3D Video Coding Using Advanced Prediction, Depth Modeling, and Encoder Control Methods

Heiko Schwarz*, Christian Bartnik*, Sebastian Bosse*, Heribert Brust*, Tobias Hinz*, Haricharan Lakshman*, Detlev Marpe*, Philipp Merkle*, Karsten Müller*, Hunn Rhee*, Gerhard Tech*, Martin Winken*, and Thomas Wiegand*‡

*Image Processing Department, Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Berlin, Germany
‡Image Communication Chair, Technical University of Berlin, Germany

***Abstract*—The presented approach for 3D video coding uses the multiview video plus depth format, in which a small number of video views as well as associated depth maps are coded. Based on the coded signals, additional views required for displaying the 3D video on an autostereoscopic display can be generated by depth image based rendering techniques. The developed coding scheme represents an extension of HEVC, similar to the MVC extension of H.264/AVC. However, in addition to the well-known disparity-compensated prediction advanced techniques for inter-view and inter-component prediction, the representation of depth blocks, and the encoder control for depth signals have been integrated. In comparison to simulcasting the different signals using HEVC, the proposed approach provides about 40% and 50% bit rate savings for the tested configurations with 2 and 3 views, respectively. Bit rate reductions of about 20% have been obtained in comparison to a straightforward multiview extension of HEVC without the newly developed coding tools.**

I. Introduction

Recent improvements in 3D video technology led to a growing interest in 3D video. The number of cinema screens capable of showing 3D movies as well as the number of movies produced in 3D has been constantly increased in recent years. With the availability of 3D-capable TV sets and Blu-ray players, the introduction of first 3D broadcast channels, and the release of 3D Blu-ray discs it has also been started to bring 3D video into consumers' homes. Autostereoscopic displays, which provide a 3D viewing experience without glasses, are consistently improved and are considered as a promising technology for future 3D home entertainment.

The state-of-the-art standard for multiview video coding is the MVC extension of H.264/AVC [1]. In MVC, one of the views is conventionally coded in conformance to the High profile of H.264/AVC. For coding the other views, the same coding tools are used, but in addition to previously coded pictures of the same view already coded co-located pictures of other views can also be used as reference pictures for inter prediction. In contrast to common stereo displays, autostereoscopic displays require not only two, but a multitude of different views for providing the 3D viewing experience. Since the bit rate required for coding multiview video with MVC increases approximately linearly with the number of coded views, MVC is not appropriate for delivering 3D content for autostereoscopic displays. A promising alternative is the transmission of 3D video in the Multiview Video plus Depth (MVD) format [2]. In the MVD format, typically only a few views are actually coded, but each of them is associated with coded depth data, which represent the basic geometry of the captured video scene. Based on the transmitted video pictures and depth maps, additional views suitable for displaying 3D video content on autostereoscopic displays can be generated using depth image based rendering (DIBR) techniques at the receiver side. Basically, such a coding format could be specified as an extension of MVC. However, the ITU-T Visual Coding Experts Group (VCEG) and the ISO/IEC Moving Pictures Experts Group (MPEG) are developing an improved video coding standard with the name High-Efficiency Video Coding (HEVC). The HEVC test model [3] already provides about 30-50% bit rate savings in comparison to H.264/AVC at the same fidelity, so that it is likely that new video applications will be based on HEVC.

In this paper, we present a 3D video coding scheme that is targeted on providing a 3D video representation suitable for autostereoscopic displays. The 3D video is coded in the MVD format using a newly developed extension of HEVC. Beside extending HEVC to multiple views and an additional coding of depth data and adding the known concept of disparity-compensated prediction, we also developed new coding tools for improving the coding efficiency for dependent views and depth maps. In March 2011, MPEG issued a Call for Proposals on 3D Video Technology [4]. Based on formal subjective tests, the developed codec was chosen as basis for the Test Model under Consideration [5] for MPEG's new standardization project on HEVC-based 3D video coding.

II. Description of the 3D Video Codec

The basic structure of the 3D video codec is shown in the block diagram of Fig. 1. Similar as for MVC, all video pictures and depth maps that represent the video scene at the same time instant build an access unit and the access units of the input MVD signal are coded consecutively. Inside an access unit, the video picture of the so-called independent view is transmitted first directly followed by the associated depth map. Thereafter, the video pictures and depth maps of other views are transmitted. A video picture is always directly followed by the associated depth map. In principle,

978-1-4577-2049-9/12/$26.00

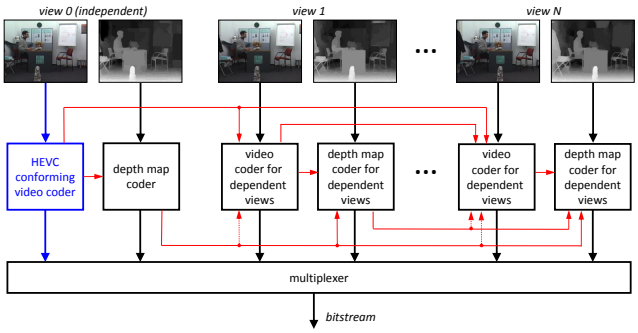

Fig. 1. Block diagram of the 3D video codec.

each component signal is coded using an HEVC-based coder. The corresponding bitstream packets are multiplexed to form the 3D video bitstream. The independent view is coded using a non-modified HEVC coder. The corresponding sub-bitstream can be extracted from the 3D bitstream, decoded with an HEVC decoder, and displayed on a conventional 2D display. The other components are coded using modified HEVC coders, which are extended by including additional coding tools and inter-component prediction techniques that employ already coded data inside the same access unit as indicated by the red arrows in Fig. 1. For enabling an optional discarding of depth data from the bitstream, e.g., for decoding a two-view video suitable for conventional stereo displays, the inter-component prediction can be configured in a way that video pictures can be decoded independently of the depth data.

A. *Coding of Dependent Views*

For coding video pictures of dependent views, the known concept of disparity-compensated prediction (DCP) has been added as alternative to motion-compensated prediction (MCP) in a similar way as for MVC. The macroblock syntax and decoding process haven't been changed for adding DCP, only the high-level syntax has been modified so that already coded video pictures of the same access unit can be inserted into the reference pictures lists. Although DCP generally improves the coding efficiency, it competes with the conventional MCP and the information in already coded views is typically used only for a small part of a picture. Except for regions that are covered or uncovered due to temporal motion, conventional MCP usually provides a more suitable prediction signal. For a more effective usage of already coded views, two additional inter-view prediction methods have been integrated, which are used together with MCP and are described in the following.

Inter-View Motion Parameter Prediction: Since the views of a multiview video sequence represent different projections of the same real world scene which are synchronously captured with multiple cameras, the motion in the different views is very similar. This fact can be employed by predicting motion parameters for a dependent view based on the coded motion parameters in an already transmitted view. In order to establish a relationship between the blocks of a current and a reference view, a depth map for the video picture to be coded is estimated. If the video pictures don't need to be independently decodable, the depth map estimate can be obtained by warping the already coded depth map of another view in the same access unit into the current view. As an alternative, which is also applicable for other configurations, the depth map of an already coded view can be estimated based on previously transmitted disparity vectors and motion parameters. Given the depth map estimate, for each block in the current view a corresponding block in the reference view can be determined and the motion parameters associated with this block are used as candidate motion parameters for the current block. These inter-view motion parameter candidates have been added to the candidate list for the so-called merge mode in HEVC. For conventional inter modes, a motion vector that is determined in the same way, but for a particular reference index, has been added to the list of motion vector predictors. A more detailed description of the concept can be found in [6].

Inter-View Residual Prediction: Not only the motion parameters, but also the reconstructed residual signal of an already coded picture in the same access unit can potentially be used for improving the coding efficiency of dependent views. For that purpose, a flag is added to the syntax of inter-coded blocks, which indicates whether inter-view residual prediction is used. If a block uses resiudal prediction, a disparity vector is determined based on a depth map estimate for the current picture, which is derived as described above. Then, similar as for motion compensation, the block of residual samples in a reference view that is located at the position given by the disparity vector is subtracted from the current residual and only the resulting difference signal is transform coded. If the disparity vector points to a sub-sample location, the residual prediction signal is obtained by interpolating the residual samples of the reference view using a bi-linear filter.

B. *Coding of Depth Maps*

Basically, the same coding tools as for coding the video pictures can be used for depth map coding. However, the HEVC design has been optimized for natural video. In contrast to natural video, depth maps are characterized by sharp edges and large regions with nearly constant values. While nearly constant regions can be well represented using the HEVC transform coding, new intra coding modes have been added for enabling a better representation of depth edges. In another added coding mode, the partitioning and motion data for a video picture are re-used for the depth map coding. Furthermore, some HEVC concepts have been modified. In order to avoid the generation of new depth values and ringing artefacts at depth map edges, the motion compensation doesn't include an interpolation. The motion vectors are coded with sample instead of quarter-sample accuracy. Furthermore, all in-loop filtering techniques are disabled for depth map coding. The two added depth coding tools are briefly described in the following.

Depth Modeling Modes: For the intra coding of depth maps, four additional coding modes have been introduced that partition a depth block into two non-rectangular regions and represent each of these regions by a constant value. Two types

of partitionings are used, namely *Wedgelets* for segmentations using a straight line and *Contours* for arbitrary segmentations. The four depth modeling modes mainly differ in the way in which the partitioning information is signaled:

- *Explicit Wedglet*: The Wedgelet block partition is explicitly signaled inside the bitstream by transmitting an index into list of candidate partitionings.
- *Intra-predicted Wedglet*: The Wedgelet block partition is predicted from already coded neighboring intra blocks and only a correction value is transmitted.
- *Inter-component Wedglet*: The location of the segmentation line is derived based on the reconstructed samples in the co-located block of the associated video picture.
- *Inter-component Contour*: The partitioning into two arbitrary regions is derived based on the reconstructed co-located block in the associated video picture.

For all four modes, the constant values for the two regions are predicted based on the reconstructed samples in neighboring blocks and the remaining difference is quantized and coded in the bitstream. Optionally, a refinement signal can be transmitted using conventional transform coding. For more details on the depth modeling modes, the reader is referred to [7].

Motion Parameter Inheritance: Since video pictures and depth maps represent different properties of the same video scene, the motion characteristics should be similar. In order to exploit this fact, a new inter coding mode for depth maps is added in which the partitioning of a block into sub-blocks as well as the associated motion parameters are inferred from the co-located block in the associated video picture. Since the motion vectors of the video signal have quarter-sample accuracy, whereas for the depth signal sample-accurate motion vectors are used, the inherited motion vectors are quantized to full-sample precision. It can be adaptively decided for each block, whether the partitioning and motion data are inherited from the co-located region of the video picture or new motion data are transmitted. For signaling the Motion Parameter Inheritance (MPI) mode, we modified the merge and skip mode. Therfore, we extended the list of merge candidates, such that in depth map coding, the first merge candidate refers to merging with the co-located block in the video picture. Addition information about the MPI mode can be found in [8].

C. Encoder Control

In modern video encoders, the decision between different coding modes is based on a Lagrangian cost measure $D+\lambda \cdot R$ that weights the distortion D obtained by coding a block in a particular mode with the number of bits R required for signaling the mode using a Lagrangian multiplier λ. The distortion is typically measured as the sum of squared differences (SSD) or the sum of absolute differences (SAD) between the original and reconstructed signal. However, coding artifacts in depth data are only indirectly perceivable in the synthesized video data. The decoded depth map itself is not visible. By considering this fact and modifying the used distortion measure for depth coding, the coding efficiency can be improved as is described in the following.

View Synthesis Optimization: In the modified encoder control for depth maps, the distortion is not directly measured in the depth map domain, but instead the resulting distortion in one or more synthesized views is analyzed. As reference views, intermediate views are synthesized using the original video pictures and depth maps. In principle, for testing a coding mode for a depth block, two variants for the reference views are synthesized using coded data. For the first variant, a depth map consisting of reconstructed depth values for already coded blocks and original depth values for the remaining blocks is used together with already coded or original video pictures for the view synthesis. The second variant is obtained in the same way, but for the current block reconstructed depth data obtained with the mode to be tested are used instead of the original depth data. Then, for both of the synthesized views, the SSD between the synthesized views and the corresponding reference views is calculated. The difference between these error measures is used as distortion measure for mode decision. In order to enable an effective calculation of the distortion measure without re-rendering the complete synthesized views for each distortion calculation, a fast rendering mechanism has been developed by which only the parts that are actually affected by a modification of the considered depth block are re-rendered. The rendering method used in the encoder supports all basic processing steps of common view synthesis algorithms, including sub-sample accurate warping, hole filling, and view blending. A more detailed description of the approach is given in [9].

As an optional encoding technique, we developed a mechanism by which regions in dependent views that can be rendered based on the transmitted independent view and depth maps are identified. These regions are encoded using a lower fidelity. At the decoder side, these regions can be identified in the same way and replaced by rendered versions. Due to the difficulties in evaluating the effectiveness of this mode by objective measures, this technique is not used in the experiments presented in the next section. For more information about this encoding technique, the reader is referred to [10].

III. Experimental Results

The effectiveness of the developed 3D video codec is evaluated by comparing it to three reference coding schemes. In the first reference coding scheme (HEVC simulcast), which was also used as anchor by MPEG for evaluating the submitted proposals, all views and depth maps are independently coded using an unmodified version HEVC. As a second reference, a straightforward multiview extension of HEVC was used. Beside a high-level signaling mechanism, this multiview extension of HEVC only uses the well-known concept of DCP, but doesn't include the newly developed coding tools. And as third reference, both the videos and depth maps are coded using the MVC extension of H.264/AVC.

The coding experiments have been conducted using the test sequences and test conditions specified in MPEG's CfP [4]. As one coding constraint restricts the interval between two successive random access points to 0.5 seconds, we used

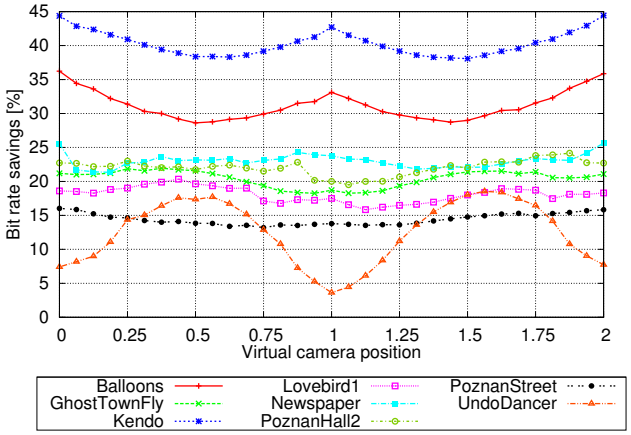

Fig. 2. Bit rate savings relative to a straightforward multiview extension of HEVC at different virtual camera positions for the 3-view test scenario.

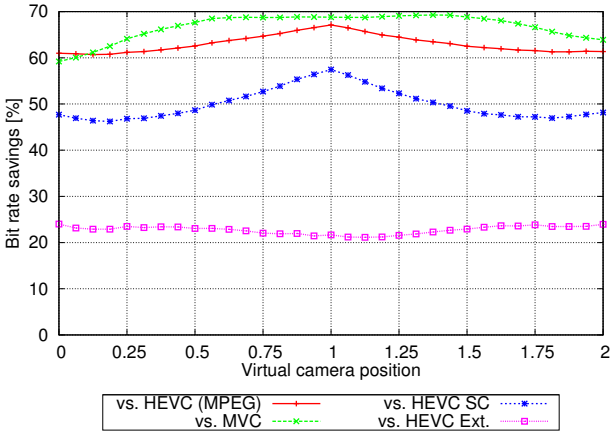

Fig. 3. Average bit rate savings relative to a straightforward multiview extension of HEVC, HEVC simulcast, and MVC for the 3-view test scenario.

hierarchical GOP structures with 12 pictures for the 25fps sequences and 15 pictures for the 30fps sequences. For the 2-view scenario, the right view is coded as the independent view, and for the 3-view scenario, the center view is coded as the independent view. As specified in the CfP, each test sequence has been coded at four different bit rates. For evaluating the coding efficiency, intermediate views have been generated at every 1/16-th position between the coded views using the decoded video pictures and depth maps. Thus, 15 and 30 intermediate views have been generated for the 2-view and 3-view scenario, respectively. As synthesis algorithm, we used a renderer which provides comparable and sometimes better visual results than the rendering algorithm provided by MPEG. The generated intermediate views are compared with intermediate views that are rendered using the original video and depth data. Then, given the determined PSNR values for the intermediate and actually coded views and the overall rates, we calculated bit rate savings for the different virtual camera positions using the Bjøntegaard delta rate [11].

In Fig. 2, the bit rate savings relative to the straightforward extension of HEVC are plotted as a function of the virtual camera position for all 8 test sequences of the 3-view scenario. Since the two compared codecs differ only in the newly developed coding tools, the plotted bit rate savings represent the coding efficiency gains that are obtained by these new tools. Fig. 3 shows the bit rate savings averaged over the test sequences in comparison to the three reference codecs, HEVC simulcast, the straightforward multiview extension of HEVC, and MVC simulcast. In addition, it also shows the bit rate savings relative to the MPEG anchors, which also use HEVC simulcast, but with a GOP of 8 pictures, since the HEVC reference software supported only dyadic GOP structures. The bit rate savings averaged over all virtual camera positions are summarized in Table I for all three reference codecs, the straightforward HEVC extension (HEVC Ext.), HEVC simulcast (HEVC SC), and MVC simulcast (MVC SC). The proposed coding scheme was identified as one of the best performing proposals in MPEG's subjective test and chosen as the basis for the Test Model under Consideration [5].

TABLE I
AVERAGE BIT RATE SAVINGS.

	HEVC Ext.		HEVC SC		MVC SC	
Sequence	2 view	3 view	2 view	3 view	2 view	3 view
PoznanHall2	20.10	22.04	35.10	44.96	70.92	76.42
PoznanStreet	11.97	14.41	37.60	51.07	57.19	60.67
UndoDancer	6.56	12.50	36.74	53.19	58.50	65.79
GhostTownFly	16.17	20.42	44.23	57.61	66.31	71.25
Kendo	37.15	40.31	42.74	53.27	62.91	67.41
Balloons	27.81	31.17	37.85	49.53	63.46	71.04
Lovebird1	16.22	18.09	34.93	47.10	53.66	59.27
Newspaper	20.28	23.02	35.20	43.16	51.86	62.53
average	19.53	22.75	38.05	49.99	60.60	66.80

IV. CONCLUSION

We presented an extension of HEVC for coding 3D video in a format suitable for autostereoscopic displays. Beside the well-known concept of disparity-compensated prediction, some new tools for the coding of dependent views and depth maps have been integrated. The experimental results indicate that the new tools provide bit rate savings of about 20%, resulting in overall bit rate savings against simulcast of about 40% and 50% for the tested 2-view and 3-view scenarios.

REFERENCES

[1] ITU-T and ISO/IEC, "Advanced video coding for generic audiovisual services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC), 2010.
[2] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," PCS, 2007.
[3] JCT-VC, "WD3: Working Draft 3 of High-Efficiency Video Coding," Doc. JCTVC-E603, 2011.
[4] ISO/IEC MPEG, "Call for Proposals on 3D Video Coding Technology," MPEG N12036, 2011.
[5] ——, "Test Model under Consideration for HEVC based 3D video coding," MPEG N12350, 2011.
[6] H. Schwarz and T. Wiegand, "Inter-View Prediction of Motion Data in Multiview Video Coding," PCS, 2012.
[7] P. Merkle, C. Bartnik, K. Müller, D. Marpe, and T. Wiegand, "3D Video: Depth Coding Based on Inter-component Prediction of Block Partitions," PCS, 2012.
[8] M. Winken, H. Schwarz, and T. Wiegand, "Motion Vector Inheritance for High Efficiency 3D Video plus Depth Coding," PCS, 2012.
[9] G. Tech, H. Schwarz, K. Müller, and T. Wiegand, "3D Video Coding using the Synthesized View Distortion Change," PCS, 2012.
[10] S. Bosse, H. Schwarz, and T. Wiegand, "Encoder Control for Renderable Regions in High Efficiency 3D Video plus Depth Coding," PCS, 2012.
[11] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, 2001.