INVITED PAPER

3-D Video Representation Using Depth Maps

These representations are able to generate many views at the receiver and allow the acquisition format and transmission constraints to be decoupled from display requirements.

By Karsten Müller, *Senior Member IEEE*, Philipp Merkle, *Student Member IEEE*, and Thomas Wiegand, *Fellow IEEE*

ABSTRACT | Current 3-D video (3DV) technology is based on stereo systems. These systems use stereo video coding for pictures delivered by two input cameras. Typically, such stereo systems only reproduce these two camera views at the receiver and stereoscopic displays for multiple viewers require wearing special 3-D glasses. On the other hand, emerging autostereoscopic multiview displays emit a large numbers of views to enable 3-D viewing for multiple users without requiring 3-D glasses. For representing a large number of views, a multiview extension of stereo video coding is used, typically requiring a bit rate that is proportional to the number of views. However, since the quality improvement of multiview displays will be governed by an increase of emitted views, a format is needed that allows the generation of arbitrary numbers of views with the transmission bit rate being constant. Such a format is the combination of video signals and associated depth maps. The depth maps provide disparities associated with every sample of the video signal that can be used to render arbitrary numbers of additional views via view synthesis. This paper describes efficient coding methods for video and depth data. For the generation of views, synthesis methods are presented, which mitigate errors from depth estimation and coding.

KEYWORDS | Advanced video coding (AVC); depth estimation; H.264; multiview coding (MVC); multiview video plus depth (MVD); platelet coding; view synthesis; 3-D video (3DV) coding

Manuscript received March 17, 2010; revised July 12, 2010; accepted October 29, 2010. Date of publication December 17, 2010; date of current version March 18, 2011.

The authors are with the Fraunhofer Institute for Telecommunications—Heinrich Hertz Institute (HHI), 10587 Berlin, Germany (e-mail: karsten.mueller@hhi.fraunhofer.de; philipp.merkle@hhi.fraunhofer.de; thomas.wiegand@hhi.fraunhofer.de).

Digital Object Identifier: 10.1109/JPROC.2010.2091090

I. INTRODUCTION

> Color television! Bah, I won't believe it until I see it in black and white—*Samuel Goldwyn*

Three-dimensional video (3DV) may be the next step in the evolution of motion picture formats. This new format allows the representation of 3-D visual information through a display that provides the illusion of depth perception. Two-dimensional video signals offer a number of monocular cues for depth perception including linear perspective and occlusion. The extension to 3DV offers the sensation of depth from two slightly different projections of the scene onto the two eyes of the viewer. This also enables binocular cues, including stereopsis [18]. The local differences between the images on the retinas of the two eyes are called disparities. In addition to that, other differences may be present between the two images that can be mainly accounted to lighting effects and occlusions.

Current display technology for 3DV consists of flat screens, only offering the illusion of depth by representing the images that are seen by the two eyes with a parallax angle [24]. The parallax is the angle between the lines of sight that leads to the disparity between the two retinal images. The technology by which the parallax is created is the deciding factor for the 3DV format. Today most 3-D displays are stereo displays that require exactly two views at each instant. These views are typically exactly those views that are acquired by a stereo camera system. A stereo display for multiple viewers requires special 3-D glasses that filter the corresponding view for the left and right eyes of each viewer [2], [38]. From a compression point of view, a simple approach towards effectively representing a stereo video signal is given by treating them as two video signals with statistical dependency. The statistical dependency between the two views can be exploited by compression

techniques known from classical video coding as we will describe later.

For some stereo displays, the disparity in the stereo camera setup may not match the best viewing parallax for natural 3-D impression at the display. Hence, one of the two views needs to be repositioned. This process is called stereo repurposing. Typically, the two acquired views provide a good basis for computing another view in between them, using additional scene geometry information like depth or disparity data. Views that are not in-between the acquired views are more critical as background content is revealed, where no information from either view is available. In addition, the problem is more severe when the view generation needs to be done using compressed views as the quantization noise typically affects the estimation process for depth or disparity data that are required for view synthesis. Another issue with depth estimation at the receiver is that it typically requires significant computational resources and that different algorithms yield varying results. A final consideration should be given to the fact that the content owner would have limited control over the resulting displayed quality in case different depth estimation and view synthesis algorithms are used at the receiving ends.

In addition to stereo displays, also multiview displays are becoming increasingly available [2], [24], [38]. As multiview displays typically do not require 3-D glasses, one of the largest obstacles in user acceptance of 3DV is overcome. However, a multiview display requires the availability of many views. For example, current prototypes emit eight or nine views. It is expected that the quality improvement of multiview displays will be governed by an increase of emitted views and we can expect displays with 50 views or more in the future. Hence, a format is needed that allows the generation of arbitrary numbers of views, while the transmission bit rate is constant. Multiview synthesis based on a stereo video signal at the receiver suffers from the same problem as stereo repurposing. Here the problem is actually worse than for stereo repurposing, as many views need to be generated for multiview displays [24], [53]. Hence, a user perceives many viewing pairs, which consist of two synthesized views for a number of viewing positions, while the viewing pair for repurposed stereo consists of one original and one synthesized view.

One approach to overcome the problems with view generation at the receiver is to estimate the views at the sender and to transmit a signal that permits straightforward view synthesis at the receiver. Such a signal has to be related to the geometry of the scene, as we will show later. In this paper, we consider depth maps in combination with stereo video signals as an efficient representation for view synthesis at the receiver. We will motivate and validate this choice.

This paper provides an overview of 3DV involving depth maps and is organized as follows. Section II gives an introduction to stereo-based 3-D systems, requirements for stereo and multiview displays, and explains the limitations of formats, multiview coding techniques and view synthesis only from video data. Section III introduces depth enhancement for 3DV data, depth provision, depth coding methods and results, depth-image-based rendering (DIBR), and advanced view synthesis methods. Also, new coding methods for 3DV are discussed in Section III.

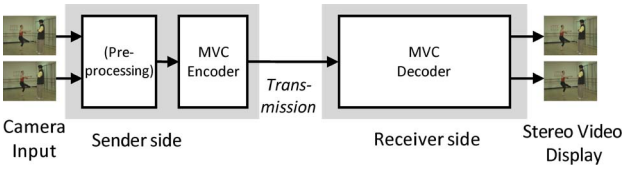

Fig. 1. ***First generation 3DV system based on stereoscopic color-only video.***

II. 3DV SOLUTIONS BASED ON STEREO SIGNALS

3DV systems based on stereo video signals are currently being commercialized for 3-D cinema, 3-D home entertainment, and mobile applications. These systems are based on stereo technology from capturing via coding and transmission to 3-D displays as shown in Fig. 1.

A stereo video signal captured by two input cameras is first preprocessed. This includes possible image rectification for row-wise left and right view alignment [12], as well as color and contrast correction due to possible differences between the input cameras. The 3-D format is called conventional stereo video (CSV) for left and right view. This format is encoded by multiview coding methods, such as specified in the stereo high profile of H.264/AVC [14], [16], where temporal dependencies in each view, as well as interview dependencies between both views are exploited for efficient compression.

Standardized solutions for CSV have already found their way to the market: 3-D cinema, Blu-Ray Disc, and broadcast. While 3-D cinema is based on JPEG-2000, the 3-D Blu-Ray Disc specification is based on the stereo high profile of H.264/AVC. For 2010, first 3-D Blu-Ray Discs and players as well as stereo broadcasting services have been announced. These stereo systems provide a robust and easy 3-D solution, since 3-D formats and coding only include stereo video data. Thus, complex processing, like provision and estimation of 3-D scene geometry or additional view synthesis are not required and coding methods can be optimized for the statistics of color video data. On the other hand, such systems are restricted to stereo displays that require glasses.

A. 3DV Display Comparison

One strong driving force towards new 3DV technology will be the availability of high-quality autostereoscopic (glasses-free) multiview displays. Here, a clear benefit of

Table 1 Comparison of Stereo and Multiview Display Properties

Property	Stereo Display	Multi-View Display
Viewing Aid	Mostly required	**Not required**
"Look around" Effect	No	**Yes**
Resolution per View	**High**	Low
Perceived Scene Depth	**High**	Low

multiview displays over current stereoscopic displays can be achieved for a multiuser audience. Although single-user autostereoscopic displays exist as well, their use is limited to niche applications. Currently, stereo and multiview display types show specific advantages and disadvantages, which are summarized in Table 1.

Here, the main properties for user acceptance are given and the boldface entries indicate which display type better fulfills them. Looking at these user preferences, multiview displays have the potential to become the first choice for 3DV, as they do not require viewing aid, like stereo glasses, for multiuser scenarios and give a more natural 3-D impression. If users move in front of the display, they expect a "look around" effect, i.e., they want to be able to see newly revealed background behind foreground objects. This can only be offered by multiview displays, as they provide multiple stereo pairs with slightly different content for each viewing position. Note that these first two properties are of systematic nature, i.e., related to the limitations of conventional stereo video, and therefore only supported by multiview displays.

However, for multiview displays to become widely acceptable, the disadvantages need to be eliminated. Such displays mostly suffer from a limited overall display resolution, as the number of available samples needs to be split over all N views. This leads to an optimization problem for the chosen number of views, since on the one hand, only few views give a higher resolution per view and on the other hand, more views are required for better 3-D viewing. The solution to this is the manufacturing of 3-D ultrahigh definition multiview displays, where, e.g., 50 views can be offered with each view in high resolution. This also improves the viewing angle problem of current multiview displays, as the viewing range becomes wider.

With such novel displays, the problem of limited depth range can be solved as well. In current displays, two contradicting requirements have to be fulfilled, which are a strong depth impression on the one hand and a seamless viewing change between neighboring stereo pairs on the other. The first condition requires a larger depth or disparity range, while the second condition requires a small range. If a 3-D ultrahigh resolution display with a large number of views will be used, the contradiction between both conditions can be resolved as follows: the disparity range between neighboring views can be made very small to provide seamless viewpoint change, but the actual stereo pairs for the user are nonneighboring pairs with a larger range for good depth perception. As an example, such a display may contain views $[1, 2, \ldots, 50]$. The disparity range between neighboring views $(1, 2)$, $(2, 3)$, etc., is very small, however a user perceives the stereo pair $(2, 6)$ with a four times higher disparity range. When moving to the left or right, the user sees pairs $(1, 5)$ and $(3, 7)$, respectively. Thus, the viewpoint change does not cause a jumping effect and a high depth perception is maintained.

B. Multiview Video Coding

For multiview displays, a high number of views has to be provided at the receiver, as discussed above. One possibility would be to transmit this high number of views, using the multiview coding (MVC) profile of H.264/AVC [14], [16], [46]. MVC is based on the single-view video compression standard H.264/AVC [14], [48]. In multiple view scenarios, the camera views share common scene content, such that a coding gain is achievable by exploiting statistical dependencies in spatially neighboring views. For this, multiview compression was investigated and the corresponding standardization activity led to the MVC extension [8], [14], [16] as an amendment to H.264/AVC.

An MVC coder basically consists of N parallelized single-view coders. Each of them uses temporal prediction structures, where a sequence of successive pictures is coded as intra (I), predictive (P), or bi-predictive (B) pictures. For I pictures, the content is only predicted referencing the I picture itself, while P and B picture content is also predicted referencing other pictures. One approach for further improving coding efficiency is the use of hierarchical B pictures [42], where a B picture hierarchy is created by a dyadic cascade of B pictures that are references for other B pictures.

For MVC, the single-view concepts are extended, such that a current picture in the coding process can have temporal as well as interview reference pictures for prediction [29]. For an example of five linearly arranged cameras, the MVC coding structure with a GOP size of eight is shown in Fig. 2.

This coding structure illustrates how the advantages of hierarchical B pictures are combined with interview prediction, without any changes regarding the temporal prediction structure. For the base view (Cam 1 in Fig. 2), the prediction structure is identical to single-view coding and for the remaining views, interview reference pictures are additionally used for prediction (red arrows). Another advantage of MVC is its backward compatibility, as the base view (Cam 1) is decodable by a legacy single-view H.264/AVC decoder. MVC provides up to 40% bit rate reduction for multiview data in comparison to single-view AVC coding [29] In addition, MVC provides a good subjective quality for stereoscopic and 3-D perception, i.e., the coding artifacts do not significantly disturb the perceived scene depth impression.

However, the bit rate resulting from MVC is linearly proportional to the number of display views N. Fig. 3

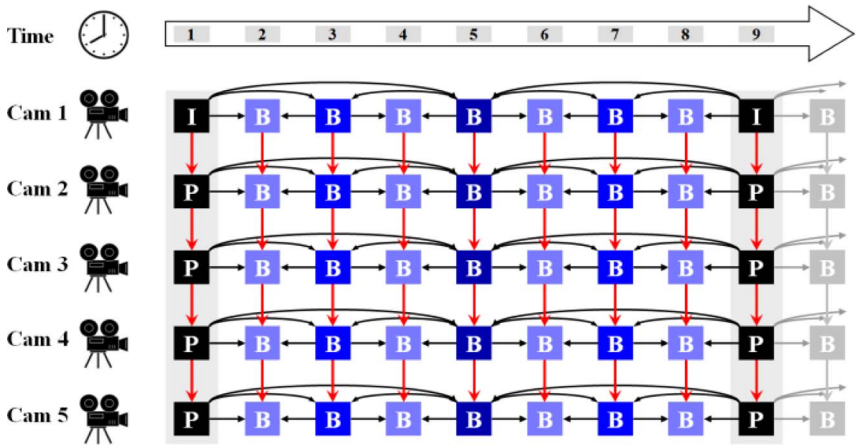

Fig. 2. *Example coding structure in MVC for linear five camera setup and GOP size of eight pictures. The red arrows indicate interview prediction.*

shows the result of an experiment on the relation between bit rate and camera density for MVC.

For this purpose, an array of 16 linearly arranged camera views from the *Rena* sequence was used. The cameras were tightly lined up with the smallest achievable baseline of 5 cm (= camera diameter). To obtain a regular camera distance refinement, a subset of nine adjacent cameras was selected. In order to minimize irregular influences of individual cameras, the experiments were carried out and averaged for all possible nine out of 16 camera subsets. Fig. 3 shows the average percentage MVC coding results using interview prediction. Here, no temporal prediction was used in order to avoid the superposition of temporal and interview coding effects, as discussed in the following. The bit rates were obtained for different quantization parameter (QP) values. The QP controls the fidelity of the coded video signal and its value is inversely proportional to fidelity. First, all nine cameras were coded. Then, every second camera was left out and MVC was applied to the five remaining camera views. Again, every second camera was omitted, such that coding was applied to the remaining three cameras and two cameras in the next step, respectively. As a reference, single-view coding without interview prediction was carried out, as shown by the dotted line in Fig. 3. Single-view coding of one view corresponds to 100% of the bit rate and consequently nine views require 900%, if coded without interview prediction. Using MVC with different QP values results in a reduction of the bit rates: thus, for nine camera views only 650% single-view bit rate is required for higher reconstruction quality (QP24) and even below 300% for low reconstruction quality (QP42). Although a significant coding gain is achieved by interview prediction, the MVC curves in Fig. 3 still show a linear increase with the number of views, although only interview coding effects were investigated. With common MVC coding conditions, as shown in the structure in Fig. 2, only up to 30% of macroblocks are predicted from the interview reference picture, as shown in [29]. Thus, the MVC curves in Fig. 3 will even be closer to the simulcast reference line. As an example, the highest bit rate reduction is 40% for low bit rates, as reported in [29]. Thus, the lowest (QP42) curve in Fig. 3 would reach 540% at nine cameras. This indicates that a reasonable data rate for nine or even 50 views cannot be realized with this concept and that the bit rate is proportional to the number of views.

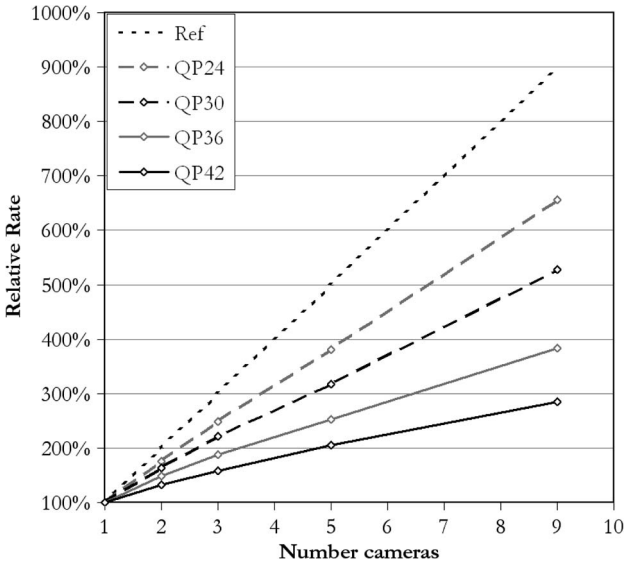

Fig. 3. *Results of coding experiments on camera density in linear camera array in terms of average rate relative to one camera rate.*

C. Color-Only View Extraction

As shown above, the required high number of views cannot be efficiently transmitted using MVC. Therefore, in

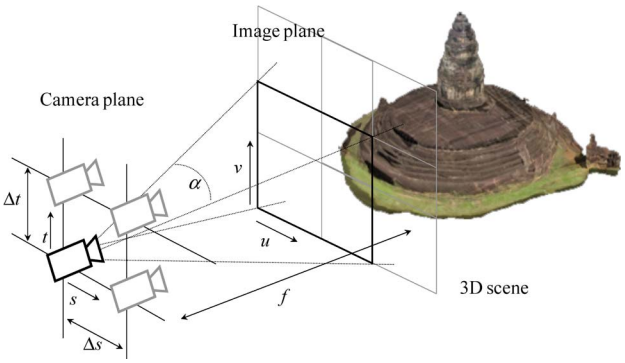

Fig. 4. *Plenoptic sampling of continuous 3-D scene by image and camera plane with horizontal and vertical camera distances $\triangle s$ and $\triangle t$, respectively.*

practical systems, the number of used input views is limited to only a few, e.g., two or three, as these can still be coded efficiently with MVC.

Because only a few original cameras are available in this case, additional views have to be synthesized from them. This leads to the question of how many cameras are required to allow error-free synthesis of views at arbitrary positions and how dense these cameras have to be spaced. To answer these questions, the original 3-D scene can be considered as a continuous signal, which is recorded by a limited number of cameras at discrete positions. In this classical sampling problem, the original continuous 3-D signal is sampled in two domains. First, the scene is sampled by the discrete sample array of horizontal and vertical sensor elements in each camera. This refers to the classical spatial sampling theory, known from 2-D images and videos. Second, the scene is also sampled by the discrete camera positions. This problem was investigated by Chai *et al.* [4] and called "plenoptic sampling." Here, the sampling by the camera sensor and camera positions is described by discrete coordinates in the image plane (u, v), as well as discrete coordinates in the camera plane (s, t), respectively, as shown in Fig. 4.

For solving the multicamera sampling problem, a continuous light field $l(u, v, s, t)$ is defined in [4], which represents all light rays of the 3-D scene that cross the image plane at (u, v) as well as the camera plane at (s, t). Next, the geometric relationship between the cameras in the array is exploited, in order to reduce the problem to one camera plane (see black border plane in Fig. 4).

For simplicity, a parallel camera setting is assumed, such that the light field can be transformed to the base camera at $(s, t) = (0, 0)$ as follows:

$$l(u, v, s, t) = l\left(u - \frac{fs}{z_0}, v - \frac{ft}{z_0}, 0, 0\right). \tag{1}$$

In (1), f represents the camera focal length and z_0 a constant scene depth value. According to classical sampling theory, sampling of a signal at discrete positions leads to repetitions of the frequency spectrum, e.g., if the spatial distance between two samples is Δs, its associated distance between two spectrum repetitions in frequency domain is $2\pi/\Delta s$. Assuming a highest frequency $\Omega_{\max} = 2\pi F_{\max}$ for that signal, nonzero frequency components exist in the interval $[-\Omega_{\max}, \Omega_{\max}]$ and adjacent spectrum repetitions have to be separated at least by $2F_{\max}$ times in order not to overlap. This is important for perfect and alias-free signal reconstruction and the classical sampling condition therefore gives $2\pi/\Delta s \geq 2\Omega_{\max}$. For the light field in (1), sampling occurs in all four coordinates u, v, s, and t. However, the sampling evaluation can be separated into the horizontal components u and s, and the vertical components v and t. As we consider horizontally linear camera arrays, the sampling condition can be reduced to the horizontal light field components, as shown in [4]

$$\frac{2\pi}{\Delta s} \geq f\Omega_{u,\max}\left|\frac{1}{z_{\min}} - \frac{1}{z_{\max}}\right|. \tag{2}$$

Here, Δs is the camera distance, f the camera focal length, $\Omega_{u.\max}$ the maximum horizontal frequency of an image, and

$$d_R = \left|\frac{1}{z_{\min}} - \frac{1}{z_{\max}}\right| \tag{3}$$

is the depth range with minimum and maximum depth value of the recorded scene. Note that $z_{\min}$ and $z_{\max}$ are positive values larger than f, as also shown in Fig. 8. Equation (2) gives the sampling condition for linear camera settings for alias-free continuous light field reconstruction from sparse color data. It especially states that the required distance between neighboring cameras Δs is inversely proportional to the full depth range of the recorded 3-D scene. The sampling condition in (2) was derived for ideal sensors, assuming full coverage of the entire 3-D scene. Real cameras, however, only record a portion of a 3-D scene, limited by their aperture angle α as shown in Figs. 4 and 8. Thus, the overlapping area of commonly recorded scene content in neighboring parallel cameras, for which the sampling condition holds, becomes smaller with increasing camera distance Δs and decreasing $z_{\min}$. Based on the aperture angle α, the maximum value of Δs is limited by the minimum scene overlap: $\Delta s < 2z_{\min} \cdot \tan(\alpha/2)$.

If a scene has a large depth range d_R, a small camera distance is necessary for alias-free continuous light field reconstruction, i.e., view synthesis at any intermediate

Fig. 5. ***Intermediate view synthesis from color-only camera data with alias (left) and from color with 8-b depth values without alias (right) for the Ballet set.***

position. Note that in (2) the highest horizontal frequency $\Omega_{u.\max}$ as well as the focal length f are usually fixed by the number of samples of a given camera sensor. If the camera distance is too large, alias in the form of double images occurs, as shown in Fig. 5 on the left-hand side.

Therefore, a good quality view synthesis from two or three camera views is not possible, if only color information is available. It will be shown in Section III that additional scene geometry information, e.g., in the form of per-sample depth data, has to be provided for high-quality view synthesis.

III. 3DV USING DEPTH MAPS

As concluded in Section II, the provision of a large number of views for multiview displays is not efficient with video data only. The efficiency can be drastically increased using scene geometry information like a depth map. Such a transmission system for 3DV using depth maps is shown in Fig. 6. It is assumed that a few cameras, e.g., two or three, are used. The 3DV encoder generates the bit stream, which can be decoded at the receiver.

The 3DV bit stream contains color and depth data corresponding only to a fixed number of views and with that the overall bit rate is limited. Given these data, a high-quality view synthesis can generate any number N of views for different displays within a given range across the transmitted views. It is assumed that the data are structured in a way such that an arbitrary number of views can be generated at the receiver. Thus, any stereo or multiview display can therefore be supported by the decoded result of the 3DV bit stream as the required number and spatial position of views can be synthesized individually for each display.

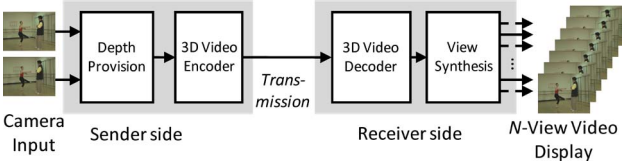

Fig. 6. ***3DV system based on depth-enhanced multiview video.***

One of the most challenging properties of this 3DV system is the interdependency between different parts of the processing chain, namely depth provision, coding, and view synthesis. These three parts influence each other, as the quality of depth maps influences coding and view synthesis, coding artifacts influence view synthesis, and a good and robust view synthesis in return can compensate depth and coding errors. The depth data are provided at the sender side for the uncompressed input video sequences as per-sample depth values. Although depth data could also be generated at the receiver side, i.e., after the MVC decoding in Fig. 1, we believe that depth data will be provided at the sender side and transmitted within the 3DV format. The advantage is that producers of 3DV content will have control over the resulting multiview display output, which therefore appears similar across different display types and a certain transmission quality can be guaranteed for comparable 3-D viewing quality. Furthermore, different methods for determining depth values can be used at the sender side, e.g., to incorporate additional data from content production formats, which are not part of the transmitted 3DV format. Such depth provisioning methods are discussed in Section III-B. At the receiver side, intermediate views are synthesized via DIBR methods, as described in Section III-C. This system architecture can be combined with coding algorithms for lossless, lossy, or even no compression.

Please note that the input to this 3DV system are captured and rectified video sequences from a few cameras and the output of the system is an arbitrary number N of view sequences. Thus, the input or capturing format is decoupled from the output format and the decoded 3DV bit stream can be used by any 2-D, stereoscopic 3-D, and multiview 3-D display.

One basic assumption to be made for this 3DV coding system is that it delivers uncoded synthesized views, which

can be used as a reference for quality comparison. This is important, as the coding of depth maps needs to consider the resulting quality for the synthesized views [21] and a sufficient picture quality has to be guaranteed for any view that might be generated, since the user might be looking exclusively at synthesized views.

A. Scene Depth Representation

As shown in Section II-C, dense view synthesis only from sparse video data leads to aliasing in the synthesized image. This is shown on the left-hand side of Fig. 5, where a double image is synthesized. However, if additional scene geometry data are available, e.g., in form of a depth value for each color sample in the camera plane, the sampling condition in (2) changes, as derived in [4]. These depth values are quantized into a number of different values. Then, the depth range d_R in (2) is split into a number of S quantization intervals $d_{R,i}$, $i = 1 \ldots S$, with $d_{R,i} = d_R/S$. Thus, an S-times larger camera distance Δs in (2) is allowed for correct view synthesis, as illustrated by the good reconstruction result on the right-hand side of Fig. 5 with the same camera distance as for the nondepth case. For example, if a depth map uses 8 b/sample, the depth range in (2) is split into 256 small subranges. Thus, the camera distance Δs for depth-enhanced 3-D formats with 8-b depth data can be 256 times larger for alias-free reconstruction than the camera distance for formats that contain only video data.

Such a depth-enhanced format for two different views is shown in Fig. 7 with color and per-sample depth information. Note that the maximum value for Δs is again limited for real-world cameras by their aperture angle, as discussed in Section II-C.

The depth data are usually stored as inverted real-world depth data $I_d(z)$, according to

$$I_d(z) = \text{round}\left[255 \cdot \left(\frac{1}{z} - \frac{1}{z_{\max}}\right) \Big/ \left(\frac{1}{z_{\min}} - \frac{1}{z_{\max}}\right)\right]. \quad (4)$$

Here, a representation with 8 b/sample and values between 0 and 255 is assumed. This method of depth storage has the following advantages: since depth values are inverted, a high depth resolution of nearby objects is achieved, while farther objects only receive coarse depth resolution, as shown in Fig. 8. This also aligns with the human perception of stereopsis [47], where a depth impression is derived from the shift between left and right eye view. The stored depth values are quantized similarly to these shift or disparity values. However, the inverse quantized depth values are not identical to disparity values, since disparity values depend on the camera distance or baseline in contrast to depth values. This difference is very important, as depth values are therefore also independent from neighboring cameras, as well as different camera sensor and image resolutions. Consequently, the stored depth representation in (4) combines the advantages of inverse quantization for more natural depth representation with the independency from camera baselines and image resolutions.

Fig. 7. *Example for depth enhanced format: two view plus depth format for the Ballet set.*

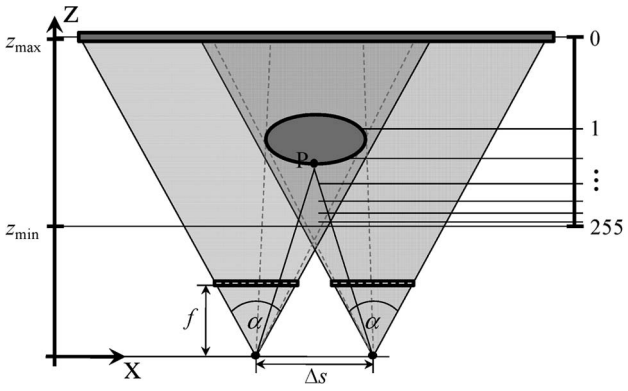

Fig. 8. *Inverse depth sampling with 8-b resolution between z_{min} and z_{max}.*

For very high image resolutions and very precise view synthesis results, more than 8 b/sample might be required for the depth signal, in order to provide disparity values with sufficient accuracy.

For retrieving the depth values z from the depth maps, the following is applied, which is typically used in synthesis scenarios:

$$z = 1/\left[\frac{I_d(z)}{255}\cdot\left(\frac{1}{z_{\min}}-\frac{1}{z_{\max}}\right)+\frac{1}{z_{\max}}\right]. \quad (5)$$

For this, the original minimum and maximum depth values z_{min} and z_{max} are required, which have to be signaled with the 3DV format for a correct geometric displacement in synthesized intermediate views.

B. Depth Provision

The provision of high-quality depth data is crucial for 3DV applications. The depth information can be obtained in different ways. One approach is to estimate the depth data based on the acquired pictures, as intensively investigated in the research community [41]. Usually, depth estimation algorithms attempt to match corresponding signal components in two or more original cameras, using a matching function [44] with different area support and size [3]. They apply a matching criterion, e.g., sum of absolute differences, cross correlation, etc., and try to optimize the estimation, based on different strategies, such as graph cuts [23], belief propagation [11], plane sweeping [7], or combined approaches [1], [22]. Recently, depth estimation has been studied with special emphasis for multiview video content and temporal consistency in order to provide depth data for 3DV applications [26], [31], [45].

Usually, depth estimation algorithms generate disparity values d in the matching process, which relate to real-world depth values z as follows:

$$d = \frac{f\cdot\Delta s}{z}. \quad (6)$$

Although depth estimation algorithms have been improved considerably in recent years, they can still be erroneous in some cases due to mismatches, especially for partially occluded image and video content that is only visible in one view.

Another method for depth provision is the use of special sensors, like time-of-flight cameras, which record low-resolution depth maps [25]. Here, postprocessing is required for interpolating depth for each video sample [6]. Such sensors currently lack accuracy for larger distances and have to be placed at slightly different positions than the video camera. It is therefore envisioned that in the future a recording device would capture high precision depth together with each color sample directly in the sensor.

For synthetic sequences, such as computer-generated scene content and animated films, scene geometry information is available, e.g., in the form of wireframe models [15] or 3-D point coordinates [49]. Thus, depth data can be extracted as the distance between a selected camera position and the given scene geometry information.

With the provided depth data, a coding format can be specified, using per-sample depth values for each input video view, as shown in Fig. 7 for a depth-enhanced two-view format. The video signal from two different perspectives is required for partially occluded data behind foreground objects in one original view, which becomes visible in an intermediate view and can be filled with the visible data from the other view. For stereo displays with the correct baseline, the two video views can be used directly without generating additional views. For correct intermediate view synthesis, each view requires its own depth data, especially for the partially occluded parts, which are only visible in one view.

Therefore, we believe that the data format should contain at least two video and two associated depth signals from different viewpoints, in order to generate the required range of N views with good quality for a multiview display, as shown in Section III-C.

C. Depth-Image-Based Rendering

With the provision of per-sample depth data, any number of views within a given range can be synthesized from a few input views. Based on the principles of projective geometry [10], [13], arbitrary intermediate views are generated via 3-D projection or 2-D warping from original camera views. This is typically referenced as DIBR [19], [40]. For the presented 3DV solution, the camera views are rectified in a preprocessing step. Thus, the complex

process of general DIBR can be simplified to horizontal sample shifting from original into newly rendered views. An example for a fast view generation method with line-wise processing and sample shift lookup table can be found in [30].

The sample shifts are obtained by calculating disparity values d from the stored inversely quantized depth values $I_d(z)$ by combining (4) and (6)

$$d = f \cdot \Delta s \cdot \frac{I_d(z)}{255} \cdot \left(\frac{1}{z_{\min}} - \frac{1}{z_{\max}} \right) + \frac{1}{z_{\max}}. \quad (7)$$

Here, the focal length f and camera baseline Δs have to be known. If Δs is given as the spatial distance between two original cameras, d represents the disparity between these cameras and has to be scaled for any intermediate view. As an example, consider two original cameras, *camera 0* and *camera 1*, with corresponding color samples c_0 and c_1 at positions (u_0, v_0) and (u_1, v_1), respectively. From this, an intermediate view between both cameras is to be synthesized with corresponding sample c_κ at position (u_κ, v_κ). Here, $\kappa \in [0 \ldots 1]$ represents the intermediate position parameter, which specifies the intermediate position between *camera 0* and *camera 1*. For instance, a value of $\kappa = 0.5$ specifies the middle view between both original cameras. As derived in [33], the view synthesis for $c_\kappa(u_\kappa, v_\kappa)$ can be described by

$$c_\kappa(u_\kappa, v_\kappa) = (1 - \kappa) \cdot c_0(u_0, v_0) + \kappa \cdot c_1(u_1, v_1). \quad (8)$$

Note that (8) describes a general interpolation case, where both original samples show the same content, visible in both views. In cases of partial occlusions, where content is only visible in one view, (8) is adapted, e.g., $c_\kappa(u_\kappa, v_\kappa) = c_1(u_1, v_1)$, if content is only visible in $c_1(u_1, v_1)$. Also, the color values from both cameras are weighted by the position parameter κ in order to allow smooth transition for a continuum of many synthesized views between *camera 0* and *camera 1*. Via the horizontal κ-scaled disparity values from (6) or (7), the sample positions in the original views can be related to the intermediate position (u_κ, v_κ)

$$c_\kappa(u_\kappa, v_\kappa) = (1 - \kappa) \cdot c_0(u_\kappa + (1 - \kappa) \cdot d, v_\kappa) + \kappa \cdot c_1(u_\kappa - \kappa \cdot d, v_\kappa). \quad (9)$$

In (9), d represents the disparity value from (u_0, v_0) to (u_1, v_1), such that $u_0 + d = u_1$. Since the view synthesis is applied after decoding, the color values c_0 and c_1 as well as the disparity value d can contain coding errors. While coding errors in color samples lead to slight color changes, erroneous disparity values cause wrong sample shifts [34]. This is especially critical at coinciding depth and color edges, where completely different color values are used for interpolation. Consequently, strong sample scattering and color bleeding can thus be present in the synthesized view. This requires a special treatment of such image areas in 3DV, e.g., via reliability-based processing, as discussed in Section III-E.

D. Depth Signal Coding

One major task in 3DV coding is the development of efficient coding methods for depth data. For the video data contained in the 3DV format, typically MVC is used, as it is optimized for this type of data. Depth data show different characteristics that are to be considered for coding these signals. Depth maps have large homogeneous regions within scene objects and abrupt signal changes at object boundaries with different depth values, as shown at the bottom of Fig. 7. Here, mostly low frequencies as well as very high frequencies are present in the depth signal. In contrast to the video signal, especially the high frequencies should not be omitted, in order to guarantee a good visual perception of intermediate views. Reconstruction errors in the video data lead to image blurring. However, reconstruction errors in the depth signal lead to wrong sample displacements in synthesized intermediate views. For video data, a certain reconstruction quality can be measured directly by comparing compressed and uncompressed signals. For depth data, a reconstruction quality can only be measured indirectly by analyzing the quality of the color information for the synthesized intermediate views.

Therefore, one of the major requirements for depth coding is the preservation of important edge information. In literature, approaches for depth signal coding have been reported. One approach is depth down-sampling before classical MVC encoding and special up-sampling after decoding to recover some of the original depth edge information [37]. A coding approach for depth data with MVC, which is based on rate-distortion optimization for an intermediate view, is shown in [21]. For better edge preservation in depth compression, also wavelet coding was applied [9], [27]. A computer graphics-based approach was taken in [20], where depth maps were converted into meshes and coded with mesh-based compression methods.

Another example for edge-aware coding is platelet coding, which is described in detail in [28]. As for any other lossy coding technique, coding artifacts occur for high compression ratios with platelets as well. However, their characteristics are different from classical coding, as shown in Fig. 9. First, a sample of the original depth map is shown in Fig. 9(a). Next, differently coded depth maps at the same bit rate are shown. With H.264/AVC, as shown in Fig. 9(b), compression artifacts appear as edge smoothing. Here, intra-only coding was used as a first direct comparison to the platelet-based coding approach. In Fig. 9(c), fully optimized MVC with temporal and interview

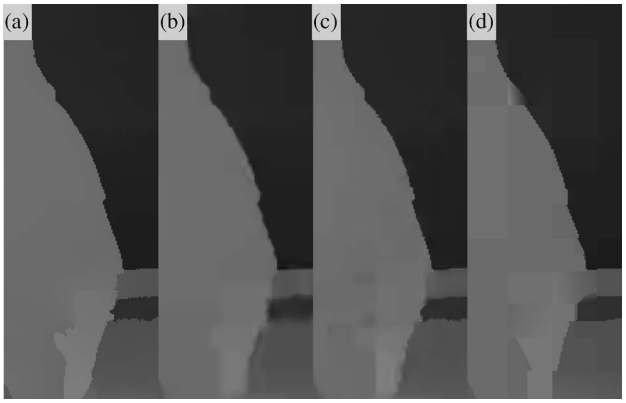

Fig. 9. *Impact of coding artifacts on depth maps for Ballet sequence: (a) original uncoded depth, (b) H.264/AVC (intra-only) coded depth, (c) MVC (fully optimized) coded depth, and (d) platelet coded depth at the same bit rate.*

prediction was used. Here, edge smoothing artifacts are also present. However, due to higher compression efficiency, a better reconstruction quality at the same bit rate is achieved for MVC in comparison to H.264/AVC simulcasting. This is especially visible at sharp edges. In contrast to block-based coding methods, the artifacts for platelet coding in Fig. 9(d) are coarser edge approximation. This becomes especially visible for the foreground/background edge towards the bottom of Fig. 9. Here, the edge preservation of the platelet coding is much closer to the original depth edge than for the other two coding methods. Thus, the height and sharpness of the depth edge is much better preserved in platelet coding.

The final quality of a depth coding method has to be evaluated for the synthesized views. This is shown in Fig. 10, where the synthesized views are presented with respect to the differently coded depth maps at the same bit rates from Fig. 9 and uncoded video data. The synthesized result from uncoded depth and video data is shown in Fig. 10(a). Note that even in this case, corona artifacts are visible, which can be removed by advanced synthesis algorithms, as discussed in Section III-E. The classical video coding approaches H.264/AVC and MVC in Fig. 10(b) and (c) show color displacement artifacts around foreground objects due to depth edge smoothing. In contrast, the foreground boundaries are much better preserved by the platelet coding approach, as shown in Fig. 10(d). Only in cases with rather complex depth edge structures, some color displacements occur. Again, a good view synthesis can help to further reduce these errors.

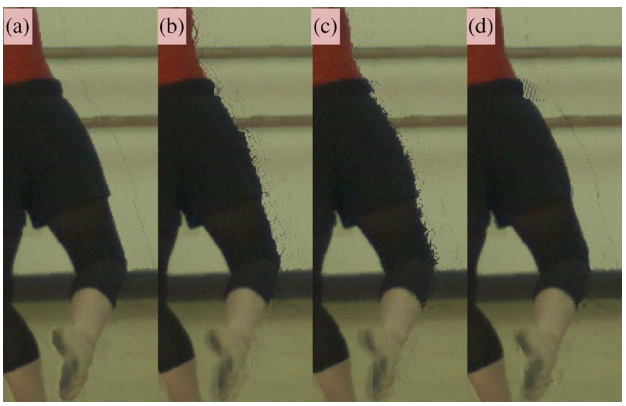

Fig. 10. *Impact of depth coding artifacts on view synthesis rendering for Ballet sequence with uncoded color data: (a) original uncoded depth, (b) H.264/AVC (intra) coded depth, (c) MVC (fully optimized) coded depth, and (d) platelet coded depth for the associated depth map from Fig. 9.*

Further ongoing research on advanced coding approaches is discussed in Section III-F.

E. Advanced View Synthesis Methods

For high-quality view synthesis, advanced processing has to be applied in addition to the sample-wise blending in (9). In the literature, a number of view synthesis improvements have been reported, which focus on the following topics. For hole filling, inpainting methods are used as described in [5], [36], and [39]. Here, surrounding texture information and statistics are analyzed and used to fill missing information in synthesized views. Postfiltering is applied in order to remove wrongly projected outliers and to provide a better overall impression [32]. In [33], [50], and [52], the use of reliability information for improved synthesis results is described.

For any view synthesis, foreground/background object boundaries are among the most challenging problems. A simple projection from original views can cause corona artifacts, as shown in Fig. 11(a) and (c). The reasons for such artifacts are certain effects, like incorrect depth values and edge samples, which contain a combination of foreground and background color samples. Also, object edges may be fuzzy and may contain semitransparent content. Therefore, special treatment in such areas has to be applied. In advanced synthesis methods, a reliability-based approach is taken with one [52] or two [33] boundary layers. Since areas along depth discontinuities in 3DV are known to produce visual artifacts in the projection process, they are processed separately.

The video data of original views are classified as "unreliable areas" along depth edges, while the remaining areas are labeled as being "reliable areas." While the method from [52] only uses one depth edge or boundary layer, the method from [33] uses two layers for foreground and background boundary data with different projection rules. The reliable areas are projected or shifted into the intermediate view first. Then, the unreliable boundary areas are split into foreground and background data. Here, foreground areas are projected next and merged with the reliable data. Afterwards, the background data are projected and also merged. The important difference between

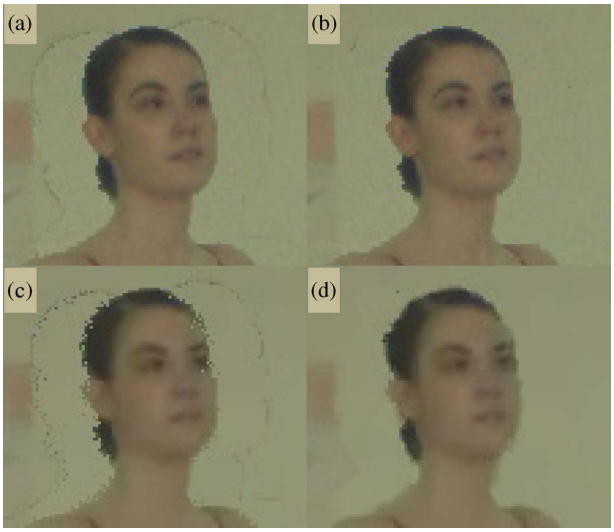

Fig. 11. ***Comparison of intermediate view quality: (a) and (c) with simple view synthesis and (b) and (d) with reliability-based view synthesis. (a) and (b) using uncompressed data and (c) and (d) using compressed data from the Ballet sequence.***

foreground and background handling is the merging process. The foreground data are merged with the reliable data in a frontmost sample approach, where the color sample with the smallest depth value is taken and with that most of the important information of the foreground boundary layer is preserved. In contrast, background information is only used to fill remaining uncovered areas. Finally, different view enhancement algorithms are applied, including outlier removal, hole filling, and natural edge smoothing. A more detailed description can be found in [33].

The results of the described advanced view synthesis are shown in Fig. 11. Here, intermediate views are synthesized from uncompressed [Fig. 11(a) and (b)] as well as compressed data [Fig. 11(c) and (d)]. To show the capability of the view synthesis, reliability-based layer projection was switched off for the results in Fig. 11(a) and (c). The reduction of corona artifacts is visible for the uncompressed as well as compressed case in Fig. 11. Thus, an advanced high-quality view synthesis can compensate some errors from depth estimation as well as coding artifacts.

F. Perspectives

For 3DV systems, depth enhanced formats have been introduced and coding methods for video and depth data have been described. In addition to coding methods that are applied separately to these signals, a number of joint coding approaches have been published. Kim *et al.* [21] investigated the quality of synthesized views for video coding optimization. View synthesis is also incorporated into the rate distortion optimization for MVC in [51]. First methods for sharing data, like motion vectors for video and depth, are presented in [35]. Other approaches introduce scalability mechanisms, which encode one reference view as base layer and warped residual data of adjacent views as enhancement layers [43]. It still needs to be shown to what extent joint approaches will offer better compression efficiency over coding approaches with separate video and depth coding. However, according to the vision of standardization bodies [17], an optimized 3DV coding design needs to address the following challenges:

- break the linear dependency of coding bit rate from the number of target views;
- provide a generic format, e.g., video and per-sample depth data of two to three original views, for support of different 3-D capturing and production environments, as well as different multiview displays;
- optimize coding approaches for consideration of depth statistics;
- consider new quality evaluation methods for intermediate views;
- provide high-quality view synthesis for continuous viewing range and optimize bit rate allocation for video and depth accordingly.

In addition to that, the migration from existing stereo services needs to be considered. Therefore, new 3DV services will also include the extraction and generation of high-quality stereo video, e.g., by signaling one bit stream portion for video data and a second portion for the depth enhancement. Based on the new coding solution for 3DV, existing stereo systems might either extract the video data portion as is or generate high-quality stereo using the improved encoder/decoder technology.

IV. CONCLUSION

This paper provides an overview on 3DV systems using depth data. First, an introduction of currently available stereo video systems is given. These systems are based on stereo formats, where the two views are coded using the stereo high profile of H.264/AVC. The extension of the stereo high profile to many views is called MVC. When applying MVC to many views, it can be shown that the bit rate is linearly proportional to the number of views.

As 3-D displays present a strong factor for enabling stereo as well as depth-enhanced 3DV solutions, we compared stereo with multiview display systems and analyzed their specific properties. It is clear that natural 3-D viewing with "look around" effects and viewing without glasses is only supported by autostereoscopic displays. Currently, these displays have restrictions on view resolution and depth perception, which should be overcome, as soon as ultrahigh resolution display technology becomes available. For this, we estimate that the number of views in upcoming autostereoscopic displays may be 50 or more. However, the transmission of 50 or more views would be very inefficient using MVC. This problem is compounded by

the fact that the improvements of autostereoscopic 3-D displays are likely to be driven by increasing the number of views. A scenario with bit rate being linearly proportional to the number of views is therefore not feasible. Trying to solve this problem by restricting the number of transmitted camera views to only a few and generating the required views from the decoded views at the receiver appears to be an unlikely scenario for various reasons, including associated complexity, estimation problems, and lack of control over the output quality.

It is rather proposed to provide means for generation of the required number of views at the encoder and to transmit these data to the decoder for simple and straightforward view synthesis. The format that can be used consists of two or three camera views together with associated depth data. The depth data are provided at the sender side for control over the output views for any display and can be estimated from the multiview video signal in the production process or rendered from synthetic objects. The determination of the numbers of necessary views is provided by plenoptic sampling theory, which can be used to show that only a few input views with depth maps are required for alias-free synthesis of any number of intermediate views. A major advantage of this solution is that it decouples the transmission format from the display format. It is no longer necessary to exactly transmit the various displayed views, and the same decoded bit stream can be provided to any 3-D display, independent of the number and spatial positions of views. The depth maps included in such a 3DV representation are then employed to synthesize the displayed views via DIBR. Thus, for high-quality rendered views, advanced view synthesis methods are discussed.

For the transmission of such 3DV formats, efficient compression is required. Regarding the video signal, MVC is currently used efficiently, as the format only contains two or three input views. For efficient coding of depth data, we show that current methods, like H.264/AVC or MVC, are not recommended, as they cause coding errors, which lead to visible distortions in synthesized views. In particular significant depth edges need to be preserved, as video sample displacement artifacts especially occur here. Therefore, coding methods for depth signals are introduced with platelet-based depth coding being presented in more detail. The results for platelet coding show that it is capable of reducing the color sample displacement artifacts significantly. Finally, new approaches for 3DV coding are discussed, which concentrate on joint color and depth coding.

In conclusion, we expect that future research and development in depth enhanced 3DV will lead to efficient and generic solutions for new services, such as high-quality 3DV on various stereoscopic and autostereoscopic displays for home entertainment and mobile applications. ■

Acknowledgment

The authors would like to thank the Interactive Visual Media Group of Microsoft Research for providing the *Ballet* data set and Nagoya University for providing the *Rena* data set.

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.*, vol. 14, *Special Issue on Immersive Telecommunications*, 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] M. Bleyer and M. Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints," *ISPRS J. Photogrammetry Remote Sens.*, vol. 59, no. 3, pp. 128–150, 2005.

[4] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proc. SIGGRAPH*, pp. 307–318, 2000.

[5] C.-M. Cheng, S.-J. Lin, S.-H. Lai, and J.-C. Yang, "Improved novel view synthesis from depth image with large baseline," in *Proc. Int. Conf. Pattern Recognit.*, Tampa, FL, Dec. 2008, DOI: 10.1109/ICPR.2008.4761649.

[6] J. Choi, D. Min, B. Ham, and K. Sohn, "Spatial and temporal up-conversion technique for depth video," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 741–744.

[7] C. Cigla, X. Zabulis, and A. A. Alatan, "Region-based dense depth extraction from multi-view video," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, Sep. 2007, pp. 213–216.

[8] 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, Jan. 2009, article ID 786015.

[9] 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.

[10] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press, 1993.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[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, 2000.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[14] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 1: May 2003, Version 2: May 2004, Version 3: Mar. 2005 (including FREXt extension), Version 4: Sep. 2005, Version 5 and Version 6: Jun. 2006, Version 7: Apr. 2007, Version 8: Jul. 2007 (including SVC extension), Version 9: Jul. 2009 (including MVC extension).

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

[16] ISO/IEC JTC1/SC29/WG11, "Text of ISO/IEC 14496-10:200X/ FDAM 1 multiview video coding," Doc. N9978, Hannover, Germany, Jul. 2008.

[17] *Vision on 3D Video*, ISO/IEC JTC1/SC29/WG11, Feb. 2009, Doc. N10357, Lausanne, CH.

[18] B. Julesz, "Binocular depth perception of computer-generated images," *Bell Syst. Tech. J.*, vol. 39, no. 5, pp. 1125–1163, 1960.

[19] 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 on 3DTV*, vol. 22, no. 2, pp. 217–234, Feb. 2007.

[20] 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, Sep. 2007, pp. V117–V120.

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

[22] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568–1583, Oct. 2006.

[23] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in

Proc. Eur. Conf. Comput. Vis., vol. 3, pp. 82–96, May 2002.

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

[25] E.-K. Lee, Y.-K. Jung, and Y.-S. Ho, "Three-dimensional video generation using foreground separation and disocclusion detection," in *Proc. IEEE 3DTV Conf.*, Tampere, Finland, Jun. 2010, DOI: 10.1109/3DTV.2010.5506602.

[26] S.-B. Lee and Y.-S. Ho, "View consistent multiview depth estimation for three-dimensional video generation," in *Proc. IEEE 3DTV Conf.*, Tampere, Finland, Jun. 2010, DOI: 10.1109/3DTV.2010.5506320.

[27] M. Maitre and M. N. Do, "Shape-adaptive wavelet encoding of depth maps," in *Proc. Picture Coding Symp.*, Chicago, IL, May 2009, DOI: 10.1109/PCS.2009.5167381.

[28] 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, no. 1+2, pp. 73–88, Jan. 2009.

[29] P. Merkle, A. Smolic, K. Mueller, 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.

[30] P. Merkle, Y. Wang, K. Müller, A. Smolic, and T. Wiegand, "Video plus depth compression for mobile 3D services," in *Proc. IEEE 3DTV Conf.*, Potsdam, Germany, May 2009, DOI: 10.1109/3DTV.2009.5069650.

[31] D. Min, S. Yea, and A. Vetro, "Temporally consistent stereo matching using coherence function," in *Proc. IEEE 3DTV Conf.*, Tampere, Finland, Jun. 2010, DOI: 10.1109/3DTV.2010.5506215.

[32] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," *Signal Process., Image Commun.*, vol. 24, *Special Issue on Advances in Three-Dimensional Television and Video*, no. 1–2, pp. 65–72, Jan. 2009.

[33] 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.*, vol. 2008, *Special Issue on 3D Image and Video Processing*, 2008, article ID 438148, DOI: 10.1155/2008/438148.

[34] K. Müller, A. Smolic, K. Dix, P. Merkle, and T. Wiegand, "Coding and intermediate view synthesis of multi-view video plus depth," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 741–744.

[35] H. Oh and Y.-S. Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," *Advances in Image and Video Technology*, vol. 4319. Berlin, Germany: Springer-Verlag, 2006.

[36] K.-J. Oh, S. Yea, and Y.-S. Ho, "Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-D video," in *Proc. Picture Coding Symp.*, Chicago, IL, May 2009, DOI: 10.1109/PCS.2009.5167450.

[37] 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.

[38] H. M. Ozaktas and L. Onural, Eds., *Three-Dimensional Television: Capture, Transmission, Display*. Heidelberg, Germany: Springer, 2007, ISBN: 78-3-540-72531-2.

[39] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting under constrained camera motion," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 545–553, Feb. 2007.

[40] 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. Visual. Transm.*, Jun. 2002, pp. 313–319.

[41] 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.

[42] 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.

[43] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multi-view video coding using 3-D warping with depth map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1485–1495, Nov. 2007.

[44] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, vol. 2, pp. 16–29.

[45] M. Tanimoto, T. Fujii, and K. Suzuki, "Improvement of depth map estimation and view synthesis," ISO/IEC JTC1/SC29/WG11, M15090, Antalya, Turkey, Jan. 2008.

[46] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/AVC standard," *Proc. IEEE, Special Issue on 3D Media and Displays*.

[47] C. Wheatstone, "Contributions to the physiology of Vision. Part the First. On some remarkable, and hitherto unobserved, phenomena of binocular vision," *Philosoph. Trans. R. Soc. Lond.*, vol. 128, pp. 371–394, 1838.

[48] 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.

[49] S. Würmlin, E. Lamboray, and M. Gross, "3d video fragments: dynamic point samples for real-time free-viewpoint video," *Comput. Graph., Special Issue on Coding, Compression and Streaming Techniques for 3D and Multimedia Data*, pp. 3–14, Elsevier.

[50] L. Yang, T. Yendo, M. P. Tehrani, T. Fujii, and M. Tanimoto, "Error suppression in view synthesis using reliability reasoning for FTV," in *Proc. IEEE 3DTV Conf.*, Tampere, Finland, Jun. 2010, DOI: 10.1109/3DTV.2010.5506260.

[51] S. Yea and A. Vetro, "View synthesis prediction for multiview video coding," *Signal Process., Image Commun.*, vol. 24, no. 1+2, pp. 89–100, Jan. 2009.

[52] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *Proc. ACM SIGGRAPH*, Los Angeles, CA, Aug. 2004, DOI: 10.1145/1015706.1015766.

[53] M. Zwicker, A. Vetro, S. Yea, W. Matusik, H. Pfister, and F. Durand, "Signal processing for multi-view 3D displays: Resampling, antialiasing and compression," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 88–96, Nov. 2007.

ABOUT THE AUTHORS

Karsten Müller (Senior Member, IEEE) received the Dr.-Ing. degree in electrical engineering and the Dipl.-Ing. degree from the Technical University of Berlin, Berlin, Germany, in 2006 and 1997, respectively.

Currently, he is heading the 3D Coding group within the Image Processing Department, Fraunhofer Institute for Telecommunications—Heinrich Hertz Institute, Berlin, Germany. He has been with the Fraunhofer Institute for Telecommunications—Heinrich Hertz Institut, since 1997 and coordinates 3D Video and 3D Coding related international projects. His research interests are in the field of representation, coding and reconstruction of 3-D scenes in free viewpoint video scenarios and coding, immersive and interactive multiview technology, and combined 2-D/3-D similarity analysis. He has been involved in ISO-MPEG standardization activities in 3-D video coding and content description.

Philipp Merkle (Student Member, IEEE) 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, Germany, in 2003 and has been employed as 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 research interests include 3-D video, representation and compression of multiview video plus depth scenes, free viewpoint video, 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.

Dr. 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.

Thomas Wiegand (Fellow, IEEE) received the Dipl.-Ing. degree in electrical engineering from the Technical University of Hamburg, Harburg, Germany, in 1995 and the Dr.-Ing. degree from the University of Erlangen, Nuremberg, Germany, in 2000.

Currently, he is a Professor at the Department of Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany, chairing the Image Communication Laboratory, and is jointly heading the Image Processing Department of the Fraunhofer Institute for Telecommunications—Heinrich Hertz Institute, Berlin, Germany. He joined the Heinrich Hertz Institute in 2000 as the Head of the Image Communication group in the Image Processing Department. His research interests include video processing and coding, multimedia transmission, as well as computer vision and graphics. From 1993 to 1994, he was a Visiting Researcher at Kobe University, Japan. In 1995, he was a Visiting Scholar at the University of California at Santa Barbara. From 1997 to 1998, he was a Visiting Researcher at Stanford University, Stanford, CA, and served as a consultant to 8x8, Inc., Santa Clara, CA. From 2006 to 2008, he was a Consultant to Stream Processors, Inc., Sunnyvale, CA. From 2007 to 2009, he was a Consultant to Skyfire, Inc., Mountain View, CA. Since 2006, he has been a member of the technical advisory board of Vidyo, Inc., Hackensack, NJ. 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.

Dr. Wiegand was appointed the Associated Rapporteur of ITU-T VCEG in October 2000. In December 2001, he was appointed the Associated Rapporteur/Co-Chair of the JVT. In February 2002, he was appointed the Editor of the H.264/MPEG-4 AVC video coding standard and its extensions (FREXt and SVC). From 2005 to 2009, he was Co-Chair of MPEG Video. In 1998, he received the SPIE VCIP Best Student Paper Award. In 2004, he received the Fraunhofer Award and the ITG Award of the German Society for Information Technology. The projects that he co-chaired for development of the H.264/AVC standard have been recognized by the 2008 ATAS Primetime Emmy Engineering Award and a pair of NATAS Technology & 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. He was elected Fellow of the IEEE in 2011 "for his contributions to video coding and its standardization." He was a Guest Editor for the IEEE Transactions on Circuits and Systems for Video Technology for its Special Issue on the H.264/AVC Video Coding Standard in July 2003, its Special Issue on Scalable Video Coding-Standardization and Beyond in September 2007, and its Special Section on the Joint Call for Proposals on High Efficiency Video Coding (HEVC) Standardization. Since January 2006, he has been an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology.