IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 14, 2020, accepted December 28, 2020, date of publication January 11, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050554

Depth Map Refinement for Immersive Video

DAWID MIELOCH, ADRIAN DZIEMBOWSKI, AND MAREK DOMAŃSKI, (Senior Member, IEEE)
Institute of Multimedia Telecommunications, Poznań University of Technology, 60-965 Poznań, Poland
Corresponding author: Dawid Mieloch (dawid.mieloch@put.poznan.pl)

This work was supported by the Ministry of Science and Higher Education of Republic of Poland.

ABSTRACT In this article, we propose a depth map refinement method that increases the quality of immersive video. The proposal highly enhances the inter-view consistency of depth maps (estimated or acquired by any method), crucial for achieving the required fidelity of the virtual view synthesis process. In the described method, only information from depth maps is used, as the use of texture can introduce errors in the refinement, mostly due to inter-view color inconsistencies and noise. In order to evaluate the performance of the proposal and compare it with the state of the art, three experiments were conducted. To test the influence of the refinement on the encoding of immersive video, four sets of depth maps (original, refined with the synthesis-based refinement, a bilateral filter, and with the proposal) were encoded with the MPEG Immersive Video (MIV) encoder. In the second experiment, in order to provide a direct evaluation of the accuracy of depth maps, the Middlebury database comparison was performed. In the third experiment, the temporal consistency of depth maps was assessed by measuring the efficiency of encoding of the virtual views. The experiments showed both a high increase of the virtual view synthesis quality in immersive video applications and higher similarity to ground-truth after the refinement of estimated depth maps. The usefulness of the proposal was appreciated and confirmed by the experts of the ISO/IEC MPEG group for immersive video and the method became the MPEG Reference Software for the depth refinement. The implementation of the method is publicly available for other researchers.

INDEX TERMS Depth map refinement, immersive video, virtual navigation, multiview stereo, inter-view consistency.

I. INTRODUCTION

In this article, we focus on new types of visual systems represented by immersive media [14], [50]. Recent rapid development of virtual reality applications [32], [34], free-viewpoint television [13], [36], [37], and standardization activities in compression of point clouds [29], [33], or immersive video [28], shows that the need for natural visual content with depth information is constantly increasing.

Although substantial efforts are made in providing new depth estimation methods [6], [30], [31], the quality of depth maps for natural content is often too low to provide a satisfactory experience for the final user of immersive video systems. Therefore, in order to improve the process of depth estimation, a variety and multitude of noteworthy depth refinement methods were presented during the last years (e.g. [2], [3], [5]).

The associate editor coordinating the review of this manuscript and approving it for publication was Huazhu Fu.

In many cases, a multimedia system does not have a possibility to control the quality of used depth maps (e.g. the multi-view encoder that has to work with any input data). The use of depth refinement methods as a pre-processing done before using the depth maps can easily increase the usability of such depth. Moreover, most refinement methods can be used with any depth maps (i.e., depth maps of any origin). For example, the depth maps that were acquired by multiple depth cameras (e.g. time-of-flight) can be easily enhanced, therefore, possible to use in applications that require inter-view consistency of depth maps.

The researchers are still mainly focused on multiview video plus depth (MVD) representation [35], therefore, further considerations presented in this article also concern MVD. Of course, multi-plane images (MPI) [7] and their variants [22] are gaining much attention, nevertheless, in these representations, depth information is still present, but in another form.

The goal of this article is to present a new method of depth refinement that will increase the quality of immersive video.

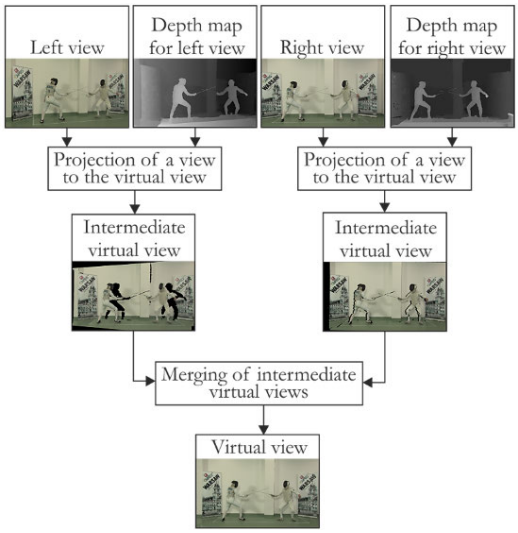

FIGURE 1. Virtual view synthesis. Black regions in intermediate virtual views are occluded in the input views.

This quality should be understood widely, as both the quality of virtual views synthesized for a viewer of immersive video and the possibility of efficient encoding of such video.

The authors believe that these two aspects are of equal importance, as achieving the highest possible quality of rendered views is important in all types of visual media, but efficient encoding of immersive video is also crucial for making it easily available for possible viewers.

As discussed in Section II, many state-of-the-art methods of refinement of depth maps lack versatility or heavily increase the time of processing. The novelty of the proposed method, which was described in Section III, can be seen in merging all the abovementioned aspects in one, not computationally-expensive post-processing algorithm that can be easily included as a part of any immersive video system.

II. MOTIVATION AND RELATED WORKS

In inter-view consistent depth maps, two points that represent the same part of a scene in different views have such values of depth that after the three-dimensional projection of these points they represent the same 3D point of a scene. Lack of the inter-view consistency of depth maps strongly deteriorates the quality of generated virtual views. A simplified process of virtual view synthesis is shown in Fig. 1. In order to correctly determine the positions of objects that were occluded in some views, it is required to use more than one view during the synthesis. Therefore, virtual view synthesis should be performed when at least 2 input views and their corresponding depth maps are available. If an object is not occluded, then the color of this object is usually averaged in the final virtual view, using the colors of the object from both (or more) real views of the scene.

Unfortunately, in the case of depth maps that are not inter-view consistent, an ambiguity of the position of this point in a virtual view can be seen. It results in visible errors in the virtual view, reducing both the objective and the subjective quality of synthesized views [10], [11].

FIGURE 2. Inter-view redundancy removal using MIV: input views (left column), pruned input views (center column), and atlas containing all preserved (non-pruned) information from input views (right column).

The previous works of the authors on the depth estimation method (described in [6]) have also shown that the inter-view consistency is crucial e.g. in free-viewpoint television applications. These results encouraged the authors to propose a method of depth map refinement that could increase this consistency as the post-processing step that could be done after any depth estimation.

The inter-view consistency of depth maps does not only influence the quality of the synthesis, but also the effectiveness of the encoding of immersive video. The state-of-the-art encoding technique for such content, i.e. the newest MPEG Immersive Video (MIV) [28], utilizes depth information to minimize the redundancy between input views of the scene (Fig. 2). The main idea behind MIV is that several base views gathering most of the information of the scene should be encoded in their entirety (second row in Fig. 2), while supplementary information (e.g., disocclusions from other views) can be transmitted in the form of a mosaic of much smaller patches. The pruner, basing on depth inter-view redundancy check, identifies and extracts regions occluded in the input views. These occluded regions are preserved in additional views, while the rest of the regions are removed (top and bottom row of central column in Fig. 2). It results in small patches preserved in the pruned additional views. The packer gathers all patches and preserved input views into atlases (right column of Fig. 2), which summarized size is usually much smaller than the summarized size of input views. The more the input depth maps are inter-view consistent, the more parts of input views can be removed.

A test model of MIV is already publicly available [21], while during the 131st meeting of ISO/IEC MPEG group the Committee Draft of International Standard for MIV was presented [28].

Inter-view consistency of depth maps can be achieved already during the depth estimation process. Existing methods can provide inter-view consistent depth maps of the quality sufficient for view synthesis. An example of such methods is [6], which utilizes graph-based multiview depth

optimization, or deep-learning-based method [41]. Unfortunately, high-quality real-time estimation is being achieved only for stereo pair camera rigs, as in [12]. Obviously, even state-of-the-art depth estimation methods can be assisted with appropriate additional post-processing.

The frequent approach used to increase the inter-view consistency comprises warping of depth maps from all views into the selected viewpoint (usually a viewpoint of the central view) and re-warping such merged depth back to input views. The performed merging can be based on the texture of input (e.g., on local edges and estimated motion of objects [18]). In method [1], the decision of which depth should be used, i.e., the depth of merged central view, or the original input depth, is based on the iterative process that involves the calculation of the quality of virtual view synthesis. Such an approach can highly increase both the inter-view consistency and fidelity of depth maps, nevertheless, is highly prone to view synthesis errors. The use of view synthesis also increases the overall processing time.

Depth refinement can be also a part of virtual view synthesis itself [19], [23], [40]. In [23], depth maps are projected to the center view, as in other abovementioned methods, but the additional optimization of a complex cost function calculated on the basis of input views, depth maps, and motion vectors is performed. The resulting quality of virtual views is significantly improved, but the complexity of this method makes it impossible to perform real-time virtual view synthesis, desirable in all immersive video applications in which such delay influences the subjective quality of immersion [27]. The method described in [40] not only uses virtual view synthesis as a method for depth refinement but also outputs the refined virtual view together with refined depth maps used during the synthesis. It shortens the overall processing time, as the standalone view synthesis is not required, but the time of performing this refinement for stereo pair is still close to one minute per frame.

Many depth refinement methods are based on depth map filtering. Very common is the use of bilateral filters (based on simultaneous utilization of color and depth), which can be used to iteratively refine depth maps [38], [44], [45]. Such methods can be easily enhanced by utilization of temporal information [2], [17] (multi-lateral filtering), or by auto-regressive models [3]. The effects of such methods on depth maps are limited to the removal of depth temporal noise or artifacts resulting from video compression but do not focus on the preserving or increasing of depth maps inter-view consistency. Lack of such consistency can be also seen in super-resolution methods, but such methods can be used to efficiently decrease the computational complexity of depth estimation, as it can be performed initially in lowered resolution. The upsampling and refinement process, especially for methods based on convolutional networks [41], [42], [43], can be relatively fast (below half of second per frame) when GPU implementation is used.

Other methods focus strictly on the refinement of depth acquired with depth cameras [4], [5]. Characteristics and typical errors of such depth maps are different than in image-based estimation, typically used in the multi-view video. Such a lack of versatility can be seen also in methods adapted to use with light field multi-array camera rigs [15].

What connects all presented methods, is that information from input views is used in some stage of processing. It makes them not suitable for color-inconsistent or highly noised input views. Such color artifacts are unavoidable in real-world sequences acquired by a set of cameras and are present in available immersive video test sequences.

Moreover, because the proposal does not utilize any color information, its usability is very high for decoder-side depth estimation [46]. The quality of depth maps estimated using encoded input views can vary significantly, depending on the available bitrate [47]. Use of the proposed refinement, independent of the quality of input views, can be particularly desirable in such applications.

III. PROPOSED DEPTH MAP REFINEMENT

The main idea of the proposal is to enhance the inter-view consistency of the depth maps. Only information from depth maps is used, as the use of texture can introduce errors in the refinement, mostly due to inter-view color inconsistencies and noise.

In the proposed method, the depth value present in the largest number of input views is being chosen as a new depth value, increasing the inter-view consistency of depth maps. In order to further increase this consistency, the refinement is performed iteratively: in the second iteration, the refined depth maps are treated as input ones.

In the first step of the proposed depth inter-view consistency refinement, the cross-view warping is performed to project depth values from all N depth maps into every of N input depth maps.

For all points $p_i(x, y)$ in each depth map i, a list $\boldsymbol{D}_i\,(x, y)$ of depth values projected from various input depth maps is created. Depth values in a list $\boldsymbol{D}_i(x, y)$ are sorted in descending order:

$$\boldsymbol{D}_i\,(x, y) = [d_1\,(x, y)\,, d_2\,(x, y)\,, d_3\,(x, y)\,, \ldots], \tag{1}$$

so $d_1(x, y) \leq d_2(x, y) \leq d_3(x, y)$, etc. Each point $p_i\,(x, y)$ is processed as presented in Fig. 3. A subset of n elements of the list $\boldsymbol{D}_i\,(x, y)$ is defined as

$$\boldsymbol{s}\,(x, y) = [d_{n_0}\,(x, y)\,, \ldots, d_{n_0+n}\,(x, y)], \tag{2}$$

where n_0 is the number of the first element of a subset. The initial value of n (i.e., the length of $\boldsymbol{s}\,(x, y)$ subset) is equal to N (the number of input views), while n_0 initially points to the first element of $\boldsymbol{D}_i(x, y)$, i.e., $n_0 = 1$. The next steps of the algorithm are as follows:

1) If the range of depth values in a given subset $\boldsymbol{s}\,(x, y)$ is smaller than the predefined threshold T (i.e., the subset $\boldsymbol{s}\,(x, y)$ contains very similar values of depth), go to step 3, else continue.

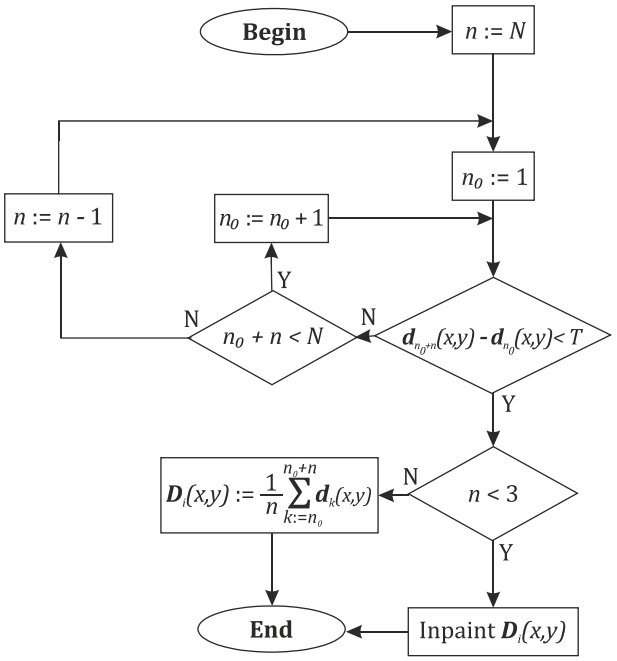

FIGURE 3. Scheme of the processing of each point of input depth maps.

2) If $n_0 + n < N$ (the last element in $\boldsymbol{s}(x, y)$ is not the last element of the $\boldsymbol{D}_i(x, y)$ list), increment n_0 and go to step 1; else decrement n, set $n_0 = 1$, and go to step 1.
3) If $n > 2$, the new depth value for the analyzed point $p_i(x, y)$ is an average value of n values from subset $\boldsymbol{s}(x, y)$; else, the depth value of the analyzed point is inpainted using depth values from neighboring points.

If the list $\boldsymbol{D}_i(x, y)$ for some point contains depth values that are significantly different (there is no subset $\boldsymbol{s}(x, y)$ containing more than two similar values of depth), these values should not be used in depth maps, as they are likely very unreliable. Such points occur mostly in areas occluded in most of the input views, areas with specular or non-Lambertian reflections, or were estimated erroneously in the process of depth estimation.

In the proposal, such unreliable points in depth maps are inpainted using 8-way depth-based inpainting. For each mentioned point, depth values of the nearest non-empty points in each direction are compared. As it was described earlier, unreliable points occur mostly in disoccluded areas of views, therefore, the farthest found depth value is copied to the analyzed point.

Through the preliminary experiments, the threshold T was set to 4% of the depth dynamic range (i.e., 40 for 10-bit depth maps).

Note that the existing methods of depth estimation for stereo pairs usually perform the inter-view consistency check, therefore, the use of the proposed refinement would not give any increase of the quality or fidelity. Nevertheless, the proposed algorithm of refinement was prepared to provide an efficient encoding of immersive video, in which the number of cameras is usually much larger than two (the number of cameras used in experiments described in Section V-A varied from 9 to 15). The use of a single stereo pair would result in a very small volume of three-dimensional space in which the virtual view synthesis could provide satisfactory quality for the final user. Therefore, stereo pairs are not in the scope of the proposal.

The proposed approach influences not only the inter-view consistency of depth maps. After the refinement, uncertain depth values are removed, reducing the number of points in depth maps that are with high probability erroneous, which indirectly also increases their temporal consistency.

The abovementioned features of depth maps, i.e., their inter-view and temporal consistency, highly influence the overall quality of synthesized virtual views [10], [11], [20], making the proposal particularly useful in immersive video applications. This usefulness was appreciated and confirmed by experts of the ISO/IEC MPEG group for immersive video. The proposed method became the MPEG Reference Software for the depth refinement [8].

IV. SOFTWARE IMPLEMENTATION OF THE METHOD

The authors provide the proposal to be freely used by other researchers. The above-described method is implemented as C++ software and can be downloaded together with a manual, configuration examples, and license details from the following repository: https://gitlab.com/dmieloch/depth-map-refinement.

The complementary software is provided for the convenience of the research community. The authors believe that the availability of this new software will be useful as an additional reference for future developments in depth refinement.

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposal and compare it with the state of the art, three experiments were conducted. The following sections describe the assessment of the influence of the proposal on the immersive video encoding, the accuracy of refined depth maps, and the influence on the temporal consistency of depth maps, respectively.

A. ASSESSMENT OF THE INFLUENCE ON THE IMMERSIVE VIDEO ENCODING

1) DESIGN OF THE EXPERIMENT

Original depth maps provided with test sequences were refined using the proposed method, the synthesis-based refinement [1], and our implementation of bilateral filter [38] (both described in Section II). In order to test the influence of the refinement on the encoding of immersive video, four sets of depth maps (original, refined with synthesis-based refinement, bilateral filter, and with the proposal) were encoded with the MPEG Immersive Video (MIV) encoder that was implemented in Test Model for Immersive Video [21].

As it was described in Section II, MIV encodes only the minimal number of parts of input views (patches) that represent the whole scene. If some area is visible in many views (e.g., the background), only one instance of this area is

encoded (Fig. 2). The matching of areas that represent the same object is performed by depth inter-view consistency test. It significantly reduces the amount of data that have to be sent, decreasing the required bitrate of encoded views. Therefore, the performance of MIV is a good determinant of both depth maps quality (as depth maps are used for virtual view synthesis in the decoder) and inter-view consistency (as it influences the process of input views pruning).

All views with corresponding depth maps are encoded with TMIV for 5 different QPs. The encoder was set as in Common Test Conditions for Immersive Video [9]. In the end, the encoded representation of scenes is used to synthesize all input views. These views are used to calculate the average BD-rate (Bjøntegaard metric) [49]. Four objective quality metrics are reported: Y-PSNR, VMAF [51], MS-SSIM [52], and IV-PSNR [53].

Five test sequences, recommended by ISO/IEC MPEG for immersive video experiments, were used in the following test: IntelFrog [24], TechnicolorPainter [25], PoznanStreet, PoznanCarpark, and PoznanHall [26]. Sequences varied in their character (e.g. outdoor scene with large depth range in PoznanStreet, indoor with moving cameras in PoznanHall), and also in arrangement and number of cameras (e.g. matrix of 16 cameras in TechnicolorPainter, 13 linearly-arranged cameras in IntelFrog). The methods used for estimation of original depth maps were also different, what tested the versatility of compared refinement techniques. All experiments were performed on one thread of Intel Core i7-5820K CPU (3.3 GHz clock) machines equipped with 64 GB of operational memory.

2) RESULTS

Results of encodings of the abovementioned sequences are presented in tables that show changes of BD-rates for high and low bitrates (i.e., BD-rates calculated for 4 lowest and 4 highest QPs). In Table 1, for the encoding of sequences with depth maps refined with the proposal, all objective metrics show a significant decrease of BD-rate in comparison with original, non-refined depth maps originally provided with test sequences. On average, after the refinement of depth maps using the proposal, test sequences can be encoded with about a 50% reduction of the bitrate.

MIV utilizes depth information to minimize the redundancy between input views of the scene. Such a great reduction of bitrate shows that inter-view consistency of depth maps was significantly increased. It is shown in Fig. 4, that in original depth maps, the depth map for the toy in the foreground is not consistent in different views. After performing the proposed refinement, depth maps are consistent in both presented views. It is not a case for depth maps after the bilateral filtering, shown in Fig. 4 f) and i), as this refinement is performed independently for each view.

Figs. 5 – 9 show that the proposal both decreases the bitrate and increases the quality of synthesized views. It was observed for all tested sequences. First of all, the bitrate required to send the whole representation of a scene to e.g. a head-mounted display is significantly reduced. It can decrease the streaming delay, which influences the subjective quality of immersion [27]. Moreover, as the quality of synthesized virtual views for natural content is still usually worse than for CGI content, the increase of the quality is of high importance for immersive video applications. For the bilateral filtering and synthesis-based refinement, the decrease of the required bitrate also can be seen in the provided figures, nevertheless, the quality varies significantly. For bilateral filter, the quality of rendered virtual views is often lower than for the original, not refined depth maps.

Table 2 presents BD-rate changes for the encoding of sequences with depth maps refined with the proposal in comparison with depth maps refined with [1]. For the Y-PSNR

TABLE 1. BD-Rate Change for the Encoding of Sequences With Depth Maps Refined With the Proposal in Comparison With Original, Non-Refined Depth Maps (Low Bitrate – Calculated for 4 Lowest QPs, High Bitrate – 4 Highest QPs).

Test sequence	Bitrate	Change of synthesis BD-rate (original depth maps vs. the proposal)			
		Y-PSNR	VMAF	MS-SSIM	IV-PSNR
Intel	High	-44.42%	-29.29%	-33.67%	-48.15%
Frog	Low	-21.35%	-14.26%	-17.20%	-27.79%
Technicolor	High	-77.98%	-32.95%	-50.26%	-69.04%
Painter	Low	-58.79%	-27.51%	-39.25%	-56.59%
Poznan	High	-56.01%	-58.58%	-35.42%	-19.66%
Carpark	Low	-33.03%	-38.05%	-16.30%	-10.03%
Poznan	High	-92.22%	-22.18%	-73.74%	-91.64%
Hall	Low	-61.39%	-11.03%	-46.72%	-85.99%
Poznan	High	-51.96%	-57.99%	-35.74%	-31.23%
Street	Low	-32.27%	-36.68%	-22.55%	-21.85%
Average:	High	-64.52%	-40.20%	-45.77%	-51.94%
	Low	-41.37%	-25.51%	-28.40%	-40.45%

TABLE 2. BD-Rate Change for the Encoding of Sequences With Depth Maps Refined With the Proposal in Comparison With Depth Maps Refined With Synthesis-Based Refinement (Low Bitrate – Calculated for 4 Lowest QPs, High Bitrate – 4 Highest QPs).

Test Sequence	Bitrate	Change of synthesis BD-rate (depth maps refined with [1] vs. the proposal)			
		Y-PSNR	VMAF	MS-SSIM	IV-PSNR
Intel	High	-23.22%	-5.86%	-20.59%	-32.29%
Frog	Low	-4.88%	4.08%	-3.16%	-12.75%
Technicolor	High	-8.93%	22.62%	0.05%	-18.43%
Painter	Low	-12.78%	0.99%	-10.55%	-14.82%
Poznan	High	-30.92%	-15.93%	-5.87%	10.92%
Carpark	Low	-5.05%	7.26%	18.31%	25.11%
Poznan	High	-81.81%	-19.53%	-14.24%	-49.06%
Hall	Low	-58.69%	-6.39%	-1.25%	-42.42%
Poznan	High	-15.32%	-10.10%	-0.52%	21.84%
Street	Low	-3.64%	0.14%	8.02%	20.82%
Average:	High	-32.04%	-5.76%	-8.23%	-13.40%
	Low	-17.01%	1.22%	2.27%	-4.81%

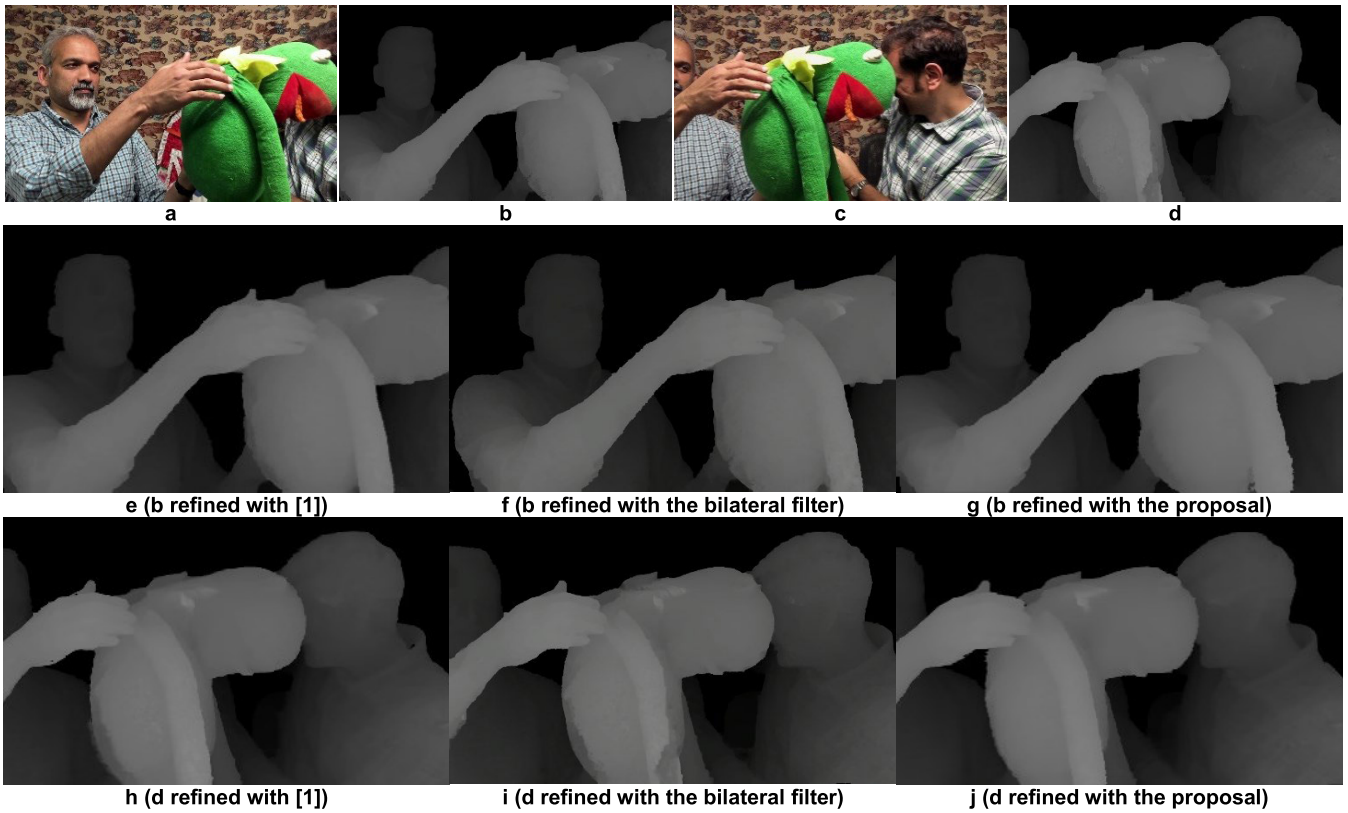

FIGURE 4. Comparison of input views and depth maps for IntelFrog test sequence: (a) input view 3, (b) original depth map for view3, (c) input view 10 (d) original depth map for view 10, (e) depth map b refined with [1], (f) depth map b refined with bilateral filter, (g) depth map b refined with proposal, (h) depth map d refined with [1], (i) depth map d refined with bilateral filter, (j) depth map d refined with the proposal.

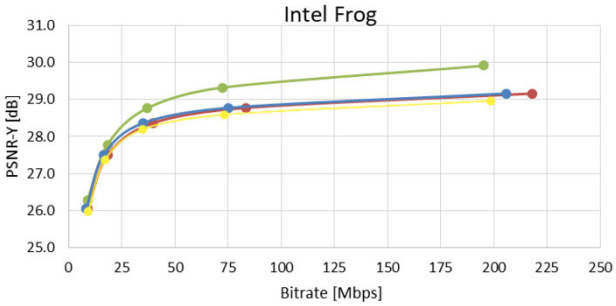

FIGURE 5. PSNR-Y RD-curves for IntelFrog sequence for encoding with original depth maps (red curve), with depth maps refined with the proposal (green), synthesis-based refinement (blue), and bilateral filter (yellow).

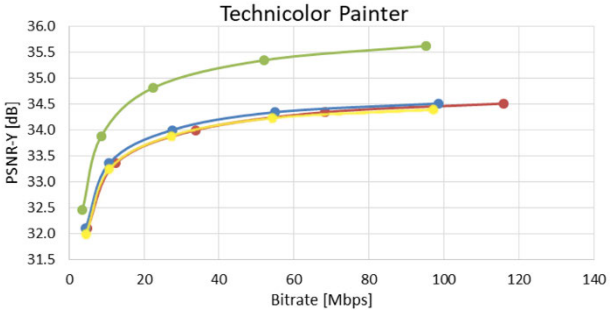

FIGURE 6. PSNR-Y RD-curves for TechnicolorPainter sequence for encoding with original depth maps (red curve), with depth maps refined with the proposal (green), synthesis-based refinement (blue), and bilateral filter (yellow).

metric, the proposal shows significantly better performance than the synthesis-based refinement. For other metrics, differences are smaller, but on average, the proposal also achieves better results. Figs. 5 – 9 show that synthesis-based refinement [1] reduces the bitrate similarly to the proposal but does not increase the quality of virtual view synthesis. As described in Section II, the method [1] is based on the iterative process of depth refinement and view synthesis. Method [1] optimizes the quality of view synthesis, but only for the synthesizer that is implemented within that method. It decreases the possible applications of such a method.

The proposal uses only depth maps, therefore, it is not influenced by possible noise and color inconsistencies in input views. Moreover, as it is shown in Table 3, the processing time for the proposal is on average more than 30 times shorter than for synthesis-based refinement and is equal to about 30 seconds per frame for all views. Such time of processing is similar to the time of estimation of depth maps in state-of-the-art methods (e.g. [6]), therefore, the proposed refinement does not significantly increase the overall time of processing. However, the further reduction of the runtime is possible with the simple use of multiple CPU cores. The proposed method is fully parallelizable, as the processing can

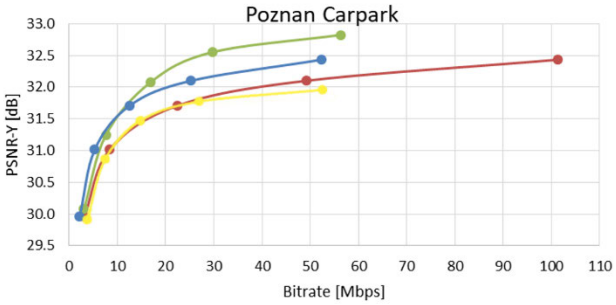

FIGURE 7. PSNR-Y RD-curves for PoznanCarpark sequence for encoding with original depth maps (red curve), with depth maps refined with the proposal (green), synthesis-based refinement (blue), and bilateral filter (yellow).

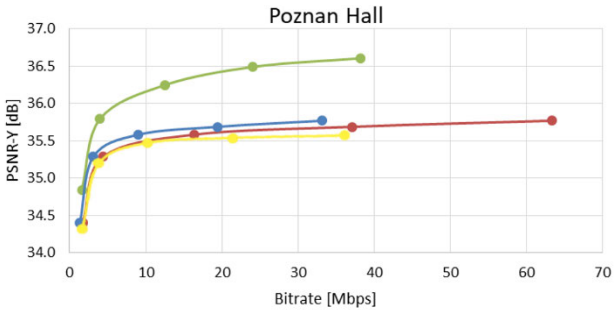

FIGURE 8. PSNR-Y RD-curves for PoznanHall sequence for encoding with original depth maps (red curve), with depth maps refined with the proposal (green), synthesis-based refinement (blue), and bilateral filter (yellow).

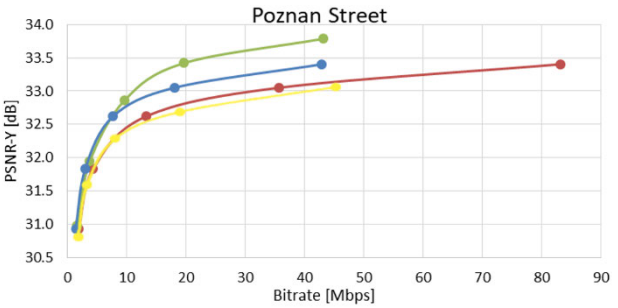

FIGURE 9. PSNR-Y RD-curves for PoznanStreet sequence for encoding with original depth maps (red curve), with depth maps refined with the proposal (green), synthesis-based refinement (blue), and bilateral filter (yellow).

TABLE 3. Processing Times for Both Refinements.

Test sequence	Processing time [s]		
	Synthesis-based refinement	Bilateral filter*	Proposal
Intel Frog	967	15	35
Technicolor Painter	656	18	71
Poznan Carpark	2101	14	15
Poznan Hall	1391	13	17
Poznan Street	802	16	14
Average:	1183	15	30

* processing time per one view

be split in order to process each view (each depth map) in a separate thread without any quality degradation. Moreover, because of no need for synchronization of data between different views (different threads), the time reduction factor will be close to the number of views, e.g. 16 for Technicolor Painter. It would result in less than 2.5 seconds per frame for refining all views.

The bilateral filter is performed independently for each view, therefore, its processing times in Table 3 are given per one view, not for all views as for other methods.

B. ASSESSMENT OF THE REFINEMENT ACCURACY

1) DESIGN OF THE EXPERIMENT

The proposed refinement method was designed for immersive video applications, which are characterized by large numbers of input views. Therefore, in order to provide a direct evaluation of the accuracy of depth maps, we use the Middlebury database [16], with two high-resolution (1800×1500) multiview images: Cones and Teddy, for which 9 views are available. Such a scenario to some degree meets the characteristics of typical immersive video, therefore, it can provide fair quantitative results for the presented method.

First of all, for two used multiview images, depth maps were estimated using [6]. Used parameters of estimation were purposely chosen to result in noisy depth maps that are not highly inter-view consistent. Then, these estimated depth maps were refined using three tested methods. Note that any depth map estimation method can be used with these methods, authors decided to use their method [6], as its source code is publicly available.

2) RESULTS

We present the percentage of bad pixels of the estimated depth maps summarized over for all available pixels of ground-truth depth maps for the error threshold of $E_T = 2.0$ and $E_T = 4.0$ (i.e., if the absolute error of estimated depth value for a pixel is larger than E_T then this pixel is considered as a bad pixel), and an average error. Fig. 10 and 11 show the ground-truth, estimated and refined depth maps, and visualizations of bad pixels for $E_T = 4.0$. The results are presented in Table 4. In immersive video applications, the accuracy of depth maps is much less important than their inter-view consistency and quality of virtual view synthesis. Nevertheless, as achieved results show, the proposed method improves also the accuracy of depth maps, both in terms of the percentage of bad pixels and the average error. For the synthesis-based refinement [1], the percentage of bad pixels for one of the test images was not decreased.

When bilateral filtering is used, the improvement is higher than for synthesis-based refinement, but still, the accuracy of depth maps is lower than in the proposal. It shows that these two methods less versatile than the proposal, as their efficiency varies in different tests.

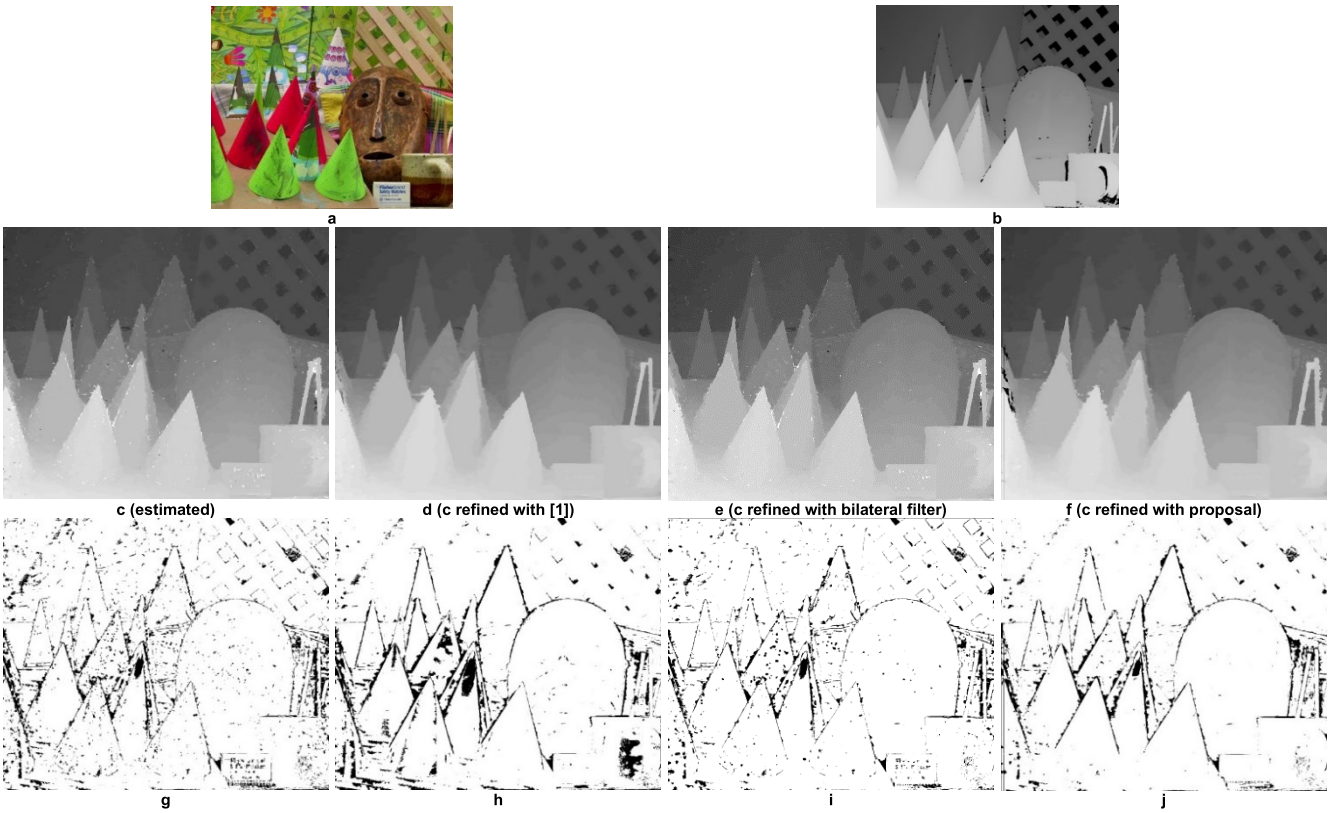

FIGURE 10. Comparison of depth maps for Cones test images: (a) input image. Depth map: (b) ground-truth, (c) estimated, (d) refined with synthesis-based refinement, (d) bilateral filter (f) proposal. Bad pixels of depth map: (g) estimated, (h) refined with synthesis-based refinement, (i) bilateral filter, (j) proposal.

TABLE 4. The Results of the Assessment of the Accuracy of Depth Maps Refined Using The Proposed Method, Bilateral Filter and Synthesis-Based Refinement [1] on the 9 Views High-Resolution Middlebury Dataset.

Test images	Depth refinement method	Percentage of bad pixels		Average error
		$E_T = 2.0$	$E_T = 4.0$	
Teddy	No refinement	44.58%	16.63%	3.27
	Refinement [1]	37.73%	12.21%	2.88
	Bilateral filter	37.43%	12.59%	2.76
	Proposal	**30.84%**	**10.77%**	**2.65**
Cones	No refinement	31.56%	8.10%	2.93
	Refinement [1]	35.01%	8.61%	2.73
	Bilateral filter	28.20%	6.17%	2.71
	Proposal	**22.77%**	**5.82%**	**2.68**

C. ASSESSMENT OF THE INFLUENCE ON THE TEMPORAL CONSISTENCY

1) DESIGN OF THE EXPERIMENT

As it was presented earlier in [6], the size of a virtual view after encoding is one of the objective measures of temporal consistency of depth maps used in virtual view synthesis. The lower the temporal consistency of depth maps, the lower the efficiency of the encoding of virtual views.

In this experiment, the same test sequences as in Section V-A were used for the evaluation of tested methods. For all sequences, the virtual view was placed in the same position as the central view. The virtual view synthesis was performed using Versatile View Synthesizer 2.0 [48], which used four nearest views to synthesize the desired virtual view.

The video encoder used for the encoding of virtual views was set in the low-delay mode, so only the first frame of virtual views was encoded as an intra frame. Such settings of the encoder increase the influence of temporal consistency of the encoded sequence on the final bitrate. In the experiments, the HEVC encoder (HM 16.15 framework [39]) was used with the MPEG common test conditions (with the exception of used test sequences) and software reference configurations.

2) RESULTS

Table 5 presents the results of the experiment, which are expressed as average luma bitrate reductions calculated using the BD-rate [49] metric in comparison to virtual views synthesized with original, not refined depth maps.

For all tested methods, the bitrate was on average decreased, but synthesis-based refinement showed the least stable results (for one sequence the bitrate was significantly increased). The proposal achieved the highest average reduction of the bitrate of encoded virtual views, proving the increase of temporal consistency of depth maps after using this method.

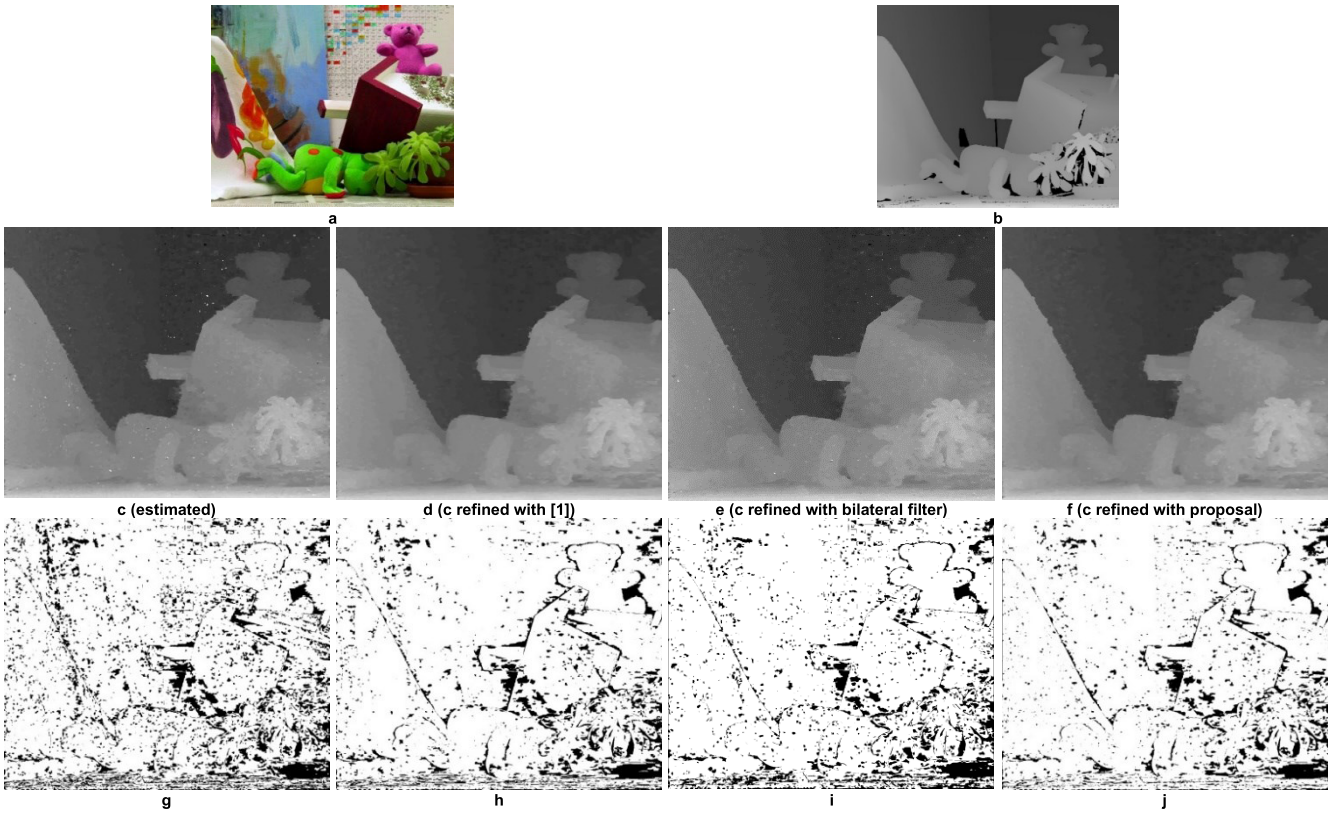

FIGURE 11. Comparison of depth maps for Teddy test images: (a) input image. Depth map: (b) ground-truth, (c) estimated, (d) refined with synthesis-based refinement, (d) bilateral filter (f) proposal. Bad pixels of depth map: (g) estimated, (h) refined with synthesis-based refinement, (i) bilateral filter, (j) proposal.

TABLE 5. Average Luma Bitrate Reductions of Encoded Virtual Views Synthesized Using Depth Maps Refined With Different Methods.

Test sequence	Encoded virtual views bitrate reduction compared to virtual views synthesized using not refined depth maps		
	Synthesis-based refinement	Bilateral filter	Proposal
Intel Frog	14.07%	-6.28%	-6.41%
Technicolor Painter	-0.84%	-8.11%	-6.94%
Poznan Carpark	-18.65%	-13.95%	-19.21%
Poznan Hall	-11.73%	-10.61%	-13.27%
Poznan Street	-16.80%	-3.91%	-3.02%
Average:	-6.79%	-8.57%	-9.77%

In this test, the bilateral filter showed a performance comparable to the proposal. When confronted with the experiment from Section V-A, where the bilateral filter achieved the worst results, it shows that for immersive video encoding the inter-view consistency is more crucial than the temporal consistency for achieving high objective quality for the final viewer.

VI. CONCLUSION

This article describes a depth map refinement method aimed at the increase of the quality of the immersive video. In the described method, only information from depth maps is used, as the use of texture can introduce errors in the refinement, mostly due to inter-view color inconsistencies, noise, and compression artifacts.

In order to evaluate the performance of the proposal and compare it with the state-of-the-art, two experiments were conducted. To test the influence of the refinement on the encoding of immersive video, the MPEG Immersive Video encoder was used. As the results show, the proposal allows, first of all, increasing the quality of virtual views synthesized for a viewer of immersive video. Moreover, by enhancing the inter-view consistency of depth maps, also crucial for further improving the fidelity of the virtual views synthesis process, the performance of the used encoder was also highly improved. Moreover, the evaluation of the time of processing shows that the proposal does not significantly increase the overall time of processing as is similar to the time of estimation of depth maps. Another important aspect of the article can be seen also in the proposed design of experiments itself, which is based on using an immersive video encoder for the evaluation. This encoding process is heavily influenced by inter-view consistency of depth maps, what makes it very valuable for future comparisons of depth maps estimation or refinement methods.

In the second experiment, in order to provide a direct evaluation of the accuracy of depth maps, we used multi-view images from the widely-recognized Middlebury database. The experiments showed higher similarity to ground-truth

after the refinement of estimated depth maps. The last experiment showed also the increase of the temporal consistency of refined depth maps.

The particular usefulness of the method for immersive video is a result of merging all the abovementioned features into one, easy-to-use software. The proposal became the MPEG Reference Software for the depth refinement, as decided by the experts of the ISO/IEC MPEG group for immersive video. The method is also available in the public repository, so the source code of the implementation can be used by other researchers as a new reference for their future works.

REFERENCES

[1] M. Kurc, O. Stankiewicz, and M. Domanski, "Depth map inter-view consistency refinement for multiview video," in *Proc. Picture Coding Symp.*, May 2012, pp. 137–140.

[2] Y. Yang, Q. Liu, X. He, and Z. Liu, "Cross-view multi-lateral filter for compressed multi-view depth video," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 302–315, Jan. 2019.

[3] J. Yang, X. Ye, and P. Frossard, "Global auto-regressive depth recovery via iterative non-local filtering," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 123–137, Mar. 2019.

[4] J. Choi, D. Min, and K. Sohn, "Reliability-based multiview depth enhancement considering interview coherence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 603–616, Apr. 2014.

[5] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, "Depth restoration from RGB-D data via joint adaptive regularization and thresholding on manifolds," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1068–1079, Mar. 2019.

[6] D. Mieloch, O. Stankiewicz, and M. Domanski, "Depth map estimation for free-viewpoint television and virtual navigation," *IEEE Access*, vol. 8, pp. 5760–5776, 2020.

[7] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, Aug. 2018.

[8] *Manual of Depth Refinement Software PDR*, document ISO/IEC JTC1/SC29/WG11 MPEG/N18708, Gothenburg, Sweden, Jul. 2019.

[9] *Common Test Conditions for Immersive Video*, document ISO/IEC JTC1/SC29/WG11 MPEG/N18789, Geneva, Switzerland, Oct. 2019.

[10] L. Fang, Y. Xiang, N.-M. Cheung, and F. Wu, "Estimation of virtual view synthesis distortion toward virtual view position," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1961–1976, May 2016.

[11] 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 Trans. Image Process.*, vol. 23, no. 1, pp. 185–199, Jan. 2014.

[12] N. Vretos and P. Daras, "Temporal and color consistent disparity estimation in stereo videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 3798–3802.

[13] O. Stankiewicz, M. Domanski, A. Dziembowski, A. Grzelka, D. Mieloch, and J. Samelak, "A free-viewpoint television system for horizontal virtual navigation," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2182–2195, Aug. 2018.

[14] M. Domanski, O. Stankiewicz, K. Wegner, and T. Grajek, "Immersive visual media—MPEG-I: 360 video, virtual navigation and beyond," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2017, pp. 1–9.

[15] Z. Lee and T. Q. Nguyen, "Multi-array camera disparity enhancement," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2168–2177, Dec. 2014.

[16] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 195–202.

[17] X. Wang, C. Zhu, S. Li, J. Xiao, and T. Tillo, "Depth filter design by jointly utilizing spatial-temporal depth and texture information," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Jun. 2015, pp. 1–5.

[18] E. Ekmekcioglu, V. Velisavljevic, and S. T. Worrall, "Content adaptive enhancement of multi-view depth maps for free viewpoint video," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 2, pp. 352–361, Apr. 2011.

[19] C.-H. Wei, C.-K. Chiang, and S.-H. Lai, "Iterative depth recovery for multi-view video synthesis from stereo videos," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Asia–Pacific*, Dec. 2014, pp. 1–8.

[20] D. Mieloch, "Depth estimation in free-viewpoint television," Ph.D. dissertation, Dept. Electron. Telecommun., Chair Multimedia Telecommun. Microelectron., Poznan Univ. Technol., Poznan, Poland, 2018.

[21] *Test model 3 for Immersive Video*, document ISO/IEC JTC1/SC29/WG11 MPEG/ N18795, Geneva, Switzerland, Oct. 2019.

[22] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec, "Immersive light field video with a layered mesh representation," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 86:1–86:15, 2020.

[23] H.-A. Hsu, C.-K. Chiang, and S.-H. Lai, "Spatio-temporally consistent view synthesis from video-plus-depth data with global optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 74–84, Jan. 2014.

[24] B. Salahieh, B. Marvar, M. Nentedem, A. Kumar, V. Popvic, K. Seshadrinathan, O. Nestares, and J. Boyce, *Kermit test sequence for Windowed 6DoF Activities*, document ISO/IEC JTC1/SC29/WG11 MPEG/M43748, Ljublana, Slovenia, Jul. 2018.

[25] D. Doyen, G. Boisson, and R. Gendrot, *[MPEG-I Visual] New Version of the Pseudo-Rectified Technicolorpainter Content*, document ISO/IEC JTC1/SC29/WG11 MPEG/M43366, Ljublana, Slovenia, Jul. 2018.

[26] D. Mieloch, A. Dziembowski, and M. Domański, *[MPEG-I Visual] Natural Outdoor Test Sequences*, document ISO/IEC JTC1/SC29/WG11 MPEG/M51598, Brussels, Belgium, Jan. 2020.

[27] A. Grzelka, A. Dziembowski, D. Mieloch, O. Stankiewicz, J. Stankowski, and M. Domanski, "Impact of video streaming delay on user experience with head-mounted displays," in *Proc. Picture Coding Symp. (PCS)*, Nov. 2019, pp. 1–5.

[28] *Text of ISO/IEC CD 23090-12 MPEG Immersive Video*, document ISO/IEC JTC1/SC29/WG11 MPEG2020/N19482, Jul. 2020.

[29] L. Cui, R. Mekuria, M. Preda, and E. S. Jang, "Point-cloud compression: Moving picture experts Group's new standard in 2020," *IEEE Consum. Electron. Mag.*, vol. 8, no. 4, pp. 17–21, Jul. 2019.

[30] T. Senoh, N. Tetsutani, and H. Yasuda, "Depth estimation and view synthesis for immersive media," in *Proc. Int. Conf. 3D Immersion (IC3D)*, Dec. 2018, pp. 1–8.

[31] C. Guo, J. Jin, J. Hou, and J. Chen, "Accurate light field depth estimation via an occlusion-aware network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[32] J. Chakareski, "UAV-IoT for next generation virtual reality," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5977–5990, Dec. 2019.

[33] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuca, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko, "Emerging MPEG standards for point cloud compression," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 133–148, Mar. 2019.

[34] Z. Lai, Y. C. Hu, Y. Cui, L. Sun, N. Dai, and H.-S. Lee, "Furion: Engineering high-quality immersive virtual reality on Today's mobile devices," *IEEE Trans. Mobile Comput.*, vol. 19, no. 7, pp. 1586–1602, Jul. 2020.

[35] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[36] C.-C. Lee, A. Tabatabai, and K. Tashiro, "Free viewpoint video (FVV) survey and future research direction," *APSIPA Trans. Signal Inf. Process.*, vol. 4, pp. 1–10, Oct. 2015.

[37] M. Tanimoto, M. Panahpour Tehrani, T. Fujii, and T. Yendo, "FTV for 3-D spatial communication," *Proc. IEEE*, vol. 100, no. 4, pp. 905–917, Apr. 2012.

[38] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[39] *HEVC Reference Codec*. Accessed: Mar. 12, 2020. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[40] G. Wu, Y. Li, Y. Huang, and Y. Liu, "Joint view synthesis and disparity refinement for stereo matching," *Frontiers Comput. Sci.*, vol. 13, no. 6, pp. 1337–1352, Dec. 2019.

[41] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.

[42] T.-W. Hui, C. C. Loyd, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 353–369.
[43] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.
[44] Q. Yang, N. Ahuja, R. Yang, K.-H. Tan, J. Davis, B. Culbertson, J. Apostolopoulos, and G. Wang, "Fusion of median and bilateral filtering for range image upsampling," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4841–4852, Dec. 2013.
[45] Q. Yang, "Hardware-efficient bilateral filtering for stereo matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 1026–1032, May 2014.
[46] P. Garus, J. Jung, T. Maugey, and C. Guillemot, "Bypassing depth maps transmission for immersive video coding," in *Proc. Picture Coding Symp. (PCS)*, Nov. 2019, pp. 1–5.
[47] A. Dziembowski, M. Domanski, A. Grzelka, D. Mieloch, J. Stankowski, and K. Wegner, "The influence of a lossy compression on the quality of estimated depth maps," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.
[48] P. Boissonade and J. Jung, *[MPEG-I Visual] Improvement of VVS1.0.1*, document ISO/IEC JTC1/SC29/WG11/MPEG2019/m46263, Marrakesh, Morocco, Jan. 2019.
[49] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD986 Curves*, document ISO/IEC JTC 1/SC 29/WG, 11, Doc. MPEG M15378, Austin, TX, USA, 2001.
[50] F. Isgro, E. Trucco, P. Kauff, and O. Schreer, "Three-dimensional image processing in the future of immersive media," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 288–303, Mar. 2004.
[51] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Nettflix Technol. Blog.*, to be published. Accessed: Mar. 12, 2020. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[52] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.
[53] *Software Manual of IV-PSNR for Immersive Video*, document ISO/IEC JTC1/SC29/WG11 MPEG/N18709, Göteborg, Sweden, Jul. 2019.

DAWID MIELOCH received the M.Sc. and Ph.D. degrees from the Poznań University of Technology, in 2014 and 2018, respectively. He is currently an Assistant Professor with 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 3-D video processing. His current research interests include free-viewpoint television, depth estimation, and camera calibration.

ADRIAN DZIEMBOWSKI was born in Poznań, Poland, in 1990. He received the M.Sc. and Ph.D. degrees from the Poznań University of Technology, in 2014 and 2018, respectively. Since 2019, he has been an Assistant Professor with the Institute of Multimedia Telecommunications. He has authored or coauthored about 30 articles on various aspects of immersive video, free navigation, and free viewpoint television systems. He is also actively involved in ISO/IEC MPEG activities towards future MPEG immersive video coding standard.

MAREK DOMAŃSKI (Senior Member, IEEE) received the M.Sc., Ph.D., and Habilitation degrees from the Poznań University of Technology, Poland, in 1978, 1983, and 1990, respectively. Since 1993, he has been a Professor with the Poznań University of Technology, where he leads the Institute of Multimedia Telecommunications. He has coauthored one of the very first AVC decoders for tv set-top boxes, in 2004, and highly ranked technology proposals to MPEG for scalable video compression, in 2004, and 3D video coding, in 2011. He authored three books and more than 300 articles in journals and conference proceedings. His research interests include image, video and audio compression, virtual navigation, free-viewpoint television, image processing, multimedia systems, 3D video and color image technology, digital filters, and multidimensional signal processing. He served as a member for various steering, program, and editorial committees of international journals and international conferences. He was a general chairman/co-chairman and host of several international conferences: Picture Coding Symposium, PCS 2012; IEEE International Conference on Advanced & Signal based Surveillance, AVSS 2013, European Signal Processing Conference, EUSIPCO 2007; 73rd and 112nd Meetings of MPEG; International Workshop on Signals, Systems and Image Processing, IWSSIP 1997 and 2004; International Conference on Signals and Electronic Systems, ICSES 2004, and others.