RAM DDR4, SSD, and 64-bit Windows 10.

As it is presented in Fig. 15, the IV-PSNR value can be calculated much faster, than other state-of-the-art metrics (especially when multithreading is enabled), simultaneously providing the highest correlation with the human visual system.

XI. Conclusions

The processing and compression of immersive video introduce distortions that are not properly assessed using state-of-the-art objective quality metrics, thus the correlation between objective and subjective quality for immersive video systems is usually not satisfactory. This paper presents a quality metric adapted for such kinds of systems. The proposed metric, IV-PSNR, is a full-reference, PSNR-based objective quality metric. It contains two main techniques, which significantly increase its correlation with the human visual system: the corresponding pixel shift, which considers the problem with the slight shifting of pixels during reprojection between views; and the global component difference, which deals with the problem of different color characteristics of views captured by different cameras of a multicamera system.

The IV-PSNR metric was compared to 31 state-of-the-art metrics in three experiments, showing its performance for immersive video coding and processing, and also in other applications, using the commonly used IQA TID2013 database. As presented, IV-PSNR clearly outperforms other metrics for immersive video applications and can still be used for other purposes. The proposed metric is efficiently implemented allowing very fast quality assessment. The IV-PSNR software was provided by the authors of this paper and is used by MPEG for the evaluation of the upcoming MPEG Immersive video (MIV) standard. The software is publicly available on the public git repository.

The IV-PSNR metric is based on PSNR, which is the most commonly used quality metric in video processing. However, the proposed immersive-video-directed techniques (corresponding pixel shift and global component difference) are metric-agnostic, thus, in the future, they could be used together with other quality metrics, also with ones, which provide a better correlation with subjective quality such as SSIM.

TABLE VI RANKS FOR ALL CONSIDERED METRICS (THE BEST METRIC FOR EACH DISTORTION TYPE IS HIGHLIGHTED); WS-PSNR$_Y$ AND WS-PSNR$_{YUV}$ METRICS WERE OMITTED, AS FOR PERSPECTIVE CONTENT, THEY PERFORM IDENTICALLY TO PSNR$_Y$ AND PSNR$_{YUV}$.

#	Distortion type \ Quality metric	**IV-PSNR**	PSNR$_Y$	PSNR$_{YUV}$	CS-PSNR	PSNR-HVS	PSNR-HVS-M	PSNR-HA	PSNR-HMA	CPP-PSNR	OV-PSNR$_{PSNR}$	OV-PSNR$_{CPP}$	OV-PSNR$_{WS}$	MP-PSNR-R	MP-PSNR-F	MW-PSNR-R	MW-PSNR-F	UIQ	SSIM	MS-SSIM	FSIM	VSNR	WSNR	SFF	SRE	SAM	3DSwIM	LPIPS	NRMWT	VIF-P	VMAF
#	Average SROCC	7	16	11	17	3	8	2	5	6	21	18	20	26	24	23	19	27	13	10	4	22	15	9	25	30	28	14	29	12	1
1	Additive Gaussian noise	1	13	12	9	4	14	2	7	11	8	5	10	15	21	6	3	27	23	22	19	25	20	17	18	30	29	26	28	24	16
2	Noise in color comp.	1	2	4	14	3	18	8	16	15	13	7	12	10	21	6	5	27	23	24	19	25	11	20	17	30	29	26	28	22	9
3	Spatially correl. noise	2	13	12	8	11	5	1	3	14	7	10	9	15	22	6	4	27	21	23	19	26	20	18	17	30	28	25	29	24	16
4	Masked noise	19	6	3	2	20	25	14	16	4	12	11	9	21	22	15	1	23	10	13	17	26	27	8	7	30	29	18	28	5	24
5	High freq. noise	5	14	15	13	2	18	1	10	16	8	9	11	12	21	6	4	28	25	22	19	26	17	20	7	30	29	27	24	23	3
6	Impulse noise	11	8	5	2	7	12	4	9	3	14	20	15	29	27	28	26	17	18	21	16	22	6	19	1	30	23	25	24	13	10
7	Quantization noise	9	11	13	16	4	2	3	1	12	6	5	7	20	25	22	21	27	24	14	15	19	8	17	18	29	28	23	30	26	10
8	Gaussian blur	18	14	16	15	11	13	10	12	17	21	19	22	26	20	25	23	24	5	1	7	8	9	3	27	29	28	6	30	4	2
9	Image denoising	1	9	10	19	4	11	2	3	8	6	5	7	24	14	23	15	27	22	17	16	21	13	20	25	29	28	18	30	26	12
10	JPEG compression	17	20	6	16	1	13	2	4	9	21	15	19	7	23	3	5	27	24	10	11	26	12	14	25	29	28	18	30	22	8
11	JPEG2000 compression	23	26	16	22	4	2	3	1	20	19	15	18	10	14	12	21	25	24	8	5	17	9	6	27	29	28	11	30	13	7
12	JPEG transm. errors	18	19	20	22	14	15	7	6	9	24	26	25	13	16	10	12	23	11	4	5	17	21	2	28	29	27	1	30	8	3
13	JPEG2000 transm. errors	20	10	16	22	3	4	1	2	7	25	23	24	12	21	5	8	26	14	9	6	19	11	13	27	30	28	17	29	15	18
14	Non ecc. patt. noise	1	23	15	16	19	14	17	12	22	28	26	27	5	20	11	18	21	13	3	4	24	6	10	7	30	25	2	29	9	8
15	Local block-wise dist.	24	16	21	27	14	18	20	22	19	30	28	29	13	10	11	9	3	1	6	4	17	25	15	26	23	2	8	12	5	7
16	Mean shift	14	10	12	11	8	5	18	17	13	6	9	7	25	15	27	4	20	1	3	19	26	2	21	24	30	28	16	29	22	23
17	Contrast change	10	18	20	22	16	19	4	3	12	15	14	13	24	5	23	6	27	11	8	7	25	21	9	28	29	26	17	30	2	1
18	Change of color saturation	5	24	18	4	10	11	20	21	6	7	9	8	27	26	29	28	2	14	15	16	22	12	1	23	25	30	3	19	13	17
19	Multipl. Gaussian noise	2	12	13	7	9	14	1	8	3	6	10	5	21	26	19	11	28	22	20	17	24	16	18	4	30	29	25	27	23	15
20	Comfort noise	21	23	5	6	2	10	1	7	14	11	9	12	18	25	17	19	27	26	22	13	20	8	15	3	29	28	16	30	24	4
21	Lossy compr. of noisy im.	5	20	17	1	7	4	10	3	16	12	11	13	19	26	18	15	27	22	21	9	25	6	14	8	30	28	23	29	24	2
22	Im. color quant. w. dither	1	3	6	20	2	7	4	5	8	11	9	10	25	21	22	16	29	23	19	15	18	12	14	26	30	13	24	28	27	17
23	Chromatic aberrations	23	3	12	28	7	16	20	21	18	25	22	24	19	5	10	11	15	2	4	13	14	17	8	27	29	26	6	30	1	9
24	Sparse sampl. and reconstr.	5	24	13	14	4	2	3	1	15	22	16	21	23	20	19	18	27	25	8	6	12	9	7	26	29	28	11	30	17	10

XII. ACKNOWLEDGMENT

The Authors would like to thank Vittorio Baroncini from EVATech and Giacomo Baroncini from GBTech for providing the results of the subjective quality evaluation of MPEG Call for Proposals. The Authors would like to thank also the proponents of MPEG CfP: Bart Kroon and Bart Sonneveldt (Philips Research Eindhoven), Vinod Kumar Malamal Vadakital (Nokia), Lu Yu and Bin Wang (Zhejiang University), Gérard Briand, Julien Fleureau, and Renaud Doré (Technicolor/ InterDigital) for providing executables and/or videos generated using their proposals for the evaluation of the proposed metric.

REFERENCES

[1] J. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V.K.M. Vadakital, and L. Yu, "MPEG Immersive Video coding standard," *Proceedings of the IEEE*, pp. 1-16, 03.2021.

[2] J.B. Jeong et al., "Towards 3DoF+ 360 video streaming system for immersive media," *IEEE Access*, vol. 7, pp. 136399-136408, Sep. 2019.

[3] A. Dziembowski et al., "Virtual view synthesis for 3DoF+ video," *Picture Coding Symposium, PCS 2019*, Ningbo, China, Nov. 2019.

[4] S. Fachada, D. Bonatto, A. Schenkel, and G. Lafruit, "Depth image based view synthesis with multiple reference views for virtual reality," *3DTV Conference 2018*, Stockholm/Helsinki, Sweden/Finland, Jun. 2018.

[5] S. Tian et al., "A benchmark of DIBR synthesized view quality assessment metrics on new database for immersive media applications," *IEEE Tr. on Multimedia*, vol. 21, no. 5, pp. 1235-1247, May 2019.

[6] L. Wang, Y. Zhao, X. Ma, S. Qi, W. Yan, and H. Chen, "Quality Assessment for DIBR-Synthesized Images with Local and Global Distortions," *IEEE Access*, vol. 8, pp. 27938-27948, 2020.

[7] L. Fang et al., "Estimation of virtual view synthesis distortion toward virtual view position," *IEEE T Im. Proc.* 25(5), pp. 1961-1976, 2016.

[8] L. Fang, N. M. Cheung, D. Tian, A. Vetro, H. Sun, and O.C. Au, "An Analytical Model for Synthesis Distortion Estimation in 3D Video," *IEEE Tr. on Image Processing*, vol. 23, no. 1, pp. 185-199, Jan. 2014.

[9] A. Dziembowski et al., "Color correction for immersive video applications," *IEEE Access*, vol. 9, pp. 75626-75640, May 2021.

[10] O. Stankiewicz et al., "A free-viewpoint television system for horizontal virtual navigation," *IEEE Tr. Mult.*, vol. 20, no. 8, pp. 2182-2195, 2018.

[11] X. Wu et al., "Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display," *IEEE Tr. on Circuits and Systems for Video Tech.*, vol. 31, no. 12, Dec. 2021

[12] Q. Liu et al., "PQA-Net: Deep No Reference Point Cloud Quality Assessment via Multi-View Projection," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4645-4660, 2021.

[13] Y. Sun et al., "Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video," *IEEE Sig. Proc. Lett.*, vol. 24, no. 9, Sep. 2017.

[14] V. Zakharchenko et al., "Quality metric for spherical panoramic video," Proc. SPIE 9970, Optics and Photonics for Inf. Proc. X, Sep. 2016.

[15] P. Gao et al., "Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Approach," IEEE Tr. Mult, Dec. 2020.

[16] K. Egiazarian et al., "New full-reference quality metrics based on HVS," *2nd Int. W. on Video Proc. and Quality Metrics*, Scottsdale, USA, 2006.

[17] N. Ponomarenko et al., "On between-coefficient contrast masking of DCT basis functions," *3rd International Workshop on Video Processing and Quality Metrics*, Scottsdale, USA, 2007.

[18] N. Ponomarenko et al., "Modified image visual quality metrics for contrast change and mean shift accounting," *CADSM*, Ukraine, 2011.

[19] X. Shang et al., "Color-Sensitivity-Based Combined PSNR for Objective Video Quality Assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1239-1250, May 2019.

[20] C. Lanaras, et al., "Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, 2018.

[21] D.M. Chandler et al., "VSNR: a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Tr Im Proc*, vol. 16, pp. 2284-2298, 2007.

[22] T. Mitsa et al., "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," *ICASSP Conference*, Minneapolis, USA, pp. 301-304, 1993.

[23] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, Jan. 2004.

[24] Z. Wang, A.C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81-84, 2002.

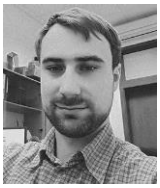
[25] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," *The 37th Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, USA, Nov. 2003.
[26] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Tr. on Image Proc.*, vol. 21, no. 4, Apr. 2012.
[27] L. Zhang et al., "FSIM: a feature similarity index for image quality assessment," *IEEE Tr. on Image Proc.*, vol. 20, pp. 2378-2386, 2011.
[28] D. Marr, *Vision*, New York, Freeman, 1980.
[29] M.C. Morrone and D. C. Burr, "Feature detection in human vision: A phase-dependent energy model," *Proceedings of the Royal Society of London. Series B*, vol. 235, no. 1280, pp. 221–245, Dec. 1988.
[30] A. Ahar, A. Barri, and P. Schelkens, "From Sparse Coding Significance to Perceptual Quality: A New Approach for Image Quality Assessment," *IEEE Tr. on Image Processing*, vol. 27, no. 2, pp. 879-893, Feb. 2018.
[31] F.A. Kruse et al., "The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging spectrometer Data," *Remote Sensing of Environment*, vol. 44, pp. 145-163, 1993.
[32] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, pp. 430-444, 2006.
[33] H.W. Chang et al., "Sparse feature fidelity for perceptual image quality assessment," *IEEE Tr. on Image Proc.*, vol. 22, pp. 4007-4018, 2013.
[34] B.A. Olshausen, "Principles of image representation in visual cortex," *The Visual Neurosc.*, Cambridge, USA, MIT Press, pp. 1603-1615, 2003.
[35] S. Li, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE T. Mult*, 13(5), pp. 935-949, Oct. 2011.
[36] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[37] Z. Li et al., "Toward a practical perceptual video quality metric," *Netflix Technology Blog, Tech. Rep.*, 2016.
[38] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video Multimethod Assessment Fusion (VMAF) on 360VR Contents," *IEEE Tr. on Consumer Electronics*, vol. 66, no. 1, pp. 22-31, Feb. 2020.
[39] R. Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *2018 CVPR Conference*, 2018, pp. 586-595.
[40] E. Bosc et al., "Towards a New Quality Metric for 3-D Synthesized View Assessment," *IEEE J. Sel. Topics in Sig. Proc.*, vol. 5, no. 7, Nov. 2011.
[41] F. Battisti et al., "Objective image quality assessment of 3D synthesized views," *Signal Proc.: Image Communication*, vol. 30, pp. 78-88, 2015.
[42] L. Li et al., "Quality Assessment of DIBR-Synthesized Images by Measuring Local Geometric Distortions and Global Sharpness," *IEEE Tr. on Multimedia*, vol. 20, no. 4, pp. 914-926, April 2018.
[43] D. Sandić-Stanković et al., "DIBR-synthesized image quality assessment based on morphological multi-scale approach," *J Im Vid Pr*, vol. 4, 2016.
[44] Z. Chen, W. Zhou, and W. Li, "Blind Stereoscopic Video Quality Assessment: From Depth Perception to Overall Experience," *IEEE Tr. on Image Processing*, vol. 27, no. 2, pp. 721-734, Feb. 2018.
[45] B. Appina et al., "Study of Subjective Quality and Objective Blind Quality Prediction of Stereoscopic Videos," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5027-5040, Oct. 2019.
[46] H. Jiang et al., "Multi-Angle Projection Based Blind Omnidirectional Image Quality Assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, early access, 2021.
[47] X. Chai et al., "Monocular and Binocular Interactions Oriented Deformable Convolutional Networks for Blind Quality Assessment of Stereoscopic Omnidirectional Images," *IEEE Transactions on Circuits and Systems for Video Technology*, early access, 2021
[48] L. Shi et al., "No-Reference Light Field Image Quality Assessment Based on Spatial-Angular Measurement," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4114-4128, 2020.
[49] Y. Tian et al., "A Light Field Image Quality Assessment Model Based on Symmetry and Depth Features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2046-2050, May 2021.
[50] S. Tian, L. Zhang, L. Morin, and O. Déforges, "NIQSV+: A No-Reference Synthesized View Quality Assessment Metric," *IEEE Tr. on Image Processing*, vol. 27, no. 4, pp. 1652-1664, April 2018.
[51] K. Gu et al., "Model-Based Referenceless Quality Metric of 3D Synthesized Images Using Local Image Description," *IEEE Tr. on Image Processing*, vol. 27, no. 1, pp. 394-405, Jan. 2018.
[52] X. Liu et al., "Subjective and Objective Video Quality Assessment of 3D Synthesized Views with Texture/Depth Compression Distortion," *IEEE Tr. on Image Proc.*, vol. 24, no. 12, pp. 4847-4861, Dec. 2015.
[53] F. Shao, Q. Yuan, W. Lin, and G. Jiang, "No-Reference View Synthesis Quality Prediction for 3-D Videos Based on Color–Depth Interactions," *IEEE Tr. on Multimedia*, vol. 20, no. 3, pp. 659-674, March 2018.
[54] L. Li, Y. Huang, J. Wu, K. Gu, and Y. Fang, "Predicting the Quality of View Synthesis with Color-Depth Image Fusion," *IEEE Transactions on Circuits and Systems for Video Technology*.
[55] D. Sandić-Stanković et al., "Fast Blind Quality Assessment of DIBR-Synthesized Video based on High-High Wavelet Subband", IEEE Tr. Im. Proc., vol. 28(11), pp. 5524-5536, 2019.
[56] G. Wang et al., "Reference-free DIBR-synthesized Video Quality Metric in Spatial and Temporal Domains," *IEEE Transactions on Circuits and Systems for Video Technology*, early access, 2021.
[57] J. Stankowski and A. Dziembowski, "Fast view synthesis for immersive video systems," *28. Int. Conf. in Central Europe on Comp. Gr., Vis. and Comp. Vis., WSCG 2020*, Plzen, Czech Republic, May 2020.
[58] R. Doré, "Technicolor 3DoF+ test materials," *ISO/IEC JTC1/SC29/WG11 MPEG M42349*, San Diego, USA, Apr. 2018.
[59] A. Leontaris et al., "Quality Evaluation of Motion-Compensated Edge Artifacts in Compressed Video," *IEEE T Im Proc*, vol. 16(4), April 2007.
[60] D. Mieloch et al., "Point-to-block Matching in Depth Estimation," in *29. Int. Conf. on Comp. Gr., Vis. and Comp. Vis. WSCG*, Online, May 2021.
[61] T. Senoh, N. Tetsutani, and H. Yasuda, "[MPEG-I Visual] Proposal of trimming and color matching of multi-view sequences," *ISO/IEC JTC1/SC29/WG11 MPEG M47170*, Geneva, Switzerland, Mar. 2019.
[62] A. Dziembowski et al, "Adaptive color correction method in virtual view synthesis," *3DTV Conf.*, Stockholm/Helsinki, Sweden/Finland, 2018.
[63] J. Snyder and P. Voxland, "An album of map projections," *US Government Printing Office*, Washington, 1989.
[64] M. Domański et al., "Immersive visual media – MPEG-I: 360 video, virtual navigation and beyond," *IWSSIP 2017*, Poznań, Poland, 2017.
[65] ITU-R, "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios," *Rec. ITU-R BT.601-7*, Mar. 2011.
[66] Z. Wang and D. Zhang, "Progressive Switching Median Filter for the Removal of Impulse Noise from Highly Corrupted Images," *IEEE Transactions on Circuits and Systems-II*, vol. 46, no. 1, pp. 78-80, 1999.
[67] ISO/IEC/IEEE, "Software engineering - Recommended practice for software acquisition," *Std. ISO/IEC/IEEE 41062:2019*.
[68] ISO/IEC, "Programming languages – C++," *Std. ISO/IEC 14882:2017*.
[69] E. Dumic et al., "Image quality of 4:2:2 and 4:2:0 chroma subsampling formats," 2009 Int. Symp. ELMAR, Zadar, Croatia, Sep. 2009.
[70] A. Neumaier, "Rundungsfehleranalyse einiger Verfahren zur Summation endlicher Summen," Z. Angew. Math. Mech., vol. 54, pp. 39-51, 1974.
[71] H.J. Lu et al., "System V Application Binary Interface," *AMD64 Architecture Processor Supplement, Version 1.0*, May 2021.
[72] A. Dziembowski, "Software manual of IV-PSNR for Immersive Video," *ISO/IEC JTC1/SC29/WG04 MPEG VC N0013*, Online, Oct. 2020.
[73] Y. Huang, H. Qi, B. Li, and J. Xu, "Adaptive weighted distortion optimization for video coding in RGB color space," *IEEE International Conference on Image Processing, ICIP 2014*, Paris, France, Oct. 2014.
[74] M.U. Müller et al., "Super-resolution of multispectral satellite images using convolutional neural networks," *ISPRS Annals.*, pp. 33-40, 2020.
[75] D. Sandić-Stanković et al., "DIBR synthesized image quality assessment based on morphological pyramids," *3DTV-CON*, Lisbon, Jul. 2015.
[76] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi–Scale Synthesized View Assessment Based on Morphological Pyramids," *Journal of Electrical Engineering*, vol. 67 (1), pp. 1–9, 2016.
[77] D. Sandić-Stanković, D. Kukolj, P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets", *Int. W. on Quality of Multimedia Experience QoMEX*, Costa Navarino, Greece, May 2015.
[78] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Proc.: Image Comm.*, vol. 30, pp. 57-77, 2015.
[79] J.L. Myers, A. Well, "Research design and statistical analysis," *Lawrence Erlbaum Associates*, London, UK, 2003.
[80] "Call for Proposals on 3DoF+ Visual," *ISO/IEC JTC1/SC29/WG11 MPEG N18145*, Marrakech, Morocco, Jan. 2019.
[81] B. Kroon, "3DoF+ test sequence ClassroomVideo," *ISO/IEC JTC1/SC29/WG11 MPEG M42415*, San Diego, USA, Apr. 2018.

[82] D. Doyen et al., "Light field content from 16-camera rig," *ISO/IEC JTC1/SC29/WG11 MPEG M40010*, Geneva, Switzerland, Jan. 2017.
[83] B. Salahieh et al., "Kermit test sequence for Windowed 6DoF activities," *ISO/IEC JTC1/SC29/WG11 MPEG M43748*, Ljubljana, Slovenia, 2018.
[84] V.K.M. Vadakital et al., "Description of Nokia's response to CFP for 3DOF+," *ISO/IEC JTC1/SC29/WG11 M47372*, Geneva, Mar. 2019.
[85] B. Kroon et al., "Philips response to 3DoF+ Visual CfP," *ISO/IEC JTC1/SC29/WG11 MPEG M47179*, Geneva, Switzerland, Mar. 2019.
[86] M. Domański et al., "Technical description of proposal for Call for Proposals on 3DoF+ Visual prepared by PUT and ETRI," *ISO/IEC JTC1/SC29/WG11 MPEG M47407*, Geneva, Switzerland, Mar. 2019.
[87] J. Fleureau et al., "Technicolor-Intel response to 3DoF+ CfP," *ISO/IEC JTC1/SC29/WG11 MPEG M47445*, Geneva, Switzerland, Mar. 2019.
[88] B. Wang et al., "Description of ZJU's response to 3DoF+ Visual CfP," *ISO/IEC JTC1/SC29/WG11 MPEG M47684*, Geneva, Mar. 2019.
[89] A. Hornbarg, "Handbook of Machine Vision," pp. 46-47, *Wiley*, 2007.
[90] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Recommendation ITU-T P.910*, Apr. 2008.
[91] V. Baroncini and G. Baroncini, "Evaluation Results of the Call for Proposals on 3DoF+ Visual," *ISO/IEC JTC1/SC29/WG11 MPEG N18353*, Geneva, Switzerland, Mar. 2019.
[92] Y. Meng and Z. Ma, "Viewport-Based Omnidirectional Video Quality Assessment: Database, Modeling and Inference," *IEEE Tran. on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 120-134, Jan. 2022.
[93] S. Tian et al., "Quality assessment of DIBR-synthesized views: An overview," Neurocomputing, vol. 423, pp. 158-178, Jan. 2021.
[94] P.T. Kovacs et al., "Big Buck Bunny light-field test sequences," *ISO/IEC JTC1/SC29/WG11 MPEG M36500*, Warsaw, Poland, Jun. 2015.
[95] D. Mieloch et al., "[MPEG-I Visual] Natural outdoor test sequences," *ISO/IEC JTC1/SC29/WG11 MPEG M51598*, Brussels, Belgium, 2020.
[96] M. Domański et al., "Multiview test video sequences for free navigation exploration obtained using pairs of cameras," *ISO/IEC JTC1/SC29/WG11 MPEG M38247*, Geneva, Switzerland, Jun. 2016.
[97] C.L. Zitnick et al., "High-quality video view interpolation using a layered representation," *ACM T Gr.*, vol. 3(23), pp. 600-608, Aug. 2004.
[98] P. Goorts, "Real-time adaptive plane sweeping for free viewpoint navigation in soccer scenes," *PhD*, Hasselt, Belgium, pp.175-186, 2014.
[99] M. Domański et al., "Poznan Blocks – a multiview video test sequence and camera parameters for free viewpoint television," *ISO/IEC JTC1/SC29/WG11 MPEG M32243*, San Jose, United States, Jan. 2014.
[100] T. Senoh et al., "View Synthesis Reference Software (VSRS) 4.2 with improved inpainting and hole filling," *ISO/IEC JTC1/SC29/WG11 MPEG M40657*, Hobart, Australia, Apr. 2017.
[101] A. Dziembowski et al., "Multiview Synthesis – improved view synthesis for virtual navigation," *PCS 2016*, Nürnberg, Germany, Dec. 2016.
[102] A. Dziembowski et al., "View and depth preprocessing for view synthesis enhancement," *Int. J. El. Tel.*, vol. 64(3), pp. 269-275, 2018.
[103] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT.500-9*, Nov. 1998.
[104] A.M. van Dijk et al., "Subjective quality assessment of compressed images," *Signal Processing*, vol. 58, no. 3, pp. 235-252, May 1997.
[105] R.K. Mantiuk et al., "Comparison of four subjective methods for image quality assessment," *Comp. Gr. Forum*, vol. 31, no. 8, Dec. 2012.
[106] E.C. Larson et al., "Most apparent distortion: full-reference image quality assessment and the role of strategy," J. El. Imaging, 19(1), 2010.
[107] Z. Sinno and A.C. Bovik, "Large-Scale Study of Perceptual Video Quality," IEEE Tr. on Im. Proc., vol. 28, no. 2, pp. 612-627, Feb. 2019.
[108] W. Sun et al., "MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment," IEEE Sel. Top. Sig. Pr., vol. 14, no. 1, 2020.
[109] M. Wien et al., "Standardization status of immersive video coding," IEEE J. Emerg. and Sel. Top. in Circ. and Syst., vol. 9, no. 9, Mar. 2019.
[110] M. Xu et al., "Viewport-based CNN: A multi-task approach for assessing 360° video quality," IEEE Tr. on Pattern Analysis and Machine Intelligence, vol. 44, no. 4, pp. 20198-2215, Apr. 2022.
[111] B. Zhang et al., "Subjective and objective quality assessment of panoramic videos in virtual reality environments," IEEE Conf. on Multimedia & Expo Workshops, Hong Kong, China, Jul. 2017.
[112] B. Bross et al. "Overview of the Versatile Video Coding (VVC) standard and its applications," IEEE Tr. on Circ. and Syst. for Vid. Tech., 2021
[113] K. Ma et al., "Geometric transformation invariant image quality assessment using convolutional neural networks," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Calgary, Canada, Apr. 2018.
[114] X. Sui at al., "Perceptual quality assessment of omnidirectional images as moving camera videos," IEEE Tr. Vis. & Comp. Gr., E.A., Jan. 2021.
[115] L. Ilola et al., "New test content for immersive video – Nokia Chess," ISO/IEC JTC1/SC29/WG11 MPEG, M50787, Geneva, Oct. 2019.
[116] Q. Jiang et al., "Unified No-Reference Quality Assessment of Singly and Multiply Distorted Stereoscopic Images," *IEEE T. Im. Proc*. 28(4), 2019.
[117] X. Wang et al. "Measuring Coarse-to-Fine Texture and Geometric Distortions for Quality Assessments of DIBR-Synthesized Images," *IEEE Transactions on Multimedia*, vol. 23, pp. 1173-1186, 2020.
[118] D. Sandić-Stanković et al., "Quality Assessment of DIBR-Synthesized Views Based on Sparsity of Difference of Closings and Difference of Gaussians," *IEEE T. Im. Proc*., vol. 31, pp. 1161-1175, 2022.

Adrian Dziembowski was born in Poznań, Poland in 1990. He received the M.Sc. and Ph.D. degrees from the Poznan University of Technology in 2014 and 2018, respectively. Since 2019 he is an Assistant Professor at the Institute of Multimedia Telecommunications. He authored and coauthored about 30 articles on various aspects of immersive video, free navigation, and FTV systems. He is also actively involved in ISO/IEC MPEG activities towards MPEG Immersive video coding standard.

Dawid Mieloch received his M.Sc. and Ph.D. from Poznań University of Technology in 2014 and 2018, respectively. Currently, he is an assistant professor at the Institute of Multimedia Telecommunications. He is actively involved in ISO/IEC MPEG activities where he contributes to the development of the immersive media technologies. He has been involved in several projects focused on multiview and 3D video processing. His professional interests include also free-viewpoint television, depth estimation and camera calibration.

Jakub Stankowski received his M.Sc. from Poznań University of Technology in 2009. Currently he works in the Institute of Multimedia Telecommunications. He has been involved in tens of scientific and R&D projects including compression and protection of video content, multiview data processing, real time 3D video delivery. His current research interests include video compression, performance-optimized video data processing, software optimization, and parallelization techniques.



Adam Grzelka was born in Śrem, Poland in 1990. He received his M.Sc. from the Poznań University of Technology in 2014. Currently, he works in the Institute of Multimedia Telecommunications. His current research interests include image processing, video compression, immersive video with high-QoE, FPGA implementation of compression algorithms and interfaces.