3D High-Efficiency Video Coding for Multi-View Video and Depth Data

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

***Abstract*— This paper describes an extension of the high efficiency video coding (HEVC) standard for coding of multi-view video and depth data. In addition to the known concept of disparity-compensated prediction, inter-view motion parameter, and inter-view residual prediction for coding of the dependent video views are developed and integrated. Furthermore, for depth coding, new intra coding modes, a modified motion compensation and motion vector coding as well as the concept of motion parameter inheritance are part of the HEVC extension. A novel encoder control uses view synthesis optimization, which guarantees that high quality intermediate views can be generated based on the decoded data. The bitstream format supports the extraction of partial bitstreams, so that conventional 2D video, stereo video, and the full multi-view video plus depth format can be decoded from a single bitstream. Objective and subjective results are presented, demonstrating that the proposed approach provides 50% bit rate savings in comparison with HEVC simulcast and 20% in comparison with a straightforward multi-view extension of HEVC without the newly developed coding tools.**

***Index Terms*— 3D video coding (3DVC), multi-view video plus depth (MVD), high-efficiency video coding (HEVC), MPEG-H, H.265.**

I. Introduction

3D VIDEO provides a visual experience with depth perception through the usage of special displays that re-project a three-dimensional scene from slightly different directions for the left and right eye. Such displays include stereoscopic displays, which typically show the two views that were originally recorded by a stereoscopic camera system. Here, glasses-based systems are required for multi-user audiences. Especially for 3D home entertainment, newer stereoscopic displays can vary the baseline between the views to adapt to different viewing distances. In addition, multi-view displays are available, which show not only a stereo pair, but a multitude of views (typically 20 to more than 50 views) from slightly different directions. Each user still perceives a viewing pair for the left and right eye. However, a different stereo pair is seen when the viewing position is varied by a small amount. This does not only improve the 3D viewing experience, but allows the perception of 3D video without glasses, also for multi-user audiences [2], [25]. As 3D video content is mainly produced as stereo video content, appropriate technology is required for generating the additional views from the stereo data for this type of 3D displays. For this purpose, different 3D video formats or representations have been considered.

First, video-only formats like conventional stereo video (CSV) and multi-view video (MVV) were proposed. Backward-compatible compression methods for efficient transmission of these video-only formats were investigated. As a result, multi-view video coding (MVC) was standardized as an extension of H.264/MPEG-4 Advanced Video Coding (AVC) [6], [20], [47], [52], [53]. MVC adds the concept of disparity-compensated prediction to the H.264/MPEG-4 AVC base standard. Due to the corresponding exploitation of inter-view dependencies, MVC provides a higher compression efficiency than a separate coding of each view. However, the bit rate that is required for a given level of video quality still increases approximately linearly with the number of coded views [33]. Thus, an MVC-based transmission of a multitude of video views suitable for multi-view displays is not feasible.

As a next step, 3D video formats with few views and associated depth information were investigated. The depth information can be provided through different methods, including direct recording by special time-of-flight cameras [28], extraction from computer animated video material from the inherent 3D geometry representation [19], or disparity estimation [1], [42]. Such depth-enhanced formats are suitable for generic 3D video solutions, where only one format is coded and transmitted while all necessary views for any 3D display are generated from the decoded data, e.g., by means of depth image based rendering (DIBR) [21], [41].

Parallel developments on improving 2D video coding have led to the high-efficiency video coding (HEVC) standard, officially approved in April 2013 as ITU-T Recommendation

Manuscript received January 1, 2013, revised April 12, 2013, accepted April 26, 2013. Date of publication May 23, 2013; date of current version July 9, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Debargha Mukherjee.

K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, H. Rhee, G. Tech, and M. Winkenare are with the Image Processing Department, Fraunhofer Institute for Telecommunications-Heinrich-Hertz-Institut, Berlin 10587, Germany (e-mail: karsten.mueller@hhi.fraunhofer.de; heiko.schwarz@hhi.fraunhofer.de; detlev.marpe@hhi.fraunhofer.de; christian.bartnik@hhi.fraunhofer.de; sebastian.bosse@hhi.fraunhofer.de; heribert.brust@hhi.fraunhofer.de; tobias.hinz@hhi.fraunhofer.de; haricharan.lakshman@hhi.fraunhofer.de; philipp.merkle@hhi.fraunhofer.de; hunn.rhee@hhi.fraunhofer.de; gerhard.tech@hhi.fraunhofer.de; martin.winken@hhi.fraunhofer.de).

T. Wiegand is with the Image Processing Department, Fraunhofer Institute for Telecommunications-Heinrich-Hertz-Institut, Berlin 10587, Germany, and also with the Department of Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin 10587, Germany (e-mail: thomas.wiegand@hhi.fraunhofer.de).

Digital Object Identifier 10.1109/TIP.2013.2264820

H.265 and ISO/IEC 23008-2 (MPEG-H Part 2), jointly developed by the ITU-T Visual Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) [18]. Experimental analysis [40] showed that HEVC is capable of achieving the same subjective video quality as the H.264/MPEG-4 AVC High Profile while requiring on average only about 50% of the bit rate.

Based on HEVC, this paper proposes a new 3D video coding framework for depth-enhanced multi-view formats. The technology was submitted as a response to the Call for Proposals (CfP) on 3D Video Technology [15], issued by MPEG in order to develop a new standard for high efficiency 3D video delivery. After extensive subjective testing, our proposed technology was selected as the starting point of the currently on-going collaborative phase in standardization of the 3D video and multi-view extensions of HEVC. The proposed framework is format scalable in the sense that sub-bitstreams representing a subset of the video views (with or without associated depth data) can be extracted by discarding NAL units from the 3D bitstream and be independently decoded. Furthermore, our proposed methods for depth coding can also be used in other hybrid coding architectures, where legacy codecs like H.264/MPEG-4 AVC or MVC are used for the coding of the video components.

The paper is organized as follows. Section II gives an overview of the 3D video coding structure and coding tools. The new methods on advanced prediction methods for coding dependent video views are discussed in Section III. Section IV explains the depth coding approaches, including new intra coding modes, motion vector coding, and motion parameter inheritance. For the optimal encoder control of the proposed 3D video codec, view synthesis optimization methods and an encoder-side rendering module are described in Section V. Section VI discusses the decoder-side view synthesis at the 3D display. Objective and subjective results are presented in Section VII and conclusions are drawn in Section VIII.

II. Overview of 3D Video Coding Extension

The presented 3D video coding (3DVC) extension of HEVC was developed for depth-enhanced 3D video formats, ranging from conventional stereo video (CSV) to multi-view video plus depth (MVD) with two or more views and associated per-pixel depth data components. The format scalability is achieved by coding each video view and associated depth map component using a 2D video coding structure that is based on the technology of HEVC [18]. The basic structure of our 3D video encoder is shown in Fig. 1.

In order to provide backward compatibility with 2D video services, the base or independent view is coded using a fully HEVC compliant codec. This includes spatial prediction within a picture, temporal motion-compensated prediction between pictures at different time instances, transform coding of the prediction residual, and entropy coding. Individual coding tools of HEVC and their rate-distortion performance are discussed in [40] and [49]. For an overview of HEVC, the reader is referred to [48].

For coding the dependent views and the depth data, modified HEVC codecs are used, which are extended by including additional coding tools and inter-component prediction techniques that employ data from already coded components at the same time instance, as indicated by the grey arrows in Fig. 1. In order to also support the decoding of video-only data, e.g., pure stereo 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. In summary, the HEVC design is extended by the following tools:

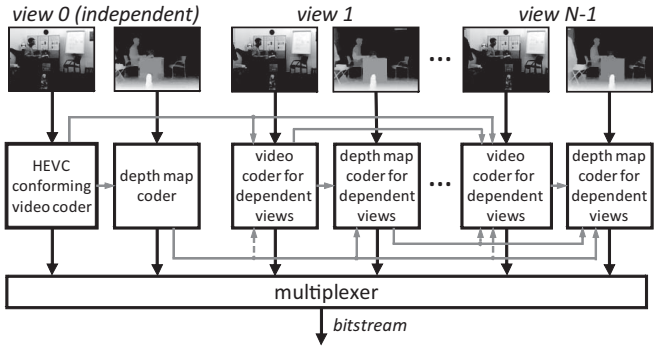

Fig. 1. Basic encoder structure with inter-view and inter-component prediction (gray arrows).

1) Coding of dependent views using disparity-compensated prediction, inter-view motion prediction and inter-view residual prediction.
2) Depth map coding using new intra coding modes, modified motion compensation and motion vector coding, and motion parameter inheritance.
3) Encoder control for depth-enhanced formats using view synthesis optimization with block-wise synthesized view distortion change and encoder-side render model.
4) Decoder-side view synthesis based on DIBR for generating the required number of display views.

Further details on these tools are provided in the following sections.

III. Coding of Dependent Views

One of the most important aspects for efficient multi-view plus depth coding is the redundancy reduction among different views at the same time instance, for which the content is usually rather similar and only varies by a slightly different viewing position. For coding dependent views, the same concepts and coding tools are used as for the independent view. However, additional tools have been integrated into the HEVC design, which employ already coded data in other views for efficiently representing a dependent view. This includes disparity-compensated prediction, inter-view motion parameter prediction as well as inter-view residual prediction. These additional tools are described in the following subsections. While disparity-compensated prediction is also supported for the coding of the depth maps of dependent views, the inter-view motion parameter and residual prediction is only used for dependent video views.

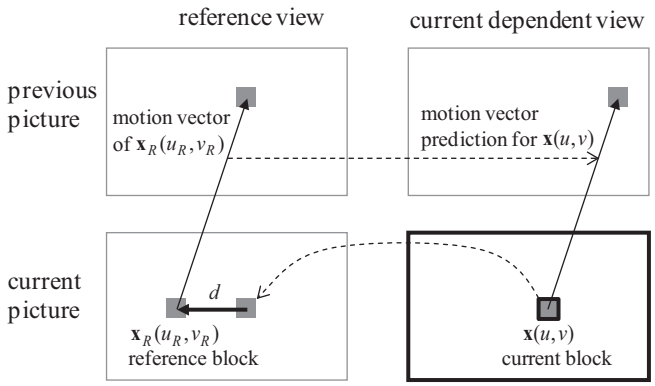

Fig. 2. Motion vector correspondences between a block in a current picture of a dependent view and an already coded reference view, using the disparity vector d from a depth map estimate.

A. Disparity-Compensated Prediction

As a first coding tool for dependent views, the concept of disparity-compensated prediction (DCP) has been added as an alternative to motion-compensated prediction (MCP). Here, MCP refers to inter-picture prediction that uses already coded pictures of the *same* view at *different* time instances, while DCP refers to inter-picture prediction that uses already coded pictures of *other* views at the *same* time instance. DCP is also used in the MVC extension of H.264/MPEG-4 AVC and similarly, the coding tree block syntax and decoding process of HEVC remain unchanged when adding DCP to our codec. 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 picture lists.

B. Inter-View Motion Parameter Prediction

The different views of an MVD sequence are typically rather similar, as they were captured from the same 3D scene at slightly different positions. Accordingly, the motion parameters of the same scene content in different views should be similar, too. Consequently, methods for depth-based inter-view prediction of motion parameters such as motion vectors have been studied [13], [24], [26], [43]. For our proposed 3DVC extension of HEVC, we added a method [45] for predicting motion parameters in a dependent view from an already coded and transmitted view at the same time instance. This concept is illustrated in Fig. 2 for a block in a currently coded dependent view.

For deriving the candidate motion parameters for a block $\mathbf{x}(u, v)$ in the current dependent view, with (u, v) being the location of the center of the block, an associated disparity vector d is used to obtain the corresponding reference sample location $(u_R, v_R) = (u + d, v)$. Based on this sample location, a reference block $\mathbf{x}_R(u_R, v_R)$ representing the prediction unit that covers (u_R, v_R) in the reference view is determined.

Note that d has only a horizontal component as 3D video sequences are typically rectified [12]. The disparity information d is calculated from a depth map estimate. This estimate can be created in two ways, depending on the selected encoder configuration with respect to the required decoder functionality:

1) Depth maps are decoded together with the video data and the depth map estimate is created by warping an already decoded depth map of another view at the same time instance into the current view.
2) The depth map estimate is created from previously transmitted disparity and motion parameters.

Note that method 1) is only applicable, if the application does not require supporting video only decoding, while method 2) can be used in all configurations.

If the reference block $\mathbf{x}_R$ is coded using MCP, the associated motion parameters (number of motion hypotheses, reference indices, and motion vectors) can be used as candidate motion parameters for the current block $\mathbf{x}$ in the current view. These inter-view motion parameter candidates have been added to the candidate list for the so-called merge mode in HEVC [14], [48]. In this mode, no motion parameters are coded. Instead, a candidate list of motion parameters is derived, which includes the motion parameters of spatially neighboring blocks as well as motion parameters that are derived based on the motion data of a temporally co-located block. The chosen set of motion parameters is signaled by transmitting an index into the candidate list. 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 predictor candidates. Finally, the derived disparity vector can also be directly used as a candidate disparity vector for DCP.

C. Inter-View Residual Prediction

In addition to similar motion parameters in the different views, also similar residual signals can be expected. In particular, a reconstructed residual signal of an already coded view can be used for further improvement of the coding efficiency in a currently coded dependent view. For this, we integrated a block-adaptive inter-view residual prediction. Similarly to the inter-view motion prediction, the inter-view residual prediction uses the same depth map estimate for the current picture, as described in the previous section III-B. Based on the depth map estimate, a disparity vector is again determined for a current block $\mathbf{x}$ to a reference sample location (u_R, v_R) for the top-left sample (u, v) of the current block.

This time, the reconstructed residual signal of the block that contains the reference sample location as the top-left sample is used for predicting the residual of the current block. 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. Finally, only the difference between the current and reference residual signal is transmitted using transform coding. At the encoder side, the inter-view residual prediction can be compared to a conventional HEVC residual transform coding for each block and selected, if a better rate-distortion performance is achieved.

IV. Depth Map Coding

For the coding of depth maps, the same concepts of intra-prediction, motion-compensated prediction, disparity-

compensated prediction, and transform coding as for the coding of the video pictures are used. However, in contrast to natural video, depth maps are characterized by sharp edges and large regions with nearly constant values [35]. Therefore, different depth coding methods have been studied in the context of H.264/MPEG-4 AVC-based 3D video coding, including wavelet coding [8], [30], mesh-based depth coding [22], subsampling of depth data [39] as well as non-rectangular block partitioning for depth maps, such as wedgelet or platelet coding [32], [34] and edge chain coding [7].

As a consequence, new intra coding techniques using so-called depth modeling modes have been developed. In addition, motion-compensated prediction and motion vector coding have been modified for the coding of depth maps. Furthermore, a depth-coding mode has been developed that directly uses the block partitioning and motion data from the associated video component. Also, all in-loop filtering techniques of HEVC are disabled for depth map coding. While for encoding of the video component the reconstruction quality is directly measured from its decoded version, the reconstruction quality of depth map coding has to be measured indirectly as the quality of synthesized views from decoded video and depth data. This has been considered for all depth-coding decisions and modes, described in this section. In particular, a special encoder process with view synthesis optimization is applied, as described in Section V.

A. Intra Coding Using Depth Modeling Modes

In contrast to video or texture data, depth maps are mainly characterized by sharp edges at boundaries between objects with different scene depth and large areas of nearly constant or slowly varying sample values within objects. For the nearly constant areas, the coding tools of HEVC, namely intra prediction and transform coding, are well-suited. In contrast, these tools may lead to significant coding artifacts at sharp edges, causing strongly visible artifacts in synthesized views along object boundaries. For a better representation of such edges, we added four new intra prediction modes for depth coding. In all four modes, a depth block is approximated by a model that partitions the area of the block into two non-rectangular regions $\mathbf{P}_1$ and $\mathbf{P}_2$ as shown in Fig. 3, where each region is represented by a constant value.

The information required for such a model consists of two elements, namely the partition information, specifying the region each sample belongs to, and the region value information, specifying a constant value for the samples of the corresponding region. Such a region value is referred to as constant partition value (CPV) in the following. Two different partition types are used, namely wedgelets [32] and contours, which differ in the way the segmentation of the depth block is derived, as shown in the top and bottom row of Fig. 3, respectively.

The new depth modeling modes are integrated as an alternative to the conventional intra prediction modes specified in HEVC. For these modes, a residual representing the difference between the approximation and the original depth signal can be transmitted via transform coding. The approximation of depth

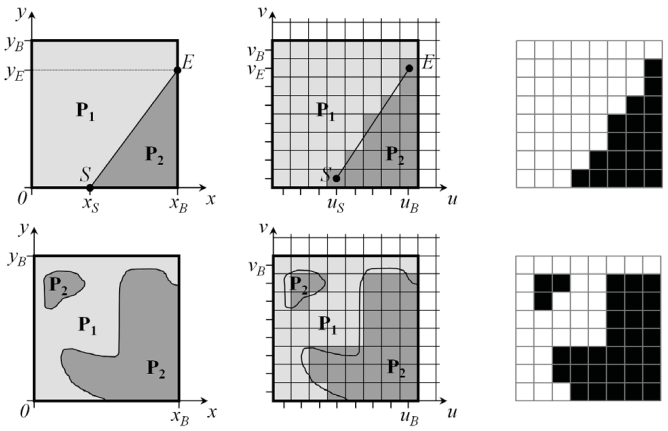

Fig. 3. Wedgelet partition (top) and contour partition (bottom) of a depth block: continuous (left) and discrete signal space (middle) with corresponding partition pattern (right).

blocks using the four new depth modeling modes to obtain specific partitioning patterns is described in more detail in subsections 1)–4) of this section.

After obtaining the optimal partitioning pattern of a depth block, coding of both CPVs of regions $\mathbf{P}_1$ and $\mathbf{P}_2$ (conf. Fig. 3) is carried out. In order to reduce the bit rate, CPVs are not transmitted directly. Instead, they are predicted from information that is also available at the decoder, namely from adjacent samples of neighboring left and top blocks [31]. First, predicted CPVs are calculated as the mean value of these corresponding sample values. Then, difference or delta CPVs are calculated between original and predicted CPVs. Finally, the delta CPVs are linearly quantized at the encoder and dequantized before reconstruction at the decoder. This method is also used in transform coding and the step size of the quantization is similarly set as a function of the quantization parameter (QP).

In the encoding process, either one of the described depth modeling modes, or one of the conventional intra/inter prediction modes is selected. If a depth modeling mode is selected, the selected mode and the associated prediction data are signaled in the bitstream in addition to a syntax element that specifies the usage of a depth modeling mode.

Note that a depth block can also be represented by piecewise linear functions for each region. For this, depth coding modes were investigated, but didn't provide much benefit in coding efficiency within the overall 3D-HEVC framework. Reasons are the quadtree-based prediction block partitioning, the intra-planar prediction mode, where the current block is interpolated from neighboring spatial blocks, and the efficient residual coding of HEVC. These tools combine well with modeling the depth signal around sharp edges by constant functions also for piecewise smooth signals.

1) Explicit Wedgelet Signaling: In this depth modeling mode, a best-matching wedgelet partition is sought at the encoder and the partition information is transmitted in the bitstream. At the encoder, a search within a defined set of wedgelet partitions is carried out using the original depth signal of the current block as a reference. During this search,

the wedgelet partition that yields the minimum distortion between the original signal and the wedgelet approximation is selected as the final partitioning pattern for the depth block. For this, the patterns of all possible combinations of start and end point positions are generated and stored in a lookup table for each block size prior to the coding process. This wedgelet pattern list contains only unique patterns. The resolution for the start and end positions (S and E in Fig. 3 top), used for generating the wedgelet patterns, depends on the block size. For 16 × 16 and 32 × 32 blocks, the possible start and end positions are restricted to locations with 2-sample accuracy. For 8 × 8 blocks, full-sample accuracy is used, and for 4 × 4 blocks, half-sample accuracy is used.

At the decoder the signal of the block is reconstructed using the transmitted partition information. Thus, the wedgelet partition information for this mode is not predicted.

2) Intra-Predicted Wedgelet Partition: In this depth modeling mode, the wedgelet partition is predicted from data of previously coded blocks in the same picture, i.e., by intra-picture prediction. For a better approximation, the predicted partition is refined by varying the line end position. Only the offset to the line end position is transmitted in the bitstream and at the decoder the signal of the block is reconstructed using the partition information that results from combining the predicted partition and the transmitted offset.

The prediction process of this mode derives the start position and gradient of the line from the information of previously coded blocks, i.e., from the neighboring blocks left and above of the current block. In this depth modeling mode, two main prediction methods have to be distinguished. The first method covers the case when one of the two neighboring reference blocks is of type wedgelet, shown in the example in Fig. 4, left. The second method covers the case when the two neighboring reference blocks are of type intra direction, which is the default intra coding type, shown in the example in Fig. 4, right. In any other case (e.g., the neighboring blocks are not available) this mode is carried out using meaningful default values.

If the reference block is of type wedgelet, the principle of this method is to continue the reference wedgelet into the current block, which is only possible if the continuation of the separation line of the reference wedgelet actually intersects the current block. In this case, the start position S_p and end position E_p (as illustrated in Fig. 4, left) are calculated as the intersection points of the continued line with block border samples.

If the reference block is of type intra direction, a gradient m_{ref} is derived from the intra prediction direction first. As the intra direction is only provided in the form of an abstract index, a mapping or conversion function is defined that associates each intra prediction mode with a gradient. Second, the start position S_p is derived from the adjacent samples of the left and above neighboring block. Among these samples, the maximum slope is obtained (as indicated by the largest grey value difference in the above neighboring block in Fig. 4, right) and S_p is assigned to the corresponding position in the current block. This information is also available at the decoder. Finally, the end position E_p is calculated from the start point and the gradient m_p, initially assuming $m_p = m_{ref}$.

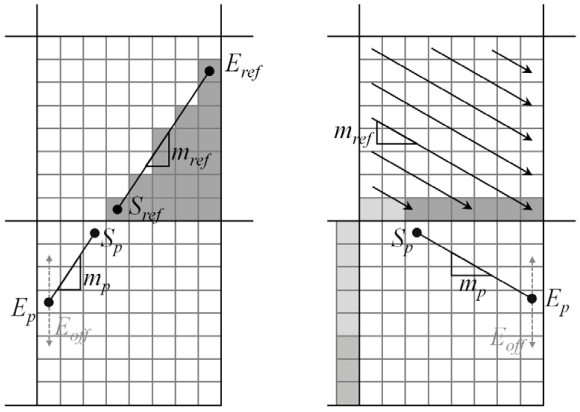

Fig. 4. Intra prediction of wedgelet partition (in bottom blocks) for the scenarios that the above reference block is either of type wedgelet partition (left) or regular intra direction (right).

The line end position offset for refining the wedgelet partition is not predicted, but sought within the estimation process at the encoder. For this, candidate partitions are generated from the predicted wedgelet partition and an offset value for the line end position E_{off} is obtained, as illustrated in Fig. 4. By iterating over a range of offset values and comparing the distortion of the different resulting wedgelet partitions, the offset value of the best matching wedgelet partition is determined using a distortion measure for the final partitioning pattern of the depth block.

3) Inter-Component Predicted Contour Partition: In this depth modeling mode, a contour partition is predicted from a texture reference block by inter-component prediction. Similar to the inter-component prediction of a wedgelet partition pattern described in the previous subsection, the reconstructed luminance signal of the co-located block of the associated video picture is used as a reference. In contrast to wedgelet partitions, a threshold method is used for the prediction of a contour partition. In this method, all samples of the reconstructed luminance block larger than the mean value of the block are categorized as region $\mathbf{P}_1$, while all samples smaller than the mean value are categorized as region $\mathbf{P}_2$, as shown in Fig. 3 bottom. The same block partitioning is carried out at the decoder, such that no partitioning or threshold information is transmitted. Finally, the obtained pattern is used as the depth block partitioning pattern.

B. Motion Compensation and Motion Vector Coding

In HEVC, eight-tap interpolation filters are used for motion-compensated interpolation. As experimentally verified, these filters are suitable for interpolating natural pictures, however they can produce ringing artifacts at sharp edges in depth maps, which are visible as disturbing components in synthesized interpolated views. Therefore, the motion-compensated prediction (MCP) as well as the disparity-compensated prediction (DCP) have been modified for depth map coding, such that no interpolation is used. Accordingly, the inter-picture prediction is always performed with full-sample accuracy. For the actual MCP or DCP, a block of samples in the reference picture is directly used as prediction signal without interpolating

any intermediate samples. In order to avoid the transmission of motion and disparity vectors with an unnecessarily high accuracy, full-sample accurate motion and disparity vectors are used for coding depth maps. The transmitted motion vector differences are coded using full-sample instead of quarter-sample precision.

C. Motion Parameter Inheritance

As the motion characteristics for the video and associated depth map in the MVD format is similar, inter-component motion vector prediction has been studied for H.264/MPEG-4 AVC in [37], and [11] for non-rectangular wedgelet partitioning, as well as in [9] for video and depth coding with H.262/MPEG-2 Video.

Accordingly, a new inter coding mode for depth maps is added in which the partitioning of a block into sub-blocks and associated motion parameters are inferred from the co-located block in the associated video picture. Since the motion vectors of the video signal are given in quarter-sample accuracy, whereas for the depth signal sample-accurate motion vectors are used, the inherited motion vectors are quantized to full-sample precision. For each block, it can be adaptively decided, whether the partitioning and motion information is inherited from the co-located region of the video picture, or new motion data is transmitted. For signaling the former case, also denoted as Motion Parameter Inheritance (MPI) mode, we modified the merge mode of HEVC. We extended the list of possible merge candidates, such that in depth map coding, the first candidate refers to merging with the corresponding block from the video signal. The usage of the merge mode syntax has the advantage that it allows very efficient signaling of the case where MPI is used without transmitting a residual signal, since the skip mode in HEVC also uses the merge candidate list [54].

V. Encoder Control

One of the main objectives of a video coding standard is to provide high compression efficiency. However, a video coding standard only specifies the decoding and parsing process together with the bitstream syntax and does not give any guarantee on the rate-distortion performance of compliant bitstreams. That means, that a corresponding video encoder can be configured in a flexible way to meet any given constraints on bit rate, quality, and computational complexity. Often, when evaluating the coding-efficiency capabilities of a given bitstream syntax, the encoder control is configured in a way to provide the best reconstruction quality at a given bit rate or vice versa, the minimum bit rate at a given quality [46]. This is equivalent to minimizing the Lagrangian cost function $J = D + \lambda \cdot R$ [3], which weights the distortion D that is obtained by coding a block in a particular mode or with a particular set of parameters with the number of bits R that is required for transmitting all data of that mode or parameter. Thus, D is an inverse measure of the reconstruction quality, while R is directly related to the total bit rate of the coded data. D and R are connected via the Lagrange multiplier λ that is usually derived based on the used quantization parameter. D is typically measured as the sum of squared differences (SSD) or the sum of absolute differences (SAD) between the original and the reconstructed sample values for video data.

Since reconstructed depth maps are only used for the synthesis of intermediate views and are not directly viewed, the coding efficiency can be improved by modifying the Lagrangian cost function. Coding errors in depth data cause artifacts in synthesized views and a modified distortion measure for depth coding is used, as explained in more detail below.

A. View Synthesis Optimization

For coding the depth maps, R is again measured as the total bit rate of the coded data. The depth distortion, however, has to be related to the distortion of synthesized views. Therefore, D needs to be calculated as the SSD between synthesized views from original and from reconstructed video and depth data. This is also referred to as synthesized view distortion (SVD) and has been studied recently in [5], [23], [29], [38]. Since standard-compliant video encoding algorithms operate block-based, the mapping of depth distortion to the synthesized view distortion must be block-based as well. Moreover, the sum of partial distortions (of sub-blocks) must be equal to the overall distortion of a block to enable an independent distortion calculation for all partitions of a subdivided block, as hierarchical block structures are common elements of modern video coding standards. However, disocclusions and occlusions prevent a bijective mapping of the distorted depth map areas to distorted areas in the synthesized view. E.g., areas in the synthesized view, which depend on depth data of an evaluated block, can become visible due to the distortions of other depth blocks; or vice versa, the distortion of a depth block has no effect on the synthesized view, since the block is occluded there. Hence, an exact mapping between the distortion of a block of the depth data and an associated distortion in the synthesized view is not possible regarding only the depth data within a currently processed block. Therefore, the SVD method needs to be extended to calculate the exact synthesized view distortion change (SVDC) [50] for a particular rendering algorithm and intermediate viewing position, as shown in the next subsection.

B. Synthesized View Distortion Change

In the SVDC method, the change of overall distortion of the synthesized view depending on the change of the depth data within a depth block is calculated while simultaneously also considering depth data outside that block. For this, the SVDC is defined as distortion difference of two synthesized textures.

Fig. 5 illustrates the SVDC calculation. First, for each tested coding mode for the current depth block, two variants of depth data are used: Variant d_1 consists of reconstructed depth values for already coded blocks and uncoded depth values for the remaining blocks, see gray and white blocks in Fig. 5 top, respectively. Variant d_2 is similar, but for the current block, the reconstructed depth values from the actual mode under test are used, as shown by the shaded area in Fig. 5 bottom. Both depth variants d_1 and d_2 are further used to synthesize portions of intermediate views v_1 and v_2 with coded and reconstructed video data t_{Cod} as explained in the next section. For the SSD

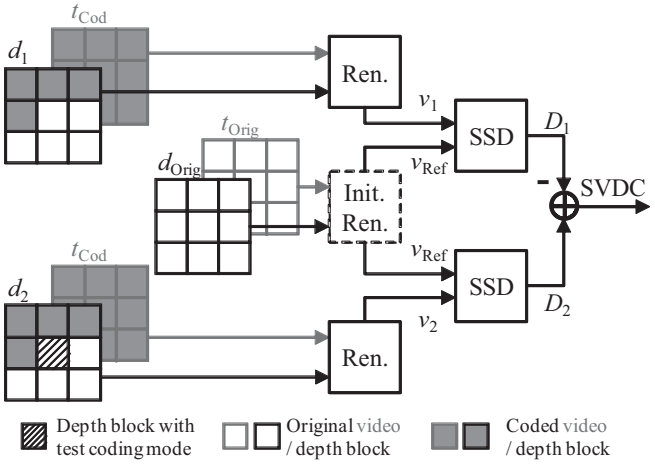

Fig. 5. Synthesized view distortion change (SVDC) calculation with respect to a currently tested depth coding mode (Ren.: view synthesis with encoder-side rendering module per block, Init.Ren.: Initial reference view synthesis per picture).

calculation, also the reference portion v_{Ref} is available. It was synthesized in the initialization phase from uncoded video and depth data t_{Orig} and d_{Orig}, as described in the next section. Next, both distortions can be calculated: $D_1 = \text{SSD}(v_1, v_{Ref})$ for depth variant 1 and $D_2 = \text{SSD}(v_2, v_{Ref})$ for depth variant 2 with the current coding mode under test. Finally, the difference between these values is used as depth distortion measure: $\text{SVDC} = D_2 - D_1$.

C. Encoder-Side Render Model

The computation of SVDC requires rendering functionalities in the encoding process, as depth data is used to shift video data to the correct position in a synthesized view. However, since computational complexity is a critical factor in distortion calculation, a simplified method has been utilized that allows minimal re-rendering of only those parts of the synthesized view that are affected by a depth distortion. This encoder-side render model provides the basic functions of most rendering approaches, including sub-pixel accurate warping, hole filling and view blending. Thus, all basic processing steps of common depth-based view synthesis algorithms are considered by the rate-distortion optimization process at the encoder. Accordingly, the rate-distortion optimization for depth maps can be performed independently of the final decoder-side view synthesis algorithm.

The render model consists of an initialization step for each picture, as well as a block-wise iteration with re-rendering and SVDC calculation. During picture-wise initialization, the reference views v_{Ref} (see Fig. 5) are synthesized. The block-wise re-rendering first warps the video data block from an original viewing position to the synthesized position using the associated depth data. Next, an up-sampling by a factor of 4 of the warped video block is carried out in order to obtain a more accurate value for the full pixel position in the synthesized view. Small holes of one pixel are filled by linear interpolation from the up-sampled video data. As the warping step is aware of the warping direction, larger holes or disocclusions are filled by the last warped pixel, which is automatically a background pixel, while occluded background areas are overridden by foreground. Thus, no z-buffer comparisons, e.g., for front-most pixel detection are required. After warping the individual views, weighted view or α-blending with the corresponding block of the second view can be carried out [36]. For this, also the second view is synthesized in the picture-wise initialization step [51]. The synthesized portion of the intermediate view is then used to calculate the SSD, using the corresponding portion of the reference view from the initialization step. Finally, the SVDC is obtained as described in the previous section. For view synthesis optimization of a set of N intermediate views, the total SVDC is calculated by averaging the individual SVDCs at each position.

To enable rate-distortion optimization using SVDC, the render model is integrated into the encoding process for depth data by replacing the conventional distortion computation with the SVDC computation in all processing steps related to the mode decision. Finally, the Lagrangian cost functional becomes $J = SVDC + l_s \cdot \lambda \cdot R$ with l_s being a constant scaling factor. In our experiments, $l_s = 0.5$ gave the best results. The Lagrangian multiplier λ was calculated similar to the calculation in the HEVC test model encoder.

VI. Depth-Based View Synthesis Algorithm

After decoding the 3D video content, a decoder-side synthesis algorithm generates the required number of dense views for a particular multi-view display. Since the proposed 3D video codec produces a view- and component-scalable bitstream, two main synthesis approaches can be applied: View synthesis from a video-only decoded bitstream and view synthesis from a full MVD decoded bitstream. The first approach only operates on the decoded video data and is described in detail in [27]. The second approach is based on classical DIBR functionality [36]. This depth-based view synthesis operates picture-wise and extends the functionality of the encoder-side rendering module, described in Section V-C. The similarities and differences of both modules are summarized in TABLE I for the individual rendering steps in processing order.

TABLE I shows, that the major parts, namely warping, hole-filling and view-blending are rather similar. The main operation of the decoder-side depth-based synthesis is similar to the encoder-side render module, as already described in Section V-C. In addition, the decoder-side depth-based synthesis operates picture-wise with an initial up-sampling of the luminance and chrominance component for higher precision before warping. Furthermore, a more complex algorithm for filling disoccluded areas is used, where a reliability map is created after warping and small hole filling. This reliability map assigns an 8 bit value from 0 to 255 for each pixel of a warped view. Here, disoccluded areas are rated with 0, i.e., unreliable. In the six-sample wide boundary area next to a disocclusion, the reliability linearly increases from 0 to 255. All remaining areas are set to 255. The reliability information is used in the view blending process in addition to the intermediate view position to obtain the blending weights.

TABLE I
METHOD COMPARISON BETWEEN DECODER-SIDE DEPTH-BASED VIEW SYNTHESIS ALGORITHM AND ENCODER-SIDE RENDER MODEL

Processing Step	Decoder-Side Depth-Based View Synthesis	Encoder-Side Render Model
Operation Mode	Picture-wise	Block-wise
Up-Sampling	4x for luminance, 8x for chrominance	Not applied
Warping	Line-wise directional for faster occlusion/ disocclusion handling	Line-wise directional for faster occlusion/ disocclusion handling
Interpolation and Hole Filling	Small holes by interpolation, disocclusions from last warped pixel as background	Small holes by interpolation, disocclusions remain
Reliability Map Creation	Reliability map for dissoclusions and object boundaries, see [36]	Not applied
Similarity Enhancement	Accumulated histogram equalization between views	Not applied
View Blending	α-blending with weights, based on distances to original viewing positions and reliability map values for dissocclusion handling	a-blending with weights, based on distances to original viewing positions
Chrominance Decimation	Down-filtering of chrominance channels, if necessary for display format (e.g. for 4:2:0)	Not applied

Accordingly, disoccluded areas with a reliability value of 0 are filled from the other view, while other areas are interpolated from either warped view, according to their blending weights. After blending, the decoder-side depth-based synthesis also needs to adapt to the required display format, e.g., by down-sampling the chrominance channels.

VII. SIMULATION RESULTS

The CfP for 3D video technology [15] specified two test categories: AVC-compatible and HEVC-compatible/unconstrained. The 3D video codec, described in this paper, was proposed for the HEVC-compatible category, i.e., the base view must be decodable by an HEVC decoder. The 3D video test set consisted of 8 sequences with the video component in 4:2:0 chroma format: 4 with a progressive HD resolution of 1920 × 1088 luma/depth samples with 25 fps and 4 with a progressive resolution of 1024 × 768 luma/depth samples with 30 fps (test sequences 1–4 and 5–8 in TABLE II, respectively). All 8 sequences were evaluated in two test scenarios: In the 2-view scenario, video and depth component of 2 views {V0,V1} were coded and a stereo pair with one original and one intermediate synthesized viewing position reconstructed and rendered. This stereo pair was evaluated on a stereoscopic display. In the 3-view scenario, video and depth component of 3 views {V0,V1,V2} were coded and three types of video data extracted. First, a central stereo pair in the middle of the 3-view range and second a random stereo pair within the viewing range were synthesized and viewed on a stereoscopic display. Third, a dense range of 28 views was synthesized and evaluated on an autostereoscopic 28-view display [15]. All results were evaluated in large-scale subjective tests [17], including the entire set of new coding tools, described in this paper. For individual coding results on single tools or a subset of tools, the reader is referred to [45] for inter-view motion parameter prediction, [54] for motion parameter inheritance, [31] for depth intra coding and [50], [51] for view synthesis optimization.

TABLE II
RATE POINTS FOR 2-VIEW AND 3-VIEW TEST SCENARIO FOR HEVC-BASED 3D VIDEO TECHNOLOGY

Test Sequence	2-View Test Scenario Bit Rates (kbps)				3-View Test Scenario Bit Rates (kbps)			
	R1	R2	R3	R4	R1	R2	R3	R4
Poznan_Hall2	140	210	320	520	210	310	480	770
Poznan_Street	280	480	800	1310	410	710	1180	1950
Undo_Dancer	290	430	710	1000	430	780	1200	2010
GT_Fly	230	400	730	1100	340	600	1080	1600
Kendo	230	360	480	690	280	430	670	1040
Balloons	250	350	520	800	300	480	770	1200
Lovebird1	220	300	480	830	260	420	730	1270
Newspaper	230	360	480	720	340	450	680	900

For the coding of the 2-view and 3-view scenario, 4 different rate points were defined prior to the call as shown in TABLE II and anchors provided at these bit rates, where each video and depth component was separately coded with the HEVC test model HM, version 3.0. For the 3D video test sets, a random access of $\leq$0.5 sec was required. Accordingly, we used temporal prediction structures with hierarchical B pictures [44] with groups of pictures (GOPs) of 12 and 15 for test sequences with 25 and 30 fps, respectively and for all coded video and depth components.

Temporal QP cascading was used similar to the HEVC test conditions [16], where pictures in random access units are coded with a given QP, while for each temporal hierarchy level, the QP is increased by 1. The independent or base view was V1 for both scenarios. Inter-view cascading is restricted to the given QP for the independent view and QP + 3 for all dependent views. The QP offset $\Delta\mathrm{QP}_D$ for the depth QP_D in relation to the Video QP ($\mathrm{QP}_D = QP + \Delta\mathrm{QP}_D$) was fixed based on subjective assessments and varies from $\Delta\mathrm{QP}_D = 0$ for video QP = 51 at lowest quality up to $\Delta\mathrm{QP}_D = 9$ for video QP $\leq$ 32 at high quality. With these settings, the same encoder configuration is used for all sequences and rate points.

A. Objective Performance

The proposed 3D video codec has been evaluated by objective comparisons with the anchor coding, as well as two improved methods. For this, intermediate views have been generated for each method at every $^1/_{16}$-th position between the coded views, such that 15 and 30 intermediate views have been generated for the 2-view and 3-view scenario respectively. Then, the PSNR values have been determined, comparing the decoded synthesized views with synthesized views from original uncoded video and depth data. From the PSNR values and the overall bit rate of all methods, the bit rate

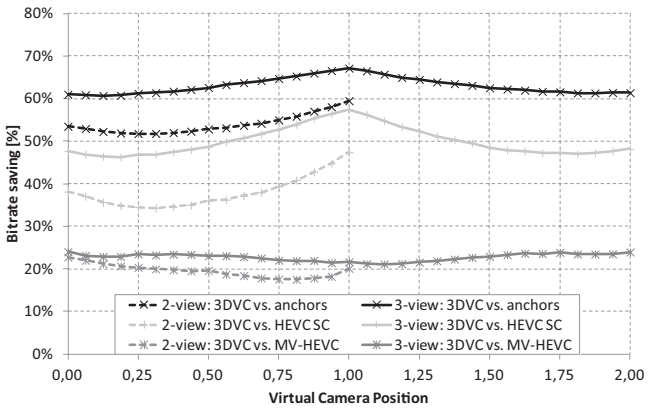

Fig. 6. Average bit rate savings of the proposed 3DVC extension of HEVC relative to the HEVC simulcast anchors with GOP8, HEVC simulcast with GOP 12/15 and a straightforward multi-view extension of HEVC with inter-view prediction, for the 2-view and 3-view test scenario.

savings of our proposed method in comparison to the different references has been calculated, using the Bjøntegaard delta rates [4]. These average bit rate savings over all test sequences are shown in Fig. 6 for the 2-view and 3-view scenario.

First, a comparison of our 3DVC extension of HEVC against the anchor coding with HEVC simulcast for each video and depth component and GOP8 is shown. This case was also evaluated within the large-scale subjective tests as shown in [17]. Here, bit rate savings of more than 60% for the 3-view case and more than 50% for the 2-view case were achieved. For the 3-view scenario, two dependent views (V0 and V2) benefit from the additional inter-view prediction tools, such that the gains are higher than for the 2-view scenario. For both scenarios, the highest average bit rate savings were achieved for the virtual camera position 1.0, i.e., at the position of the independent base view. For this view, most of the bit rate has been spent, as is analyzed in Section VII-C. Beside the disparity-compensated prediction and the additional coding tools in our proposed 3DVC extension of HEVC, the obtained bit rate savings can also partly be attributed to the usage of a larger GOP sizes in comparison to the anchors.

Therefore, we second compared the coding efficiency also to an HEVC simulcast version (labeled as "HEVC SC" in Fig. 6) that uses the same GOP structure as the developed HEVC extension for a fair comparison of the developed codec with HEVC simulcast. The corresponding gains result only from the usage of disparity-compensated prediction and new coding tools in our proposal. Again, the gains for the 3-view scenario are significantly higher than for the 2-view scenario due inter-view prediction for two dependent views.

Third, bit rate savings are shown for our codec in comparison to a straightforward HEVC extension to multiple views (MV-HEVC), which only includes disparity-compensated prediction as additional coding tool. These average gains of 23% for the 3-view scenario and 20% for the 2-view scenario thus show the improvements that were achieved by the new coding tools, described in the previous chapters.

Next, bit rate savings of the HEVC 3DVC extension vs. HEVC SC and MV-HEVC for the 2- and 3-view scenario have been averaged in TABLE III over all virtual camera positions. In addition, the obtained savings are also given for each test sequence.

TABLE III
AVERAGE BIT RATE SAVINGS COMPARED TO STRAIGHTFORWARD HEVC MULTI-VIEW EXTENSION (MV-HEVC) AND HEVC SIMULCAST (HEVC SC)

Test Sequence	2-View Test Scenario Bit Rate Savings (%)		3-View Test Scenario Bit Rate Savings (%)	
	MV-HEVC	HEVC SC	MV-HEVC	HEVC SC
Poznan_Hall2	20.10	35.10	22.04	44.96
Poznan_Street	11.97	37.60	14.41	51.07
Undo_Dancer	6.56	36.74	12.50	53.19
GT_Fly	16.17	44.23	20.42	57.61
Kendo	37.15	42.74	40.31	53.27
Balloons	27.81	37.85	31.17	49.53
Lovebird1	16.22	34.93	18.09	47.10
Newspaper	20.28	35.20	23.02	43.16
Average	19.53	38.05	22.75	49.99

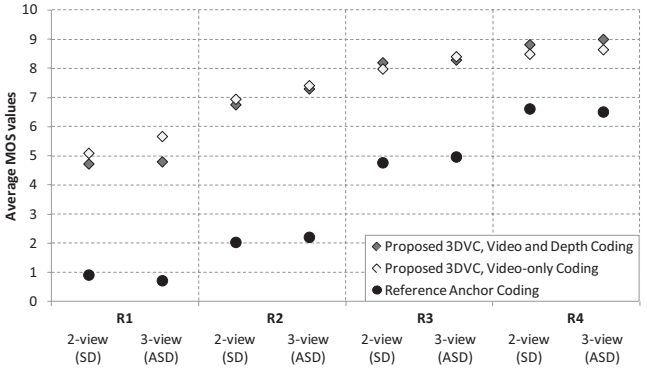

Fig. 7. Averaged MOS scores over all eight test sequences at four different bit rates R1-R4 according to Table II: Evaluation of 2-view scenario on stereoscopic display (SD), evaluation of 3-view scenario on auto-stereoscopic 28-view display (ASD).

Besides the improved compression efficiency of the proposed 3DVC extension of HEVC, also the complexity has been measured as the average runtime in comparison to the HEVC anchor configuration. For the 2-view scenario, the proposed codec showed a relative runtime of 172% for the encoder, 121% for the decoder and 51% for the decoder-side view synthesis (in comparison to the view synthesizer software provided by MPEG). For the 3-view scenario, the relative runtimes are 283% for the encoder, 128% for the decoder and 48% for the view synthesis.

B. Subjective Performance

All proposals, submitted to the CfP, were evaluated in large-scale subjective tests [17]. From this, the viewing results in terms of average MOS values over all sequences are shown in Fig. 7 for our proposed 3DVC extension of HEVC.

Here, the viewing results of the 2-view scenario on a stereoscopic display with polarized glasses (SD) as well as results of the 3-view scenario on an auto-stereoscopic 28-view display (ASD) are given for each rate point. Fig. 7 shows

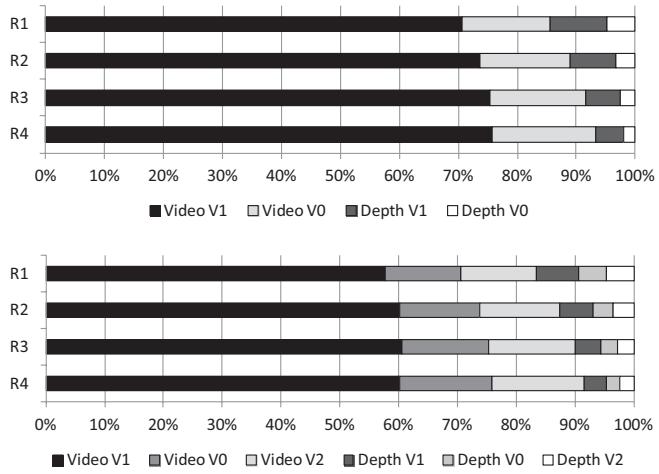

Fig. 8. Average bit rate distribution in percent of total rate over all test sets for the video and depth components of the 2-view scenario with views V0, V1 (top) and 3-view scenario with views V0, V1, V2 (bottom) for target rate points R1-R4. In both scenarios, V1 is the independent base view.

the average MOS values of the anchor coding with HEVC simulcast and our proposal with two settings: Video and depth coding, as well as video-only coding with corresponding decoder-side rendering, using image domain warping [10]. The proposed codec in both settings significantly outperforms the anchor coding, such that, e.g., a similar subjective quality of 5 for the anchor coding at R3 is already achieved by our proposal at R1. Comparing the associated individual bit rates of each sequence at R3 and R1, as given in TABLE II, average bit rate savings of 57% and 61% for the 2- and 3-view case respectively are achieved. This is also similar to the bit rate savings based on objective PSNR measure, as reported in the previous section.

Comparing the two settings of the proposed 3DV codec in Fig. 7, the video-only setting was rated slightly better at the lower rate points R1 and R2. Contrary, the video and depth setting was rated slightly better at the high rate point R4.

Comparing the 2- and 3-view scenario in Fig. 7, the subjective MOS results are very consistent for each rate point. Since very different 3D displays were used for the viewing, the proposed codec is thus able to provide a display-independent reconstruction quality from a decoded generic 3D video format.

C. Bit Rate Distribution

For our 3DVC extension of HEVC, the average bit rate distribution in percent of the total rate between the video and depth components over all test sets is shown in Fig. 8 for the 2- and 3-view scenario at all rate points.

Fig. 8 first shows, that most of the bit rate is distributed to the video component of the independent base view V1 with more than 50% for the 2-view as well as for the 3-view scenario. That means, efficient 3D video transmission of MVD data can be achieved at less than double the bit rate of a 2D video transmission based on HEVC.

Fig. 8 also shows that most of the bit rate is allocated to the video data. For the 2-view scenario, the video/depth rate distribution varies from 86%/14% at the lowest rate point R1 to 93%/7% at the highest rate point R4 on average. For the 3-view scenario the video/depth rate distribution ranges from 83%/17% at R1 to 92%/8% at R4. Thus, depth data can be coded very efficiently. Therefore, if multi-view video with depth data is transmitted at the same overall bit rate as video-only data, the video bit rate portion in MVD only needs to be around 8% for a good-quality as provided by rate point R4 in the simulations. The perceived video quality in both scenarios is almost identical. In addition, the MVD transmission provides depth data that are available at the decoder for high quality view synthesis in a 3D display. Overall, the MVD approach was even slightly better subjectively rated at the high rate point R4, as shown in Fig. 7.

VIII. Conclusion

We presented a 3D video codec as an extension of the HEVC standard for coding of multi-view video plus depth formats for stereoscopic and autostereoscopic multi-view displays. Besides disparity-compensated prediction, a number of new tools were developed, including inter-view motion parameter and inter-view residual prediction for the video component of dependent views. Furthermore, novel intra coding modes, modified motion compensation and motion vector coding, motion parameter inheritance as well as a new encoder control for the depth data were presented. The encoder control uses view synthesis optimization, which guarantees that high quality intermediate views can be generated at the decoder. At the decoder, a subset of the coded components can be extracted separately, such that 2D video, stereo video or full MVD can be reconstructed from the 3D video bitstream. In addition, our depth coding methods can also be used together with other hybrid video coding architectures such as, e.g., H.264/AVC or MVC.

The described technology was submitted to the MPEG Call for Proposals on 3D Video Coding Technology. As an outcome of the corresponding subjective tests, our proposed 3D video coding extension of HEVC performed best among all submissions. Furthermore, objective results were shown, where the proposed codec provides about 50% bit rate savings in comparison to HEVC simulcast and about 20% in comparison to a straightforward multi-view extension of HEVC (MV-HEVC) without the developed coding tools. Consequently, our technology was selected as the starting point for the standardization of the 3D video and multi-view extensions of HEVC. All described coding and synthesis tools were included in the first version of the reference software 3D-HTM for the standardization process.

Appendix A

Downloadable Resources Related to This Paper

All JCT-3V documents are publically available and can be accessed through the JCT-3V document management system at http://phenix.it-sudparis.eu/jct3v/.

ACKNOWLEDGMENT

The authors would like to thank Disney Research Zurich for providing the view generation technology, which was used for rendering output views from decoded video-only data.

They would also like to thank Poznan University of Technology, Nokia Research, Nagoya University, Gwangju Institute of Science and Technology, Electronics and Telecommunications Research Institute, MPEG-Korea Forum ETRI for providing the *Poznan_Hall, Poznan_Street, Undo_Dancer, GT_Fly, Kendo, Balloons, Lovebird1* and *Newspaper* 3D video test data sets.

REFERENCES

[1] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing," *IEEE Trans. Circuits Syst. Video Technol., Special Issue Immersive Telecommun.*, vol. 14, no. 3, pp. 321–334, Mar. 2004.

[2] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. V. Kopylow, "A survey of 3DTV Displays: Techniques and technologies," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1647–1658, Nov. 2007.

[3] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[4] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in *Proc. VCEG-M33 Meeting*, 2001, pp. 1–4.

[5] G. Cheung, J. Ishida, A. Kubota, and A. Ortega, "Transform domain sparsification of depth maps using iterative quadratic programming," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 129–132.

[6] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 8, Jan. 2009.

[7] I. Daribo, G. Cheung, and D. Florencio, "Arithmetic edge coding for arbitrarily shaped sub-block motion prediction in depth video compression," in *Proc. IEEE Int. Conf. Image Process.*, Orlando, FL, USA, Oct. 2012, pp. 1541–1544.

[8] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Adaptive wavelet coding of the depth map for stereoscopic view synthesis," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Cairns, Australia, Oct. 2008, pp. 34–39.

[9] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Motion vector sharing and bit rate allocation for 3D video-plus-depth coding," *EURASIP J. Adv. Signal Process., Special Issue 3DTV*, vol. 2009, no. 258920, pp. 1–13, Jan. 2009, doi:10.1155/2009/258920.

[10] M. Farre, O. Wang, M. Lang, M. Stefanoski, A. Hornung, and A. Smolic, "Automatic content creation for multiview autostereoscopic displays using image domain warping," in *Proc. IEEE Int. Conf. Multimedia Exposit.*, Barcelona, Spain, Jul. 2011, pp. 1–6.

[11] R. Ferreira, E. Hung, R. de Queiroz, and D. Mukherjee, "Efficiency improvements for a geometric-partition-based video coder," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 1009–1012.

[12] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Mach. Vis. Appl.*, vol. 12, no. 1, pp. 16–22, Jan. 2000.

[13] X. Guo, Y. Lu, F. Wu, and W. Gao, "Inter-view direct mode for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 12, pp. 1527–1532, Dec. 2006.

[14] P. Helle, S. Oudin, B. Bross, D. Marpe, M. O. Bici, K. Ugur, J. Jung, G. Clare, and T. Wiegand, "Block merging for quadtree-based partitioning in HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1720–1731, Dec. 2012.

[15] *Call for Proposals on 3D Video Coding Technology*, Standard ISO/IEC JTC1/SC29/WG11, Mar. 2011.

[16] *Joint Call for Proposals for Video Coding Technology*, Standard ISO/IEC JTC1/SC29/WG11, Jan. 2010.

[17] *Report of Subjective Test Results from the Call for Proposals on 3D Video Coding*, Standard ISO/IEC JTC1/SC29/WG11, Nov. 2011.

[18] *Text of ISO/IEC DIS 23008-2 High Efficiency Video Coding*, Standard ISO/IEC JTC1/SC29/WG11, July 2012.

[19] *The Virtual Reality Modeling Language*, Standard ISO/IEC DIS 14772-1, Apr. 1997.

[20] *Advanced Video Coding for Generic Audiovisual Services*, Standard ISO/IEC JTC 1, Mar. 2012.

[21] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Process., Image Commun., Special Issue 3DTV*, vol. 22, no. 2, pp. 217–234, Feb. 2007.

[22] S.-Y. Kim and Y.-S. Ho, "Mesh-based depth coding for 3D video using hierarchical decomposition of depth maps," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, USA, Sep. 2007, pp. V-117–V-120.

[23] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," *Proc. SPIE*, vol. 7543, pp. 75430B-1–75430B-10, Jan. 2010.

[24] J. Konieczny and M. Domanski, "Depth-based inter-view prediction of motion vectors for improved multiview video coding," in *Proc. IEEE True Vis., Capture, Transmiss. Display 3D Video*, Tampere, Finland, Jun. 2010, pp. 1–4.

[25] J. Konrad and M. Halle, "3-D displays and signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 97–111, Nov. 2007.

[26] H.-S. Koo, Y.-J. Jeon, and B.-M. Jeon, "Motion skip mode for MVC," ITU-T and ISO/IEC JTC1, Hangzhou, China, Tech. Rep. JVT-U091, Oct. 2006.

[27] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Trans. Graph.*, vol. 29, no. 3, pp. 75:1–75:10, Jul. 2010.

[28] E.-K. Lee, Y.-K. Jung, and Y.-S. Ho, "3-D video generation using foreground separation and disocclusion detection," in *Proc. IEEE True Vis., Capture, Transmiss. Display 3D Video*, Tampere, Finland, Jun. 2010, pp. 1–4.

[29] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model," *Signal Process., Image Commun.*, vol. 24, no. 8, pp. 666–681, Aug. 2009.

[30] M. Maitre and M. N. Do, "Shape-adaptive wavelet encoding of depth maps," in *Proc. Picture Coding Symp.*, Chicago, IL, USA, May 2009, pp. 1–4.

[31] P. Merkle, C. Bartnik, K. Müller, D. Marpe, and T. Wiegand, "3D video: Depth coding based on inter-component prediction of block partitions," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 149–152.

[32] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Signal Process., Image Commun.*, vol. 24, nos. 1–2, pp. 73–88, Jan. 2009.

[33] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461–1473, Nov. 2007.

[34] Y. Morvan, D. Farin, and P. H. N. de With, "Platelet-based coding of depth maps for the transmission of multiview images," *Proc. SPIE*, vol. 6055, p. 60550K, Jan. 2006.

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

[36] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3D video systems," *EURASIP J. Image Video Process., Special Issue 3D Image Video Process.*, vol. 2008, no. 438148, pp. 1–11, 2008, doi:10.1155/2008/438148.

[37] H. Oh and Y. Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," in *Proc. Pacific-Rim Symp. Image Video Technol.*, Hsinchu, Taiwan, Dec. 2006, pp. 898–907.

[38] B. T. Oh, J. Lee, and D.-S. Park, "Depth map coding based on synthesized view distortion function," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1344–1352, Nov. 2011.

[39] K.-J. Oh, S. Yea, A. Vetro, and Y.-S. Ho, "Depth reconstruction filter and down/up sampling for depth coding in 3-D video," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 747–750, Sep. 2009.

[40] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[41] A. Redert, M. O. de Beeck, C. Fehn, W. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman, "ATTEST-advanced three-dimensional television system techniques," in *Proc. Int. Symp. 3D Data Process., Visualizat. Transmiss.*, Jun. 2002, pp. 313–319.

[42] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, May 2002.

[43] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, D. Marpe, P. Merkle, K. Muller, H. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D video coding using advanced prediction, depth modeling, and encoder control methods," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 1–4.
[44] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. IEEE Int. Conf. Multimedia Expo*, Toronto, ON, Canada, Jul. 2006, pp. 1929–1932.
[45] H. Schwarz and T. Wiegand, "Inter-view prediction of motion data in multiview video coding," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 101–104.
[46] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 2163–2177, Jul. 1948.
[47] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-D warping with depth map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1485–1495, Nov. 2007.
[48] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[49] T. K. Tan, A. Fujibayashi, Y. Suzuki, and J. Takiue, "[AHG 8] objective and subjective evaluation of HM5.0," Joint Collaborative Team on Video Coding, San Jose, USA, Tech. Rep. JCTVC-H0116, Feb. 2012.
[50] G. Tech, H. Schwarz, K. Müller, and T. Wiegand, "3D video coding using the synthesized view distortion change," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 25–28.
[51] G. Tech, H. Schwarz, K. Müller, and T. Wiegand, "Effects of synthesized view distortion based 3D video coding on the quality of interpolated and extrapolated views," in *Proc. ICME IEEE Int. Conf. Multimedia Exposit.*, Melbourne, Australia, Jul. 2012, pp. 634–639.
[52] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/AVC standard," in *Proc. IEEE, Special Issue 3D Media Displays*, vol. 99, no. 4, pp. 626–642, Apr. 2011.
[53] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.
[54] M. Winken, H. Schwarz, and T. Wiegand, "Motion vector inheritance for high efficiency 3D video plus depth coding," in *Proc. Picture Coding Symp.*, Krakow, Poland, May 2012, pp. 53–56.

Karsten Müller (M'98–SM'07) received the Dr.-Ing. degree in electrical engineering and Dipl.-Ing. degree from the Technical University of Berlin, Berlin, Germany, in 2006 and 1997, respectively.

He has been with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Berlin, since 1997, where he is currently the Head of the 3-D Coding Group, Image Processing Department. His current research interests include representation, coding and reconstruction of 3-D scenes in free viewpoint video scenarios and coding, multi-view applications, and combined 2-D/3-D similarity analysis. He has been involved in international standardization activities, successfully contributing to the ISO/IEC Moving Picture Experts Group for work items on multi-view, multi-texture, and 3-D Video Coding. He co-chaired an ad hoc group on 3-D video coding from 2003 to 2012.

Heiko Schwarz received the Dipl.-Ing. degree in electrical engineering and the Dr.-Ing. degree from the University of Rostock, Rostock, Germany, in 1996 and 2000, respectively.

He joined the Image and Video Coding Group, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin, Germany, in 1999. He has contributed successfully to the standardization activities of the ITU-T Video Coding Experts Group (ITU-T SG16/Q.6-VCEG) and the ISO/IEC Moving Picture Experts Group (ISO/IEC JTC 1/SC 29/WG 11—MPEG).

Dr. Schwarz has been a Co-Editor of ITU-T Rec. H.264 and ISO/IEC 14496-10 and a Software Coordinator for the SVC reference software. He has been the Co-Chair of several ad hoc groups of the Joint Video Team of ITU-T VCEG and ISO/IEC MPEG.

Detlev Marpe (M'00–SM'08) received the Dipl.-Math. degree (Hons.) from the Technical University of Berlin (TUB), Berlin, Germany, in 1990, and the Dr.-Ing. degree from the University of Rostock, Rostock, Germany, in 2004.

Before joining the Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute (HHI), Berlin, in 1999, he was a Research Assistant with TUB, University of Applied Sciences, Berlin, and University Hospital Charité, Berlin. He is currently the Head of the Image and Video Coding Group, Fraunhofer HHI. He has successfully contributed to the standardization activities of the ITU-T Visual Coding Experts Group, the ISO/IEC Joint Photographic Experts Group, and the ISO/IEC Moving Picture Experts Group for still image and video coding. In the development of the H.264/MPEG-4 AVC standard, he was a Chief Architect of the CABAC entropy coding scheme, as well as one of the main technical and editorial contributors to the so-called fidelity range extensions with the addition of the High Profile in H.264/MPEG-4 AVC. He was one of the key people in designing the basic architecture of scalable video coding and multi-view video coding as algorithmic and syntactical extensions of H.264/MPEG-4 AVC. During the recent development of the H.265/MPEG-H HEVC base standard, he made significant contributions to the design of its fundamental building blocks. He has authored or co-authored more than 200 publications in the areas of image coding and signal processing. He holds numerous internationally issued patents and patent applications. His current research interests include still image and video coding, signal processing for communications as well as computer vision, and information theory.

Dr. Marpe was a co-recipient of two Technical Emmy Awards as a Key Contributor and a co-editor of the H.264/MPEG-4 AVC standard in 2008 and 2009. He was nominated for the German Future Prize 2012 and he received the 2011 Karl Heinz Beckurts Award, the 2009 Best Paper Award of the IEEE Circuits and Systems Society, the Joseph-von-Fraunhofer Prize in 2004, and the Best Paper Award of the German Society for Information Technology in 2004. As a co-founder of the Berlin-based daviko GmbH, he received the Prime Prize of the 2001 Multimedia Start-Up Competition of the German Federal Ministry of Economics and Technology.

Christian Bartnik received the Dipl. Ing. (FH) degree in communications engineering from the University of Applied Sciences, Berlin, Germany, in 2009.

In 2008, he joined the Image and Video Coding Group, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin. His current research interests include video coding, 3-D video, and compression of multiview video plus depth scenes.

Sebastian Bosse received the Diploma degree from RWTH Aachen University, Aachen, Germany, in 2008.

He is currently with the Image and Video Coding Group, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin, Germany. His current research interests include video compression, computer vision, and human visual perception.

Heribert Brust received the Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin, Germany, in 2009.

He has been with the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute Berlin, Berlin, Germany, since 2007. His current research interests include video coding, 3-D television, and stereoscopic video for mobile devices.

Tobias Hinz received the Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin, Berlin, Germany, in 1999.

He is currently a Research Engineer with the Department of Image Processing, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin. His current research interests include processing, coding and transmission of video and audio content, as well as software design and optimization.

Haricharan Lakshman is a Research Associate with Fraunhofer HHI, Berlin, Germany, and the Ph.D. degree with the Technical University of Berlin, Berlin, and the Bachelor of Engineering degree from NITK Surathkal, Surathkal, India, in 2002. From 2002 to 2006, he was an Engineer with Ittiam Systems, Bangalore, India. He received the M.S. degree from the University of Erlangen-Nuremberg, Erlangen, Germany, in 2008, while as a Research Assistant with Fraunhofer IIS, Erlangen. From 2011 to 2012, he was a Visiting Researcher with Stanford University, Stanford, CA, USA. His current research interests include image processing, video coding, 3-D video, and semantic analysis.

Philipp Merkle (S'06–M'12) received the Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin, Berlin, Germany, in 2006.

He joined the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Berlin, in 2003, and he has been a Research Associate since 2006. He has been involved in several projects focused on multiview video coding, 3-D television, free viewpoint video, and 3-D scene reconstruction. His current research interests include 3-D video, representation and compression of multiview video plus depth scenes, free viewpoint video, and 2-D and 3-D video-based rendering. He has been involved in ISO standardization activities where he contributed to the development of the MPEG-4 multiview video coding standard.

Philipp Merkle received the "Rudolf-Urtel Award" of the German Society for Technology in TV and Cinema (FKTG) for his work on multiview video coding in 2006.

Franz Hunn Rhee received the Dipl.-Ing. degree in electrical engineering from the Technical University of Hamburg-Harburg, Hamburg, Germany, in 2008.

He is currently with the 3-D Coding Group, Fraunhofer Institute for Telecommunications - Heinrich Hertz Institute, Berlin, Germany. His current research interests include signal processing, in particular multi-view video coding.

Gerhard Tech received the Dipl.-Ing. degree in electrical engineering from RWTH Aachen University of Technology, Aachen, Germany, where he has been involved in medical image processing and noise reduction.

He joined the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Berlin, Germany, in 2008, and he is a Research Assistant. His current research interests include video and image coding and processing, including stereo, multiview, and video plus depth representations. He has been involved in MPEG activities.

Martin Winken received the Diploma degree in computer engineering from the Technical University of Berlin, Berlin, Germany, in 2006. Currently, he is pursuing the Ph.D. degree in video compression technology.

He is currently a Research Engineer with the Image and Video Coding Group, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin. He has published conference contributions and contributed to standardization activity in video compression technology.

Thomas Wiegand (M'05–SM'08–F'11) received the Dipl.-Ing. degree in electrical engineering from the Technical University of Hamburg-Harburg, Hamburg, Germany, in 1995, and the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Erlangen, Germany, in 2000.

He is currently a Professor with the Department of Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany, a Chairing the Image Communication Laboratory, and is jointly heading the Image Processing Department, Fraunhofer Institute for Telecommunications–Heinrich Hertz Institute, Berlin. He joined the Heinrich Hertz Institute as the Head of the Image Communication Group, Image Processing Department in 2000. Since 1995, he has been an active participant in standardization for multimedia with successful submissions to ITU-T VCEG, ISO/IEC MPEG, 3GPP, DVB, and IETF. His current research interests include video processing and coding, multimedia transmission, and computer vision and graphics.

Dr. Wiegand was a recipient of the SPIE VCIP Best Student Paper Award in 1998, and the Fraunhofer Award and the ITG Award of the German Society for Information Technology in 2004. He co-chaired for development of the H.264/MPEG-4 AVC standard have been recognized by the 2008 ATAS Primetime Emmy Engineering Award and by a pair of NATAS Technology and Engineering Emmy Awards. In 2009, he received the Innovations Award of the Vodafone Foundation, the EURASIP Group Technical Achievement Award, and the Best Paper Award of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. In 2010, he received the Eduard Rhein Technology Award. In 2011, he received the Best Paper Award of EURASIP and the Karl Heinz Beckurts Award. In October 2000, he was the Associated Rapporteur of ITU-T VCEG. In December 2001, he was the Associated Rapporteur/Co-Chair of the JVT. In February 2002, he was the Editor of the H.264/MPEG-4 AVC video coding standard and its extensions (FREXt and SVC). From 2005 to 2009, he was the Co-Chair of MPEG Video. He is a recipient of the 2012 IEEE Masaru Ibuka Technical Field Award.