High Efficiency 3D Video Coding Using New Tools Based on View Synthesis

Marek Domański, *Member, IEEE*, Olgierd Stankiewicz, Krzysztof Wegner, Maciej Kurc, Jacek Konieczny, Jakub Siast, Jakub Stankowski, Robert Ratajczak, Tomasz Grajek

***Abstract*—We propose a new coding technology for 3D video represented by multiple views and the respective depth maps. The proposed technology is demonstrated as an extension of the recently developed High Efficiency Video Coding (HEVC). One base view is compressed into a standard bitstream (like in HEVC). The remaining views and the depth maps are compressed using new coding tools that mostly rely on view synthesis. In the decoder, those views and the depth maps are derived via synthesis in the 3D space from the decoded base view and from data corresponding to small disoccluded regions. The shapes and locations of those disoccluded regions can be derived by the decoder without any side information transmitted. In order to achieve high compression efficiency, we propose several new tools like Depth-Based Motion Prediction, Joint High Frequency Layer Coding, Consistent Depth Representation and Nonlinear Depth Representation. The experiments show high compression efficiency of the proposed technology. The bitrate needed for transmission of even two side views with depth maps is mostly below 50% of the bitrate for a single-view video.**

***Index Terms*—3D video, coding, compression, MVD representation, HEVC, depth maps.**

I. Introduction

Recently, 3D video technology is evolving towards new systems that include glassless displays and provide realistic impression of depth as well as controllable stereoscopic base-line distance. In such 3D video systems the representation of a 3D scene should be richer than a stereo pair. Therefore, a need has been identified for new video formats and for a respective compression standard. This need was discussed in the ISO/IEC Moving Pictures Experts Group (MPEG) [1], which decided to start the respective standardization project on compression of 3D video in the multiview plus depth format (the MVD format). The features, possibilities and importance of this format have been already discussed in many papers, e.g. [2-9], therefore these issues will be omitted in this paper, which is focused on compression of 3D video in this format.

Some applications will need numerous views to be available at the receiver. For example, future autostereoscopic displays are expected to present simultaneously even 50 different views corresponding to virtual cameras with parallel optical axes spaced within few human inter-ocular distances. Such dense spacing of the views yields strong similarity between the neighboring views that can be exploited for compression. Moreover, in the receiver many virtual views may be efficiently synthesized using the Depth-Image-Based Rendering (DIBR) [3,10,11] and for transmission the MVD format often may be limited to only 2-3 views accompanied with the corresponding depth maps [1]. In a realistic example of a system with an autostereoscopic display only 3 views with 3 depth maps are transmitted (Fig. 1).

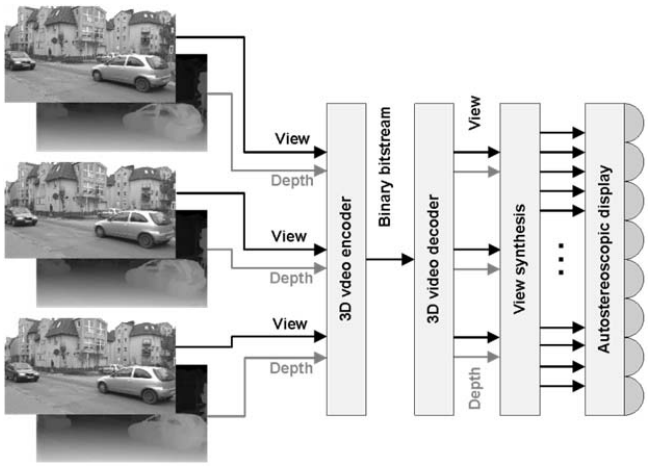


Fig. 1. An example of a 3D video system where 3 views with 3 depth maps are transmitted and used for synthesis of many virtual views.

MPEG has also identified the requirement that the prospective 3D video coding technology should exploit either Multiview Video Coding (MVC) or High Efficiency Video Coding (HEVC) standards [1]. The MVC standard [12] is an extension of the Advanced Video Coding (AVC) standard [13] and it takes advantage of additional inter-view disparity-compensated predictions and provides bitrate reduction of up to 20-30% as compared to the independent AVC-compliant compression of all views. HEVC is currently at the final stage of standardization by ISO and ITU [14]. It provides about 50% bitrate reduction as compared to AVC. Independent HEVC compression of multiple views results in 30% bitrate reduction as compared to MVC [15].

With the aim to search for a new technology that would be potentially used as a base for the prospective international

Manuscript received October 15, 2012; revised March 21, 2013. This work was supported by public funds in the years 2010-2012.

The authors are with the Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology, 60965 Poznań, Poland (e-mail: domanski@et.put.poznan.pl, ostank@multimedia.edu.pl, kwegner@multimedia.edu.pl).

Digital Object Identifier

standard, in early 2011 MPEG announced a Call for Proposals (CfP) on 3D Video Coding Technology [16]. Each proposed technology had to meet the condition that a standard bitstream (either AVC-compliant or HEVC-compliant) for a single view may be extracted from the entire bitstream.

In response to the CfP, over 20 proposals were submitted in two categories: backward compatible with AVC and compatible with HEVC. The proposals were ranked using subjective quality assessment of the decoded and rendered video clips. The assessments were done in a large experiment that involved 12 laboratories and over 600 viewing subjects who assessed about 2700 video clips. The results were disclosed during the MPEG meeting in Geneva in November 2011. In the HEVC-compatible class, the proposals from Fraunhofer Institute – H. Hertz Institute, Berlin, and from Poznań University of Technology, were qualified as the best performing ones.

The latter proposal exploits dependencies between views and depth maps in order to increase the overall coding efficiency (similarly as in [17-19]). The 3D scene information contained in depth maps was also already used to speed up the encoder mode decision [9]. Other approaches use special techniques for depth coding [7,8]. Some of them rely on independent coding of depth that allows to use the standard coding technology for the MVD format but provides limited compression efficiency.

In 2012, the standardization projects for MVD video ware developed by MPEG and by ITU-T/ISO/IEC Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V). For MVC, independent depth coding [20] was included into a standard draft [21]. A more sophisticated MVC extension for MVD is also under development [22]. The MVC-like inter-view disparity-compensated prediction was also implemented on top of the HEVC technology. The adoption of MVC coding scheme to HEVC provides coding gains similar or better than MVC gains over AVC [23]. The respective standard draft is under development [24]. In the course of these standardization activities, some new tools described in this paper were also implemented in the test models [25,26] and used as a starting point for development of 3D video coding standards.

II. Assumptions and Overview of the Paper

The paper presents a detailed description of the compression technology proposed by Poznań University of Technology in the response to the call [16] of MPEG. The novelty of this proposal consists in new coding tools and in a the selection of the tools resulting from extensive experiments. Hitherto, this technology was only briefly described in an MPEG document [27] and in conference contributions [28-30] that concerned some selected aspects only. The aim of this paper is to present the entire proposed technology and the most important coding tools. Some tools have been already described in the conference papers [19,31]. For those tools only brief descriptions are given in this paper; more details can be found in the references. This paper also presents the new extensive experimental results obtained by the authors.

The proposed codec is compliant with the requirements that were defined by MPEG in the Call for Proposals (CfP) [1,16], and which resulted from studies of potential applications. In particular, one of the views – called the base view – is coded in compatibility with HEVC syntax, which allows extraction of a base view by a legacy decoder (Fig. 2). The remaining views are called the side views. These views and all depth data are coded with the use of new tools that are described further in this paper. It may be pointed out that encoding and decoding of the side views and depth maps exploits information from the decoded base view.

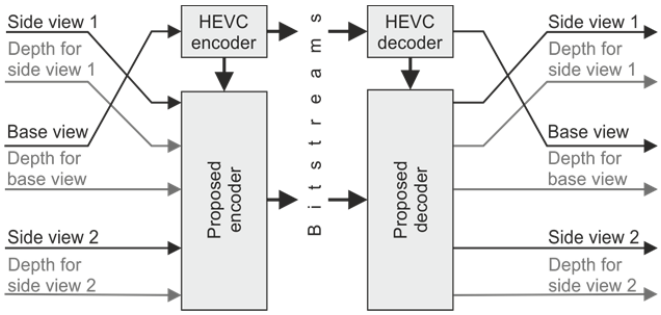


Fig. 2. Proposed codec structure for 3-view MVD (views and their depth maps are denoted in black and gray, respectively).

For the sake of conciseness, further considerations will be given only for the case of 3 views; however. those considerations can be generalized onto an arbitrary number of views. In an independent comparisons of the responses to MPEG CfP [16], the proposed technology has been proved high to have compression efficiency for two coding scenarios: 2 views with 2 depth maps and 3 views with 3 depth maps. For encoding of 3 views, using exhaustive experiments, the we found that the choice of the central view as the base view is usually optimal from the point of view of the overall rate-distortion efficiency. Detailed discussion of this issue is left beyond the scope of the paper.

The next sections of the paper deal with the new tools developed by the authors. These considerations start with Disoccluded Region Coding (Section III) that is the most important coding tool in the proposed 3D video compression technology. Then, two other tools are described: Joint High-Frequency Layer Coding and Depth-Based Motion Prediction (Sections IV and V, respectively). The common feature of these tools is that they all use view synthesis at some stage. Then, two tools for efficient depth representation are considered in Section VI. The first of these tools – Consistent Depth Representation – also exploits view synthesis at a certain step of the depth representation production. The proposed compression technology uses various new tools that exploit view synthesis at some stage. The entire proposed compression technology is described in Section VII. The experimental results for compression performance (Section VIII) are followed by the general conclusions (Section IX).

III. Disoccluded Region Coding

As already mentioned in Introduction, the encoded views are assumed to correspond to the camera locations that are relatively close one to each other. Therefore the views are highly correlated, and the coding tools that bring most of the compression gains are those that exploit inter-view similarities. Of course, those tools are only used in the side views and are not used for the base view that is encoded using the standard HEVC technique. Here, we are going to describe such a new tool called Disoccluded Region Coding, which was proposed by the authors in order to obtain very efficient representations of the side views [29,30].

In Disoccluded Region Coding, view synthesis is used as a primary inter-view prediction mechanism. With reference to the already encoded views and the respective depth maps, a virtual side view is synthesized in the position of the view currently being coded. For the case of 3-view coding, the side views are synthesized from the base view (Fig. 3). For view synthesis the Depth-Image-Based Rendering (DIBR) is used. It is assumed that the intrinsic and extrinsic parameters of cameras are present in the transmitted bitstream for all views. In the experiments (see Section VIII) the we have used a state-of-the-art DIBR implementation developed by MPEG in View Synthesis Reference Software [32].

In the straightforward implementation of the view-synthesis prediction, residuals of the prediction (i.e. prediction errors) are sent to the decoder. Alternatively, the omission of the residuals is signaled in the bitstream. A major drawback of such a solution is considerable yet unjustified transmission payload that either comes from the transmitted residuals or from the signaling of their presence or absence.

Disoccluded Region Coding was proposed in order to omit this drawback. Its idea is based on the observation that a side view consists of two types of distinctive regions (Fig. 3):

- Regions that can be synthesized in the decoder from the decoded base view also usually no prediction error needs to be transmitted.
- Disoccluded regions that cannot be synthesized from the base view and must be transmitted in the bitstream.

The locations and borders of those disoccluded regions can be estimated as a side-product of the DIBR-based synthesis of the side view. After synthesis of a side view, the pixels in an occluded region remain undefined. In a frame of a synthesized view, all pixels with undefined values can be easily detected. In that way, the locations and borders of those disoccluded regions are derived using the already transmitted base view and its depth map. This is doable in the same way both in the encoder and in the decoder. Therefore, there is no need to transmit the locations and borders of the disoccluded regions.

The above-mentioned observation yields that that a side view consists of two distinctive regions:

- Synthesizable regions that are not transmitted but synthesized in the decoder.
- Disoccluded regions are transmitted in the bitstream.

The encoding and decoding of the disoccluded regions is performed in Coding Units (CUs) which are defined in the HEVC draft standard [14] as generalization of macroblocks. The CUs are rectangular blocks of variable size between 64×64 and 8×8 pixels. If a given CU can be entirely synthesized from the base view, it is not transmitted at all. On the contrary, if there are any disoccluded pixels (i.e., pixels that cannot be synthesized from the base view) inside the given CU, it is encoded (Figs. 3 and 4). Because in a typical case most of the scene is the same in all views, only small portions of pictures are disoccluded in subsequently coded views, and thus only a small number of CUs is encoded (Figs. 3 and 4). In the decoder, most of CUs are reconstructed using view-synthesis prediction from the base view. Missing disoccluded regions are filled with the content transmitted in CUs in the bitstream.

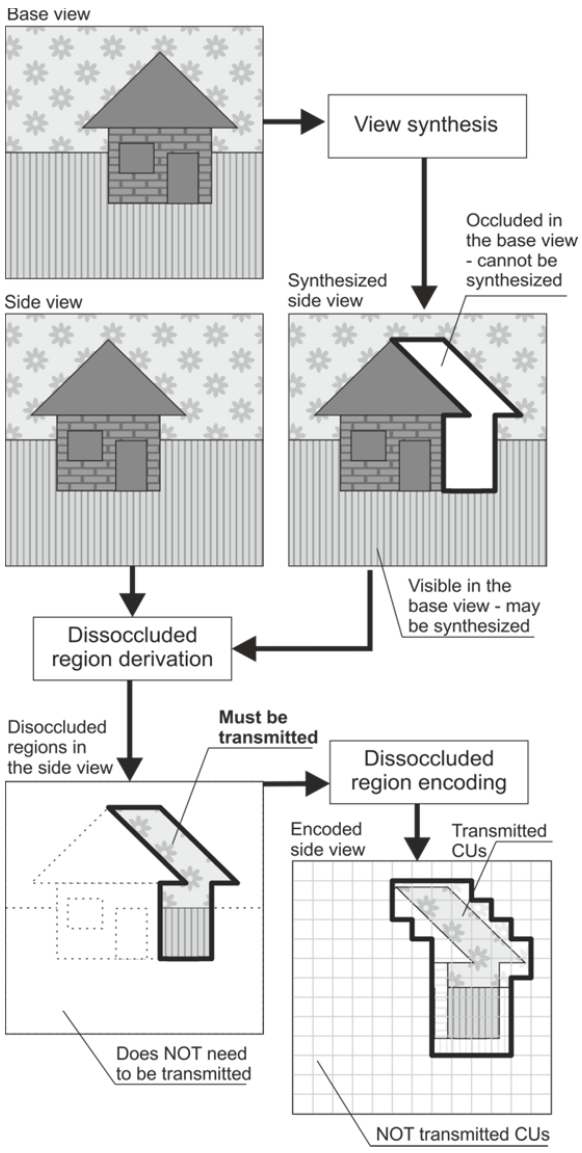


Fig. 3. The principle of Disoccluded Region Coding.

The Coding Units with some disoccluded pixels are encoded using the inter-view disparity-compensated prediction well-known from the MVC extension of the AVC [12]. Note that prediction is only applicable to the disoccluded Coding Units in the side views, i.e. to a relatively small portion of pictures.

Adaptation of the CU size is an important optimization mechanism in HEVC, and it is also exploited by Disoccluded Region Coding. Roughly, the goal is to choose the numbers and sizes of the Coding Units corresponding to disoccluded regions in such a way that all those Coding Units jointly cover the shapes of disoccluded regions (see Figs. 3 and 4). Both the encoder and the decoder use the same view synthesis algorithm with the same reference views. Therefore also the CU adaptation can be performed exactly in the same way both in the encoder and in the decoder. Moreover, the positions and the sizes of the Coding Units that contain pixels from the disoccluded regions can be derived in the decoder in exactly the same way as in the encoder. Therefore, there is no need for additional signaling in the bitstream.

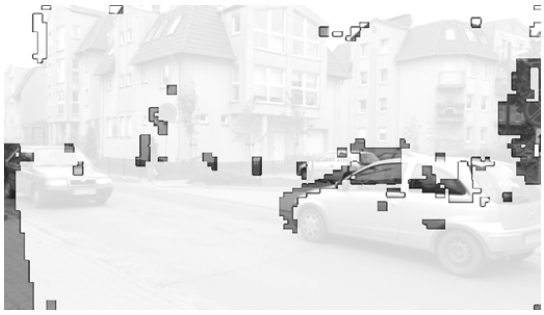


Fig. 4. Disoccluded Coding Units (CUs) transmitted for example *Poznan Street* test sequence.

Application of view-synthesis with Disoccluded Region Coding leads to very efficient representation of video in the side views. That property has been verified experimentally despite of the fact that an adaptive arithmetic encoder is likely to suffer from a small amount of context updating information when the number of the encoded CUs is small. Also, because the number of encoded CUs is highly reduced, and thus representation of the side views and depth maps is so compact, both encoder and decoder can run much faster, resulting in overall reduction of computational complexity.

Note that a reconstructed view, resulting from such encoding, is composed of two distinct areas which contain artifacts of different nature. The first area consists of synthesized regions (Fig. 3) which suffer from view synthesis artifacts. The second area consists of disoccluded regions coded in CUs which suffer from typical compression artifacts. On the boundaries between those two areas, annoying artifacts may occur. In order to remove these artifacts, a technique similar to in-loop deblocking filter is applied inside the loop of the codec.

The above-mentioned in-loop processing is performed as linear interpolation of pixel values within CUs containing disoccluded regions. The linear interpolation provides smooth fading between coded and synthesized regions (Fig. 5).

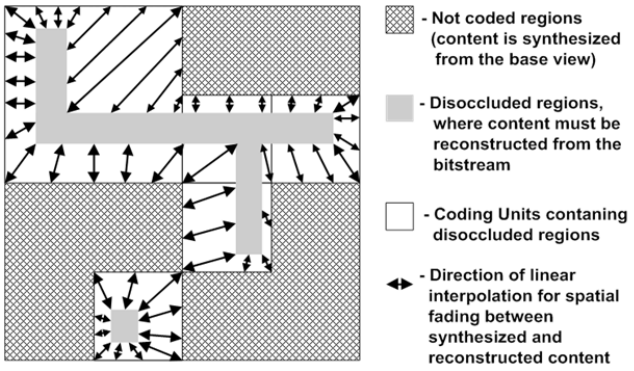
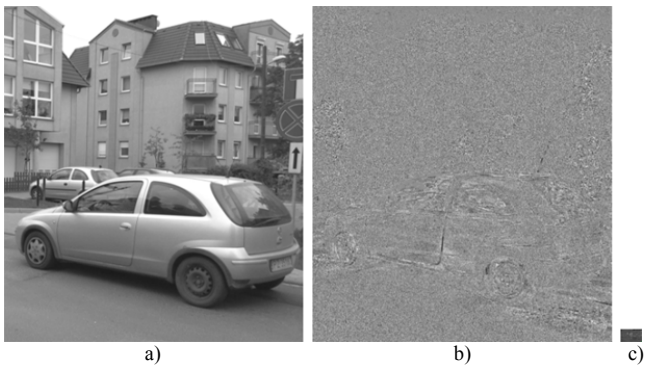


Fig.5. Illustration of in-loop processing for reduction of artifacts caused by Disoccluded Region Coding.

IV. JOINT HIGH-FREQUENCY LAYER REPRESENTATION

In video, temporal high-frequency components are mostly moderate or small (Figs 6a and 6b) but their autocorrelation functions usually exhibit very small values. Therefore these components are not compressed efficiently using standard video coding methods. On the other side, temporal high-frequency components are important for natural appearance of video. Therefore a special tool was proposed to represent efficiently temporal high-frequency components of multiple views.



a) b) c)

Fig. 6. A cropped frame from the test sequence *Poznan Street*: the low-frequency layer (a), the high-frequency layer – for presentation the sample values are amplified 8 times (b) and spatial distribution of energy (c).

The idea is to divide the input image into two spectral bands (layers) in the temporal domain and to encode them separately. The low-frequency layer is obtained by temporal motion-adaptive low-pass FIR filtering [33]. The high-frequency layer is the residual. Frequency division is selected in a way that the high-frequency layer contains only noise.

The low-frequency layer is encoded like video but the high-frequency layer is modeled as a non-stationary random process. There are two components of the model that need to be encoded (Fig. 7): spatial energy distribution (SDE) and spectral envelope. The spatial energy distribution is estimated for each frame. For this purpose, a frame from the high-frequency layer is divided into rectangular non-overlapping blocks. In each of those blocks energy is measured. Energy values, associated with respective blocks, constitute a frame of spatial energy distribution, whose resolution is smaller than resolution of the input video (for example, it fits into one CU for a HD frame, see Fig. 6c).

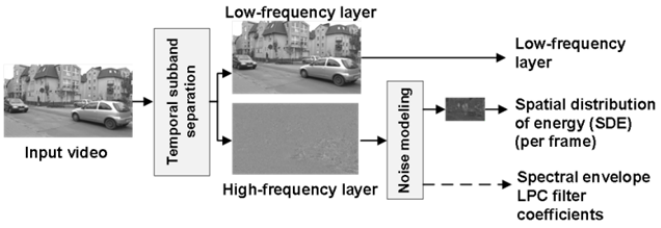


Fig. 7. High-Frequency Layer Representation in an encoder.

This estimated spatial distribution of energy is used in order to normalize the high-frequency layer.

For high-frequency layer, the second component is spectral envelope. It is estimated from energy-normalized high-frequency subband using a technique similar to LPC. The resulting set of separable IIR filter coefficients (in horizontal and vertical direction) is encoded using LAR (log-area-ratio [34]) 8-bit representation. A set of those filter coefficients is estimated for each frame and transmitted to the decoder.

Parameters of the noise model are highly correlated among the views. The frames of the spatial distribution of energy of all views are mapped through view synthesis to a position of the base view, and then averaged. This operation results in only one joint spatial distribution of energy (SDE). Similarly, the energy envelopes of all of the views are averaged, resulting in one joint spectral envelope.

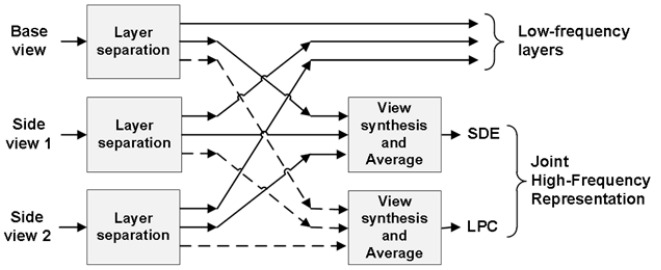


Fig. 8. Layered representation of views: low-frequency layer and Joint High-Frequency Layer Representation (SDE and LPC) – video and parameters are denoted by solid and dashed lines, respectively.

In a decoder, pseudo-random white noise is generated and then modulated by the upsampled spatial energy distribution transmitted in the bitstream and then filtered with IIR filters that reflect the envelope of the original high-frequency layer spectrum. The resultant video, which resembles the original high-frequency subband, is added to the reconstructed low-frequency layer in order to restore the high-frequency components.

V. Depth-Based Motion Prediction (DBMP)

All of the views in a multview video sequence are projections of the same scene. Therefore, motion of the objects in the subsequent viewsis almost the same. The basic idea of this tool is to reuse motion field of the base view in the side views.

The concept of DBMP was previously described in [19] for the MVC codec. Further the tool was extended [35] as a coding tool for the HEVC-based multiview video codec [36,37]. In the proposed 3D video coding technology, some modifications of this DBMP tool are introduced to maximize its performance.

During encoding of the base view sparse block-based motion field is estimated and transmitted as a part of motion compensation prediction in the HEVC encoder. This motion field of a base view (along with the corresponding depth map) is used for synthesis of dense motion field for side views. Such synthetized motion field is then used during the encoding of the side view.

Before encoding and decoding of a frame, for each pixel in the side view, motion information (motion vector and reference picture indices) is synthesized directly from already encoded CUs in the base view at the same time instant (Fig. 8). As a consequence, the motion vectors and reference indices do not need to be transmitted as they can be derived through the DIBR technique at the decoder from the base view. For the CU that uses synthesized motion field, motion compensation is done on pixel basis.

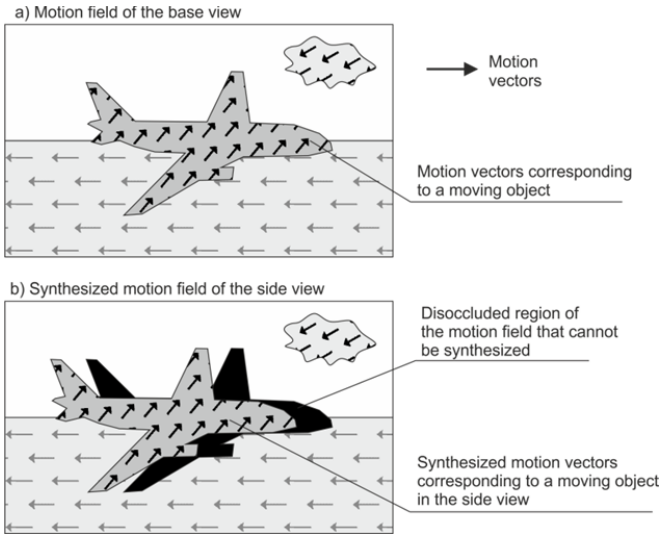


Fig. 9. Independent derivation of motion information for each point of encoded CU from a corresponding point in the reference view.

In HEVC, the most efficient modes for encoding motion information are based on the concept of block merging [35]. In this approach motion vectors and reference picture indices for each currently encoded CU are inferred from motion information of a reference CU. As a result, no motion information or reference picture indices need to be transmitted in the bitstream for the encoded CU. For this reason, DBMP tool has also been implemented as an additional merge candidate. Obviously, the proposed DBMP merge candidate is not added in the base view, which is HEVC-compatible.

VI. Depth Representation

One of the challenges in 3D video coding is efficient representation of the depth data. Many depth views may exist the MVD format, however, the depth maps acquired at the same time instant, but from different views, usually exhibit a significant level of spatial inconsistency. This inconsistency results in different depth values for the corresponding points in the individual views. Also, the depth value at a point may be

considered as noisy, especially after classical coding which introduces uniform quantization error. This is especially disappointing for objects that reside in close distance to the camera. We propose two tools that cope with these problems: Uniform Depth Representation and Nonlinear Depth Representation.

A. *Consistent Depth Representation*

As mentioned before, the inter-view prediction based on view synthesis is widely used for both video and depth. Unfortunately, if the view synthesis algorithm uses inconsistent depth maps, it renders very annoying artifacts in the synthesized video. For that reason, the first step of the 3D video compression algorithm is the depth map inter-view consistency refinement that produces Consistent Depth Representation. The refinement technique consists of four steps: depth map smoothness improvement, depth map synthesis, inter-view depth information exchange and depth value restoration.

The first step of this processing is to improve depth map smoothness by using the Mid-Level Hypothesis algorithm [38]. This algorithm increases sub-pixel precision of the artificially estimated depth maps and provides some level of alignment between the depth and the corresponding texture. This algorithm is described in more detail in [31].

The depth map synthesis is performed similarly to the view synthesis. Each depth map is synthesized from each other depth map independently. For 3-view MVD, this results in two alternative depth maps for each view.

The inter-view depth information exchange consists in median filtering of alternative depth values for each view. This results in depth values that are aligned consistently in the views.

The depth value restoration step ensures that the depth modifications imposed by the algorithm do not affect the overall 3D model of the scene represented by the depth maps.

In general, these four steps could be repeated iteratively but the experimental results [31] indicate that the major improvement is made in the first iteration. In real-life video coding case, it is enough to do one or two iterations.

B. *Nonlinear Depth Representation*

Human perception of depth depends on absolute distance between viewed objects and the observer. A typical observer is more sensitive to depth variations close at hand than to those far away. Also, view synthesis algorithms are much more vulnerable to depth mismatch in the foreground (which results in doubling of borders of the objects) than in the background (typically resulting merely in a small displacement of the whole plane). Therefore, we propose to employ a tool called Nonlinear Depth Representation. This tool produces nonlinear representation of the depth values for compression, which consequently is equivalent to non-uniform quantization of depth.

Depth samples are transformed according to a nonlinear function. An inverse transformation is done after decoding (Fig. 9). This impacts other coding and decoding operations in the tools that rely on depth values. For example, the encoder and the decoder have to be aware about the depth transformation when performing view synthesis.

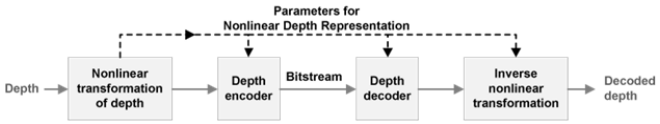


Fig. 10. Nonlinear Depth Representation scheme.

The internal depth representation is non-linear, i.e. closer objects are represented more accurately than distant ones. Internal depth sample values are defined by the following power-law expression, similar as in the case of the well-known video gamma correction:

$$d_{nonlinear} = \left(\frac{d_{linear}}{M_{linear}}\right)^{\gamma} \cdot M_{nonlinear} \quad (1)$$

where: $d_{nonlinear}$ is a nonlinear disparity value used internally inside the codec, normalized to the range from 0 to $M_{nonlinear}$, d_{linear} is a linear disparity value at the input and at the output of the codec, normalized to the range from 0 to M_{linear} and γ is a constant parameter that controls non-linearity of the process.

The exponent γ is automatically chosen by the encoder and sent to decoder in the encoded bitstream. In our experiments it was set in range 1.2÷1.6, depending on the quantization step. The depth range was assumed to be constant, but recently it was proposed to update it on frame basis [39].

VII. INTEGRATION OF THE TOOLS INTO THE CODEC STRUCTURE

For both video and depth, hierarchical view coding structure similar to MVC [13] is used: the already coded views are used as references for the coding of the subsequent views. The proposed 3D video codec includes the new tools described in this paper. These tools exploit the presence of the depth information using the approaches known from computer vision. These tools are integrated with the MVC structure and basic low-level HEVC compression tools like intra-frame prediction, inter-frame motion-compensated prediction, transform coding, in-loop filtering and others. The encoder and decoder structures are depicted in Fig. 11.

In general, the composite 3D-video bitstream consists of 4 types of video sub-streams (see Fig. 11):

- low-frequency video layer of the base view,
- low-frequency video layers of the side views (more than one such sub-stream may exist),
- depth maps for individual views (more than one such sub-stream may exist),
- high-frequency layer representations – may be sent for individual views but, as already mentioned, even all views may share one such component.

For each time instant, video and depth frames from the base view are encoded first, then the depth maps are encoded for the side views. Afterwards, the video for side views and the residual layer are encoded.

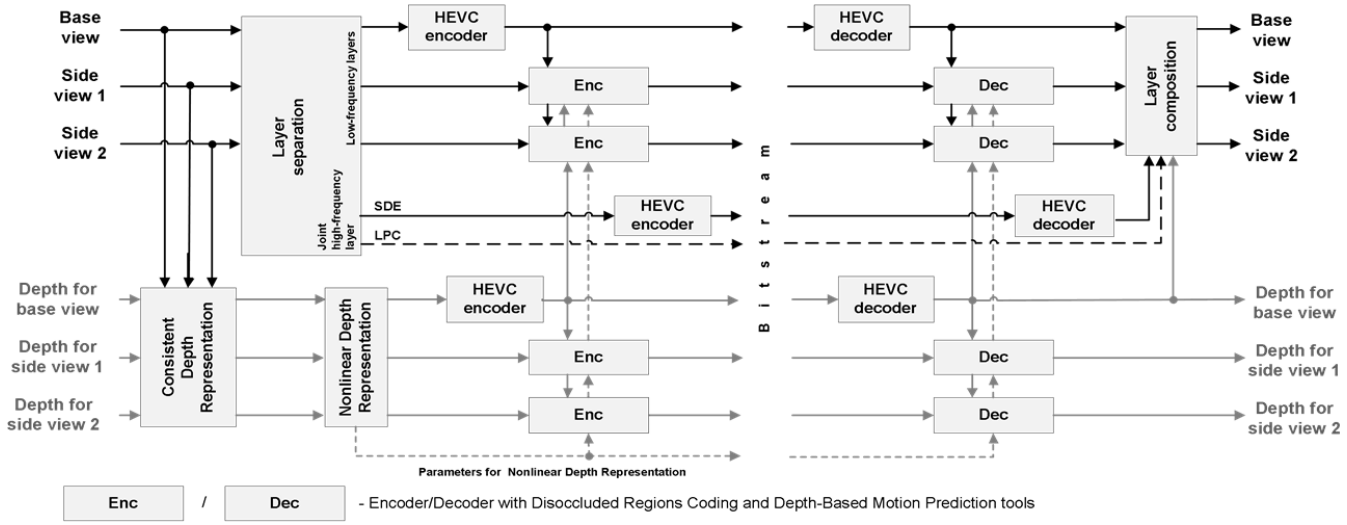


Fig. 11. Proposed encoder (left) and decoder (right) structures.

The encoder produces a bitstream in the form of a sequence of NAL units. As mentioned before, the bitstream of the base view is fully compliant with the HEVC syntax. Other streams, which are not HEVC-compatible, are encapsulated in transparent NAL units, so that they can be skipped by a basic HEVC decoder. A full 3D decoder can use them to decode all of the input views or only some of them.

VIII. Experimental Assessment and Subjective Evaluation

The final versions of individual tools and the codec structure have been obtained by extensive iterative experiments that are beyond the scope of this paper. Here, we report the experimental results that allow to estimate the overall compression performance of the technology as well as to estimate the efficiency of the individual tools. This is done using subjective and objective tests of decoded video quality. The methodology resembles the one used by MPEG for evaluation of the responses to Call for Proposals on 3D Video Coding. For that purpose MPEG has selected 8 test MVD sequences. Four of them were in 1920×1080 resolution (*Poznan Hall 2, Poznan Street, GT Fly* and *Undo Dancer*). Other four test MVD sequences were in the 1024×768 format. For the sake of brevity, in this paper, detailed results will be presented for the 1920×1080 sequences only (Table 1).

Table 1. MVD test sequences used for experiments.

Name	Length	Type	Provider
S01 *Poznan Hall2*	8 secs	Natural - indoor	Poznań University of Technology [40]
S02 *Poznan Street*	10 secs	Natural – outdoor	
S03 *Undo Dancer*	10 secs	Synthetic – indoor	Nokia Corporation [41,42]
S04 *GT Fly*	10 secs	Synthetic – outdoor	

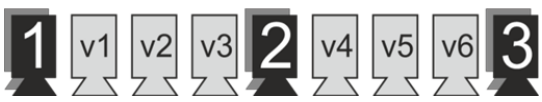


Fig. 12. The arrangement of the views: the v are marked in black while the views synthesized in the receiver are marked in gray.

For our experiments we used 3 views from each sequence (with texture depth maps) and encoded them at four different bitrates. Based on the decoded data we synthesized six virtual views (Fig. 12 – "v1", ... ,"v6") uniformly placed between the original views (Fig 12 – "1", ... ,"3"). Similarly, we have synthesized six virtual views at the same spatial positions from uncompressed data as a reference for average PSNR calculations for luma (Tables 2 and 3, Fig. 11). Average bitrate reductions versus HEVC simulcast were calculated using the Bjontegaard formula [43]. Synthetic references were used instead of the video captured by real cameras because our aim was to assess the quality degradation caused by the coding technology, not caused by the view synthesis algorithm itself.

For the view synthesis algorithm, we used MPEG Synthesis Reference Software [32] with the default configuration.

In all cases, original (not pre-preprocessed) sequences have been used as references for quality measurement – both objective (PSNR) and subjective (MOS).

The subjective tests have been carried out [44] in accordance with the general rules of ITU Recommendation BT.500 [45]. A total number of 62 young persons were viewing each stereopair (composed from virtual views "v3" and "v4", see Fig. 12) on a 46'' Hyundai S465D polarization monitor. The Double Stimulus Method was selected for the subjective quality assessment that followed the rules used by the MPEG for evaluation of the proposals for the 3D video coding technology in 2011 [16].

In our experiments, the number of subjects involved was higher than in the official MPEG evaluation. The high number of subjects yielded that 95% confidence intervals were of order of ±(0.1÷0.25), i.e. very small. Therefore, those intervals were not depicted on the plots (Fig. 13).

Figure 13 shows objective evaluation results (PSNR versus bitrate - BD-rates - in Table 2) and Figures 17-20 show subjective evaluation results (11-point MOS versus bitrate - BD-rates – Table 3) for virtual synthesized views for all four testes sequences.

Note that both subjective and objective quality assessments lead to somewhat similar conclusions. Application of Nonlinear Depth Representation (Tables 2 and 3 – column A) may result in more than 20% bitrate reduction.

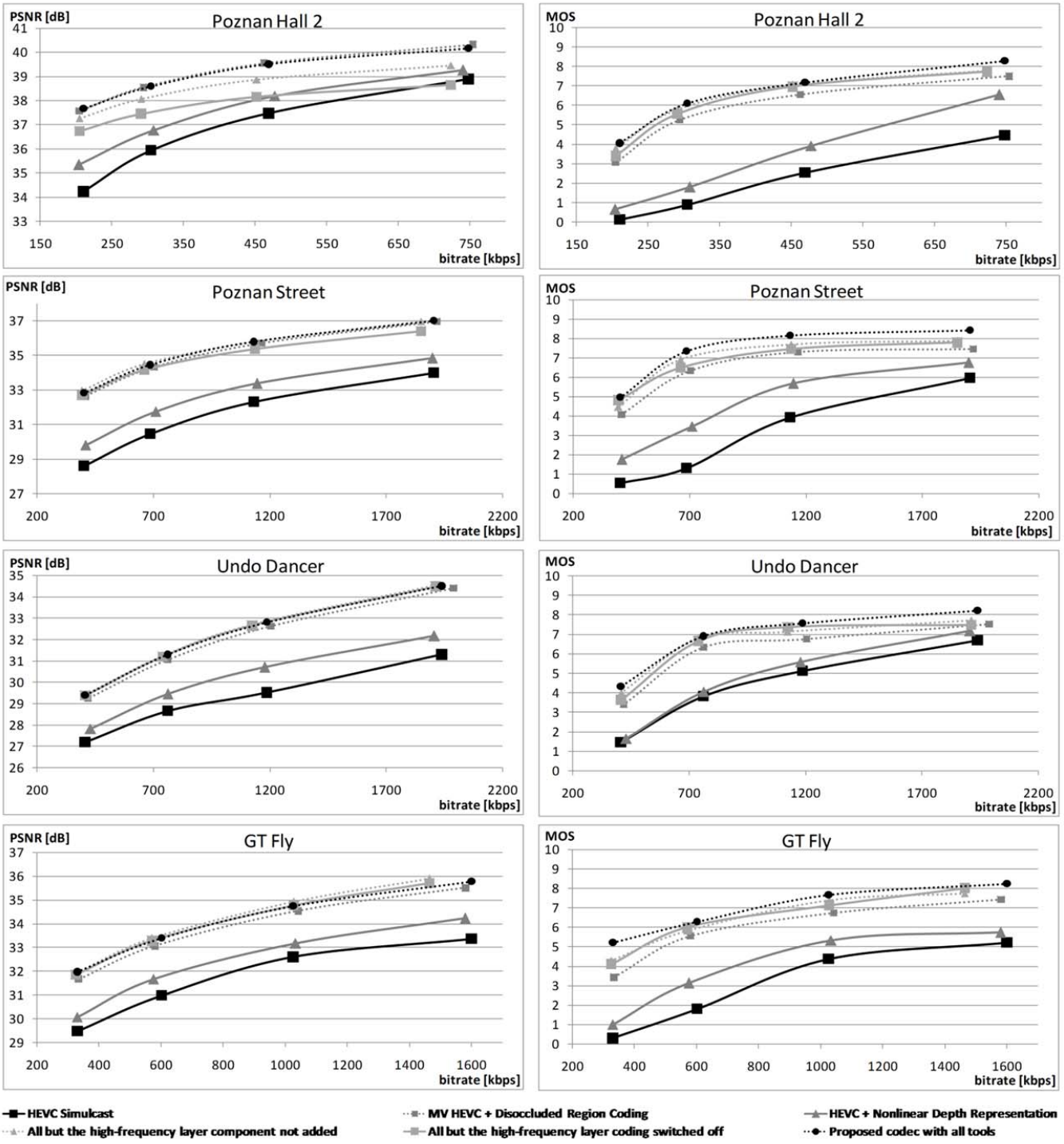


Fig. 13. Objective (left) and subjective (right) evaluation results for S01-S04 sequences.

View-synthesis inter-view prediction combined with MVC-toolset (column C) yields about 50 – 60 % bitrate reduction. Disoccluded Region Coding implemented in a standard HEVC without the MVC toolset provides similar bitrate reductions (column B) of about 45%. The discrepancy between the results obtained by subjective and objective video quality assessment is the most noticeable in the case of *Poznan Hall 2* sequence (S01) which probably results from low quality of associated depth maps.

The application of all devised tools except the Joint High-Frequency Layer Representation results in 50% or more of bitrate reduction (column D). When High-Frequency Layer coding is used, we can consider two cases: column E - when the high frequency layer is modeled but finally not reconstructed, which is more reliable for objective evaluation (PSNR comparison with synthetic noise would be irrelevant) - the gains here are about 50% (objective quality measures) and over 60% (subjective quality assessment).

In another case - column F in Tables 2 and 3 - the synthetic noise is additionally summed with the output video (more important for subjective viewing), which improves the overall gains by about 1 percent. Such gain is not much, but also adequate to the bitrate cost of High-Frequency Layer (Fig. 13).

Therefore, the main gains of the whole proposal come from: Disoccluded Region Coding, Nonlinear Depth Representation and disparity-compensated prediction (MVC-like). It is also worth noting that the remaining tools together contribute a substantial gain of approximately 2÷4% of the overall bitrate as compared to simulcast HEVC (Tables 2 and 3).

The decoded stereoscopic sequences for 2-view and 3-view scenarios and bitstreams (for 8 test sequences), and decoder executables can be found at the website: http://3d-codec.multimedia.edu.pl.

Table 2. Average bitrate reductions calculated as Bjontegaard rates for luma PSNR [dB] versus original (not preprocessed) sequences.

Test sequence	Average (over bitrates and sequences) bitrate reduction versus HEVC simulcast					
	A. HEVC + Nonlinear Depth Representation	**B.** HEVC + Disoccluded Region Coding	**C.** MV HEVC + Disoccluded Region Coding	**D.** All but Joint High-Freq.Layer Repr. switched off	**E.** All but high-freq. layer not added	**F.** Proposed codec with all tools
S01	-19.6	-20.3	-26.1	-14.7	-16.9	-23.7
S02	-27.2	-55.7	-56.8	-58.0	-62.8	-59.8
S03	-29.1	-57.0	-58.0	-60.9	-61.1	-60.7
S04	-23.2	-48.8	-49.4	-54.0	-55.4	-53.7
Avg.	**-24.8**	**-45.4**	**-47.6**	**-49.1**	**-49.1**	**-49.5**

Table 3. Average bitrate reductions calculated as Bjontegaard rates for Mean Opinion Score (MOS) versus original (not preprocessed) sequences.

Test sequence	Average (over bitrates and sequences) bitrate reduction versus HEVC simulcast				
	A. HEVC + Nonlinear Depth Representation	**C.** MV HEVC + Disoccluded Region Coding	**D.** All but Joint High-Freq.Layer Repr. switched off	**E.** All but high-freq. layer not added	**F.** Proposed codec with all tools
S01	-24.5	-65.2	-67.2	-69.4	-70.1
S02	-35.7	-67.5	-72.2	-72.6	-74.8
S03	-8.0	-52.3	-57.4	-61.4	-62.7
S04	-29.6	-62.0	-69.0	-68.8	-67.2
Avg.	**-24.5**	**-61.7**	**-66.4**	**-68.1**	**-68.7**

IX. CONCLUSIONS

In the paper, we have proposed a 3D video coding technology which consists of several new coding tools. The benefits and costs of individual tools have been discussed with the reference to the respective experimental results. The proposed compression technology provides bitrate reduction of the order of 60% as compared to HEVC simulcast. This figure was obtained by systematic subjective tests. It proves high compression performance of the proposed technology that allows very efficient coding of the side views. The bitrate needed for even two side views with the corresponding depth maps is mostly below 50% of the bitrate for single-view video.

REFERENCES

[1] "Applications and Requirements on 3D Video Coding", ISO/IEC JTC1/SC29/WG11, Doc. N11829, Geneva, CH, March 2011.

[2] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. of IEEE Int. Conf. Image Proc.(*ICIP'07), Austin TX, USA, September 2007, pp.I-201-204.

[3] P. Kauff, K. Müller, A. Smolic, et al. "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," Signal Proc.: Image Comm,, vol. 22, no. 2, 2007.

[4] F. Shao, G. Jiang, M. Yu, K. Chen, Y.-S. Ho, "Asymmetric Coding of Multi-View Video Plus Depth Based 3-D Video for View Rendering," IEEE Trans. on Multimedia, vol. 14, 2012, pp. 157- 167.

[5] K. Müller, P. Merkle, T. Wiegand, "3-D video representation using depth maps," *in Proc. of the IEEE,* vol. 99, no.4, 2011, pp. 643-656.

[6] "Overview of 3D video coding," ISO/IEC JTC1/SC29/WG11, Doc.N9784, Archamps, FR, May 2008.

[7] P. Merkle, Y. Morvan, A. Smolic, et al. "The effects of multiview depth video compression on multiview rendering," Signal Processing: Image Communication, vol. 24, no. 1+2, January 2009, pp. 73-88.

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

[9] Y. Lin, and J. Wu, "A depth information based fast mode decision algorithm for color plus depth-map 3D videos, " *IEEE Trans.* on Broadcasting, vol.57, no.2, June 2011, pp.542-550.

[10] D. Tian, P.-L. Lai, P. Lopez, C. Gomila, "View Synthesis Techniques for 3D Video," App. of Digital Image Proc. XXXII. Proc, of the SPIE, Volume 7443 (2009), 2009.

[11] A. Smolic, K. Müller, T. Wiegand, et al. "Intermediate View Interpolation Based on Multiview Video Plus Depth for Advanced 3D Video Systems," IEEE Int. Conf. on Image Processing, ICIP2008,SanDiego, CA, USA, October 2008.

[12] A. Vetro, T. Wiegand, and G.J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC Standard," *Proc. of the IEEE*, vol. 99, no.4,2011, pp. 626-642.

[13] Generic coding of audio-visual objects – Part 10: Advanced Video Coding, ISO/IEC 14496-10, 6nd Ed., 2010, [also:] ITU-T Rec. H.264, Edition 5.0, 2010

[14] "High-Efficiency Video Coding (HEVC) text specification draft 10", JCTVC Document, JCTVC-L1003, Geneva, CH, Jan. 2013.

[15] K. Wegner, O. Stankiewicz, K. Klimaszewski, M. Domański "Comparison of multiview compression performance using MPEG-4 MVC and prospective HVC technology", ISO/IECJTC1/SC29/WG11 MPEG M17913, Geneva, CH, July 2010.

[16] "Call for proposals on 3D video coding technology", ISO/IEC JTC1/SC29/WG11, Doc. N12036, Geneva, CH, March 2011.

[17] J. Seo, H. Wey, S. Lee, K. Sohn, "Motion information sharing mode for depth video coding," *in Proc. of* 3DTV Conference: The True Vision-Capture,Transmission and Display of 3DVideo(3DTV'2010), 2010.

[18] K. Klimaszewski, K. Wegner"Joint Intra Coding of Video and Depth Maps" IEEE Int. Conf. Signals and Electronic Systems – ICSES 2010, Gliwice, PL, Sept. 2010.

[19] J. Konieczny, M. Domański, "Extended inter-view direct mode for Multiview Video Coding", IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 845-848, Prague, CZ, May 2011.

[20] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," *in Proc. of* IEEE Intern. Conf. on Image Processing(ICIP'07), TX, USA, 2007, pp.I-201-204.

[21] T. Suzuki, M. M. Hannuksela, Y. Chen, S. Hattori , "Text of ISO/IEC 14496-10:2012/PDAM 2 MVC extensions for inclusion of depth maps", W12731, MPEG meeting, Geneva, May 2012.

[22] M. M. Hannuksela, Y. Chen, T. Suzuki, J.-R. Ohm, G. J. Sullivan, "3D-AVC draft text 5", ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Doc.JCT3V-C1002, Geneva, CH, Jan. 2013.

[23] M. Domański, K. Klimaszewski, O. Stankiewicz, K. Wegner et al. "Multiview HEVC – experimental results" JCT-VC (MPEG/VCEG) Doc. JCTVCG582,Geneva, November 2011.

[24] G. Tech, K. Wegner, Y. Chen, M. Hannuksela, J. Boyce, " MV-HEVC Draft Text 3", JCT3V Document, JCT3V-C1004, Geneva, CH, Jan. 2013.

[25] D. Rusanovskyy, F. C. Chen, L. Zhang, T. Suzuki, "3D-AVC Test Model 5", ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11,Doc. JCT3V-C1003, Geneva, CH, Jan. 2013.

[26] H. Schwarz, K. Wegner, " Test Model under Consideration for HEVC based 3D video coding", ISO/IEC JTC1/SC29/WG11/ MPEG2011/N12559, San Jose, CA, USA, Feb. 2012.
[27] M. Domański, O. Stankiewicz, K. Wegner, et al. "Technical Desciption of Poznan University of Technology proposal for Call on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11, Doc. M22697, Geneva, CH, November 2011.
[28] M. Domański, O. Stankiewicz, K. Wegner et al. "New Coding Technology for 3D Video within Depth Maps as Proposed for Standardization within MPEG," Int. Conf. Systems, Signals and Image Processing IWSSIP 2012, Vienna, AT, April, 2012, pp. 415-418.
[29] M. Domański, O. Stankiewicz, K. Wegner et al. "Coding of Multiple Video+Depth Using HEVC Technology and Reduced Representations of Side Views and Depth Maps," 29th Picture Coding Symposium, Kraków, PL, May 2012, pp. 5-8.
[30] M. Domański, O. Stankiewicz, K. Wegner et al. "3D video compression by coding of disoccluded regions," 2012 IEEE Int. Conf. Image Proc., ICIP 2012, Orlando, Florida, USA, Sept. – Oct. 2012.
[31] M. Kurc, O. Stankiewicz, M. Domański, "Depth map inter-view consistency refinement for multiview video", Picture Coding Symposium, Kraków, PL, 2012.
[32] ISO/IEC JTC1/SC29/WG11, "View synthesis algorithm in view synthesis reference software 3.0 (VSRS3.0)," Doc. M16090 , Feb.2009.
[33] M. Fizick (A. Balakhnin), Tsp, TSchniede "Mv-tools web-page", http://avisynth.org.ru/mvtools/mvtools.html.
[34] J.A. Martins "Low bit rate LPC vocoders using vector quantization and interpolation" International Conference on Acoustics, Speech, and Signal Processing, 1991, vol. 1 pp. 597- 600.
[35] J. Konieczny, M. Domański, "Depth-based inter-view motion data prediction for HEVC-based multiview video coding", Picture Coding Symposium, PCS 2012, Kraków, PL, May 2012, pp.33-36.
[36] M. Domański et al., "Multiview HEVC – experimental results", ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Doc.JCTVC-G582, Geneva, CH, 2011.
[37] J. Stankowski, M. Domański, O. Stankiewicz, K. Wegner et al."Extensions of the HEVC technology for efficient Multiview Video Coding", IEEE Intern. Conf. on Image Processing, Orlando, USA, 2012.
[38] O. Stankiewicz, M. Domański, K. Wegner, "Stereoscopic Depth Refinement by Mid-Level Hypothisis", IEEE Int. Conf. on Multimedia & Expo, Singapore, SG, July 2010.
[39] I. Lim, H. Wey, D. Park, "3D-CE7.a Improved Nonlinear Depth Representation", ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Doc. JCT3V-C0094, Geneva, CH, Jan. 2013.
[40] M. Domański, O. Stankiewicz, K. Wegner et al. "Poznań multiview video test sequences and camera parameters," ISO/IEC JTC1/SC29/WG11, Doc. M17050, Xian, China, October 2009.
[41] J. Zhang, Ri Li, H. Li, D. Rusanovskyy and M.M. Hannuksela, "Ghost Town Fly 3DV sequence for purposes of 3DV standardization," ISO/IEC JTC1/SC29/WG11, Doc. M20027, Geneva, CH, March 2011.
[42] D. Rusanovskyy, P. Aflaki and M. M. Hannuksela, "Undo Dancer 3DV sequence for purposes of 3DV standardization," ISO/IEC JTC1/SC29/WG11, Doc. M20028, Geneva, CH, March 2011.
[43] G. Bjontegaard, "Calculation of Average PSNR Differences between RD-curves," ITU-T SG16, Doc. VCEG-M33, April 2001.
[44] F. Lewandowski, M. Paluszkiewicz, T, Grajek, K. Wegner, "Subjective Quality Assessment Methodology for 3D Video Compression Technology", Int. Conf, of Signals and Electronic Systems, 2012.
[45] Methodology for the subjective assessment of the quality of television pictures, ITU-R Rec. BT.500-11, 2002.

Marek Domański (M'91) received M.S., Ph.D. and Habilitation degrees from Poznań University of Technology, Poland in 1978, 1983 and 1990 respectively. Since 1993, he is a Professor at Poznań University of Technology, where he leads the Chair (Department) of Multimedia Telecommunications and Microelectronics. Since 2005 he is the head of Polish delegation to MPEG. He authored highly ranked technology proposals submitted in response to MPEG calls for scalable video compression (2004) and 3D video coding (2011). He also leaded the team that developed one of the very first AVC decoders for tv set-top boxes (2004) and various AVC and AAC HE codec implementations and improvements. He is an author or co-author of 3 books and over 200 papers in journals and proceedings of international conferences. The contributions were mostly on image, video and audio compression, image processing, multimedia systems, 3D video and color image technology, digital filters and multidimensional signal processing. He was General Chairman and host of several international conferences: Picture Coding Symposium, PCS 2012; European Signal Processing Conference, EUSIPCO 2007; 73rd Meeting of MPEG; Int. Workshop on Signals, Systems and Image Processing, IWSSIP 1997 and 2004; Int. Conf. Signals and Electronic Systems, ICSES 2004 and others. He served as a member of various steering, program and editorial committees of international journals and international conferences. He served as Area Editor of Signal Processing: Image Communications journal in 2005-2010.

Olgierd Stankiewicz received his M. Sc. degree from the Faculty of Electric Engineering, Poznań University of Technology in 2006. In 2005 he won the second place in IEEE *Computer Society International Design Competition (CSIDC*), held in Washington D.C. Currently, he is an assistant at the Chair of Multimedia Telecommunications and Microelectronics, where he is working toward the Ph.D. degree. He is co-author of several papers on free view television, depth estimation and view synthesis and papers related to hardware implementation in FPGA. His professional interests include signal processing, video compression algorithms, computer graphics and hardware solutions. He is also involved in ISO standardization activities where he contributes to the development of the 3D video coding standards.

Krzysztof Wegner received the M.Sc. degree from Poznań University of Technology in 2008. Currently he is working towards his Ph.D. there. He is co-author of several papers on free view television, depth estimation and view synthesis. His professional interests include video compression in multipoint view systems, depth estimation form stereoscopic images, view synthesis for free view television, face detection and recognition. He is involved in ISO standardization activities where he contributes to the development of the future 3D video coding standards.

Jacek Konieczny received the M.Sc. and Ph.D. degrees from Poznań University of Technology, Poznań, Poland, in 2008 and 2013, respectively. He has been involved in several projects focused on multiview and 3D video coding. His research interests include representation and coding of multiview video scenes, free-viewpoint video, and 2-D and 3-D video-based rendering. He is involved in ISO standardization activities where he contributes to the development of the 3D video coding standard.

Maciej Kurc received his M.Sc. (2008) from the Faculty of Electronics and Telecommunications, Poznań University of Technology, PL, where he currently is a Ph.D. student. His main areas of research are video compression and FPGA logic design.

Jakub Siast received the M.Sc. degree (2009) from the Faculty of Electronics and Telecommunications, Poznań University of Technology, PL, where he is Ph.D. student. His current research interests include image processing and coding, developing of video coding algorithms, FPGA and microprocessor architecture design.

Jakub Stankowski received the M.Sc. degree (2009) from the Faculty of Electronics and Telecommunications, Poznań University of Technology, PL, where he is a Ph.D. student. His current research interests include video compression, performance optimized video processing algorithms, software optimization techniques.

Robert Ratajczak received the M.Sc. degree from the Faculty of Electronics and Telecommunications, Poznań University of Technology, PL, in 2010, where he is currently a Ph.D. student. His current research interests include stereoscopic images processing and coding, 3D surface reconstruction, object classification and detection.

Tomasz Grajek received his M.Sc. and Ph.D. degrees from Poznań University of Technology in 2004 and 2010 respectively. At present he is an assistant at the Chair of Multimedia Telecommunications and Microelectronics. He is author and co-author of several papers on digital video compression, entropy coding and modeling of advanced video encoders. He has been taking part in several projects for industrial research and development projects. His current researches encompass implementing and optimizing video codecs and modeling of advanced video coders.