Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV

Christoph Fehn

Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)
Einsteinufer 37, 10587 Berlin, Germany

ABSTRACT

This paper presents details of a system that allows for an evolutionary introduction of depth perception into the existing 2D digital TV framework. The work is part of the European Information Society Technologies (IST) project "Advanced Three-Dimensional Television System Technologies" (ATTEST), an activity, where industries, research centers and universities have joined forces to design a backwards-compatible, flexible and modular broadcast 3D-TV system. At the very heart of the described new concept is the generation and distribution of a novel data representation format, which consists of monoscopic color video and associated per-pixel depth information. From these data, one or more "virtual" views of a real-world scene can be synthesized in real-time at the receiver side (i. e. a 3D-TV set-top box) by means of so-called depth-image-based rendering (DIBR) techniques. This publication will provide: (1) a detailed description of the fundamentals of this new approach on 3D-TV; (2) a comparison with the classical approach of "stereoscopic" video; (3) a short introduction to DIBR techniques in general; (4) the development of a specific DIBR algorithm that can be used for the efficient generation of high-quality "virtual" stereoscopic views; (5) a number of implementation details that are specific to the current state of the development; (6) research on the backwards-compatible compression and transmission of 3D imagery using state-of-the-art MPEG (Moving Pictures Expert Group) tools.

Keywords: ATTEST, 3D-TV, Depth-Image-Based Rendering (DIBR), "Virtual" View Synthesis, Stereoscopy, Coding and Transmission, Human-Factors Evaluations, Moving Pictures Expert Group (MPEG).

1. A NEW APPROACH ON 3D-TV

As early as in the 1920s, John Logie Baird, one of the TV pioneers, dreamed of developing high-quality, three-dimensional (3D) color TV, as only such a system would provide the most natural viewing experience. Today, eighty years later, the first black-and-white television prototypes have evolved into high-definition digital color TV, but the hurdle of 3D still remains to be taken. The reasons for this anything but satisfying status quo are multifaceted but new hope arises from recent advances in a number of key technologies, with the following developments being of particular importance: (a) The introduction and increasing propagation of digital TV in Europe, Asia and the United States; (b) The particulary promising latest achievements in the area of single- and multiview autostereoscopic 3D display technologies; (c) The increased interest in the investigation of the human-factors requirements for high-quality 3D-TV systems.

Building on the foundation of these trends, the ambitious aim of the European IST project ATTEST is to design a novel, backwards-compatible and flexible broadcast 3D-TV system.[1] In contrast to former proposals, which often relied on the basic concept of "stereoscopic" video, i. e. the capturing, transmission and display of two separate video streams – one for the left eye and one for the right eye –, this novel idea is based on a more flexible joint transmission of monoscopic color video and associated per-pixel depth information. From this data representation, one or more "virtual" views of a real-world scene can then be generated in real-time at the receiver side by means of so-called depth-image-based rendering (DIBR) techniques. The modular architecture of the proposed system provides important features, such as backwards-compatibility to today's 2D digital TV, scalability in terms of receiver complexity and adaptability to a wide range of different 2D and 3D displays.[2]

To allow for an easier understanding of the main ideas, the envisioned signal processing and data transmission chain of the ATTEST 3D-TV concept is illustrated in Fig. 1. It consists of four different functional building blocks: 1) 3D content creation; 2) 3D video coding; 3) Transmission; 4) "Virtual" view generation and 3D display.

Christoph.Fehn@hhi.fraunhofer.de; phone: +49 30 31002-611; fax: +49 30 3927200; www.hhi.fraunhofer.de

Stereoscopic Displays and Virtual Reality Systems XI, edited by Andrew J. Woods, John O. Merritt, Stephen A. Benton, Mark T. Bolas, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5291 © 2004 SPIE and IS&T · 0277-786X/04/$15

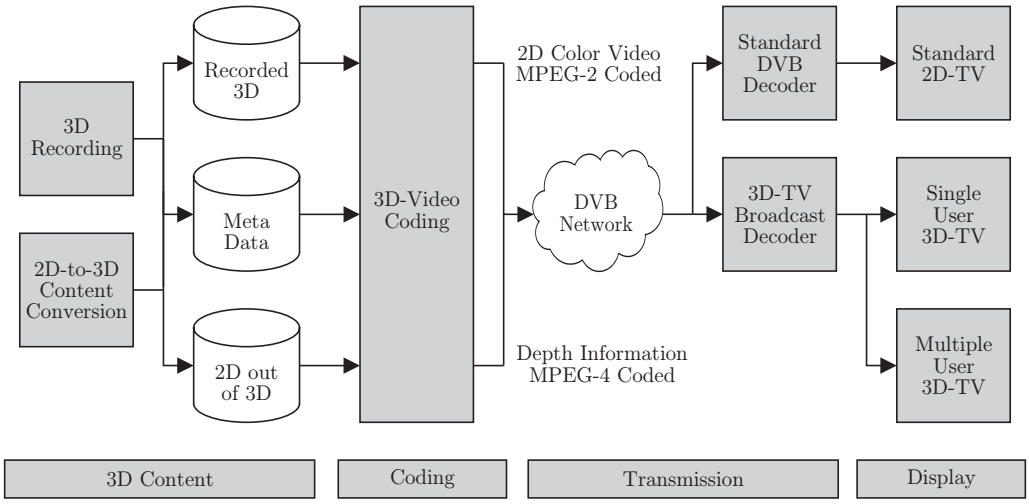

Figure 1. The ATTEST signal processing and data transmission chain. It consists of four different functional building blocks: 1) 3D content generation; 2) 3D video coding; 3) Transmission; 4) "Virtual" view generation and 3D display.

1.1. 3D Content Creation

For the generation of future 3D content two complementary approaches are anticipated. In the first case, novel three-dimensional material is created by simultaneously capturing video and associated per-pixel depth information with an active range camera such as the so-called Zcam™ developed by 3DV Systems.[3] Such devices usually integrate a high-speed pulsed infrared light source into a conventional broadcast TV camera and they relate the *time of flight* of the emitted and reflected light walls to direct measurements of the depth of the scene. However, it seems clear that the need for sufficient high-quality, three-dimensional content can only partially be satisfied with new recordings. It will therefore be necessary – especially in the introductory phase of the new broadcast technology – to also convert already existing 2D video material into 3D using so-called "structure from motion" algorithms.[4,5] On principle, such (offline or online) methods process one or more monoscopic color video sequences to: (a) establish a dense set of image point correspondences from which information about the recording camera as well as the 3D structure of the scene can be derived, or (b) infer approximate depth information from the relative movements of automatically tracked image segments*.

Whatever 3D content generation approach is used in the end, the outcome in all cases consists of regular 2D color video in European digital TV format (720×576 luminance pels, 25 Hz, interlaced) and an accompanying depth-image sequence with the same spatio-temporal resolution. Each of these *depth-images* stores depth information as 8-bit grayvalues with the graylevel 0 specifying the furthest value and the graylevel 255 defining the closest value (see Fig. 2). To translate this data representation format to real, metric depth values – which are required for the "virtual" view generation (see also Sect. 3 and 4) – and to be flexible with respect to 3D scenes with different depth characteristics, the grayvalues are normalized to two main depth clipping planes. The *near clipping plane* Z_{near} (graylevel 255) defines the smallest metric depth value Z that can be represented in the particular depth-image. Accordingly, the *far clipping plane* Z_{far} (graylevel 0) defines the largest representable metric depth value. In case of a linear quantization of depth, all other values can simply be calculated from these two extremes as:

$$Z = Z_{far} + \nu \cdot \frac{Z_{near} - Z_{far}}{255} \quad \text{with } \nu \in [0, \ldots, 255] \ , \tag{1}$$

where ν specifies the respective graylevel value.

*It must be noted that there exist a large number of other methods to generate 3D content – either in real-time or in an offline process – such as for example the joint 3D analysis of video sequences captured by synchronized multi-camera systems.[6] However, these approaches weren't explicitly considered within the ATTEST project.

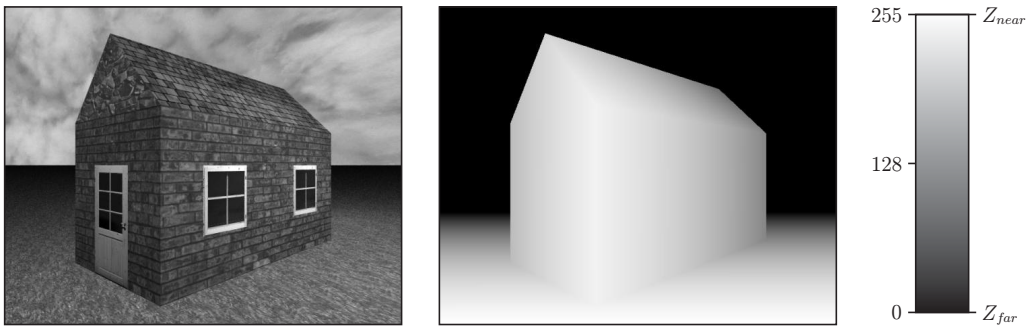

Figure 2. The ATTEST data representation format. It consists of: (a) Regular 2D color video in European digital TV format (720×576 luminance pels, 25 Hz, interlaced); (b) Accompanying 8-bit depth-images with the same spatio-temporal resolution. The depth-images are normalized to a near clipping plane Z_{near} and a far clipping plane Z_{far}.

1.2. 3D Video Coding

To provide the future 3D-TV viewers with the three-dimensional content, the monoscopic color video and the associated per-pixel depth information have to be compressed and transmitted over the conventional 2D digital TV broadcast infrastructure. To ensure the required backwards-compatibility with existing 2D-TV set-top boxes, the basic 2D color video has to be encoded using the standard MPEG-2 tools currently required by the DVB (Digital Video Broadcast) project in Europe. The supplementary depth-images, on the other hand side, can be compressed using any of the newer, more efficient additions to the MPEG family of standards such as MPEG-4 Visual or Advanced Video Coding (AVC).[7,8]

1.3. Transmission

The Digital Video Broadcast (DVB) project, a consortium of industries and academia responsible for the definition of today's 2D digital TV broadcast infrastructure in Europe, requires the use of the MPEG-2 *Systems Layer* specification for the distribution of audio-visual data via cable (DVB-C), satellite (DVB-S) or terrestrial (DVB-T) transmitters. Because of its almost universal acceptance and world-wide use, it is of major importance for any future 3D-TV system, to also build its distribution services on this transport technology. This was realized from the start by the ATTEST project consortium and explicitly considered during the design of the utilized three-dimensional data representation format.

1.4. "Virtual" View Generation and 3D Display

At the receiver side of the proposed ATTEST system, the transmitted data is decoded in a 3D-TV set-top box to retrieve the decompressed color video- and depth-image sequences (as well as the additional meta data). From this data representation format, a depth-image-based rendering (DIBR) algorithm generates "virtual" left- and right-eye views for the three-dimensional reproduction of a real-world scene on a stereoscopic- or autostereoscopic, single- or multiple user 3D-TV display. The backwards-compatible design of the system ensures that viewers that don't want to invest in a full 3D-TV set are still able to watch the two-dimensional color video without any degradations in quality using their existing digital 2D-TV set-top boxes and displays.[9,10]

2. A COMPARISON WITH "STEREOSCOPIC" VIDEO

Compared with the classical approach of "stereoscopic" video, the proposed 3D-TV system has a number of advantages, with the most important being the following:

+ The 3D reproduction can be adjusted to a wide range of different stereoscopic displays and projection systems. As the required left- and right-eye views are only generated at the 3D-TV receiver, their appearance in terms of 'perceived depth' can be adapted to the particular viewing conditions. This allows to provide the viewer with a customized 3D experience that is comfortable to watch on any kind of stereoscopic- or autostereoscopic 3D-TV display.[9,10]

+ 2D-to-3D conversion techniques based on "structure from motion" approaches can be used to generate the required depth information for already recorded monoscopic video material.[4, 5, 11] This is a very important point, as it seems clear that the success of any future 3D-TV broadcast system will depend to a great extent on the timely availability of sufficient interesting and exciting 3D video material.[1]

+ Head-motion parallax (HMP) could – on principle – be supported to provide an additional extrastereoscopic depth cue. This would also eliminate the well-known "shear-distortions" that are usually experienced with stereoscopic- or autostereoscopic 3D-TV systems.[12] In addition to that, this feature could equally well be used to provide an increased sensation of depth on conventional, monoscopic 2D-TV displays.[9, 10]

+ Due to the local smoothness characteristics of most "real-world" objects surfaces as well as due to the additional Gaussian filtering that is used to deal with the disocclusion problem (see also Sect. 5.3), the per-pixel depth information doesn't contain a lot of high frequency components. Therefore, it can be compressed much more efficiently than an additional color video channel (which would be required to represent the second view in a conventionally recorded stereoscopic image). This feature makes it possible to introduce a 3D-TV service that builds on the ATTEST concept with only a very small transmission overhead (below 10-20% of the basic color video bitrate) compared to today's conventional 2D digital TV.

+ Photometrical asymmetries, e. g. in terms of brightness, contrast or color, between the left- and the right-eye view, which can destroy the stereoscopic sensation,[13] are eliminated from the first, as both views are effectively synthesized from the same original image.

+ The system allows the viewer to adjust the reproduction of depth to suit his/her personal preferences – much like every conventional 2D-TV set allows the viewer to adjust the color reproduction by means of a (de-)saturation control.[1] This is an important system feature taken into account the fact that there is a difference in depth appreciation over age groups. A recent study conducted by Norman *et al.* for example demonstrated that older adults were less sensitive to perceiving stereoscopic depth than younger adults, in particular, when screen parallax was higher.[14]

+ The ATTEST data representation format of monoscopic video plus associated per-pixel depth information is ideally suited to facilitate 3D post-processing. It enables automatic object segmentation based on depth-keying and allows for an easy integration of synthetic 3D objects into "real-world" sequences (augmented reality).[15] This is an important prerequisite for advanced television features such as 'virtual advertisement' as well as for all kinds of real-time 3D special effects.

Despite all the just-mentioned advantages, it must not be concealed that the ATTEST concept also has a number of potential disadvantages that must be taken into serious consideration during the design of a commercially successful 3D-TV system:

– The quality of the "virtual" stereoscopic views surely depends on the accuracy of the per-pixel depth values of the original imagery. Therefore, it must be examined very carefully how different depth-image artifacts, which might be either due to the 3D content generation or result from the compression and transmission, translate into visually perceivable impairments in the synthesized images (see also Sect. 6 and 7).

– An inherent problem of the ATTEST approach on 3D-TV is due to the fact that areas that should be visible in the "virtual" left- and right-eye views are occluded in the original image. This requires the development of suitable "hole-filling" techniques as well as an experimental assessment of the typical artifacts that these methods introduce during the stereoscopic view synthesis (see also Sect. 5).

– Atmospheric effects like fog or smoke, semi-transparent objects like certain types of glass or plastic as well as view-dependent effects like shadows or reflections can not be handled adequately by the described ATTEST concept. However, recent work on image-based rendering (IBR) as well as experiences from traditional Computer Graphics (CG) seem to indicate that this might not be a problem in most situations.

– The complexity of a set-top box that implements the ATTEST approach on 3D-TV is higher than the complexity of a conventional, "stereoscopic" video system as, in addition to the decoding of two video streams, the integration of high-quality, real-time depth-image-based rendering (DIBR) algorithms is required.

3. DEPTH-IMAGE BASED RENDERING

Depth-image-based rendering (DIBR) is the process of synthesizing "virtual" views of a scene from still- or moving color images and associated per-pixel depth information.[16,17] Conceptually, this novel view generation can be understood as the following two-step process: At first, the original image points are reprojected into the 3D world, utilizing the respective depth data. Thereafter, these 3D space points are projected into the image plane of a "virtual" camera, which is located at the required viewing position. The concatenation of reprojection (2D-to-3D) and subsequent projection (3D-to-2D) is usually called *3D image warping* in the Computer Graphics (CG) literature and will be derived mathematically in the following paragraph.

3.1. 3D Image Warping

Consider a system of two cameras and an arbitrary 3D space point M with the projections m and m' in the first-, resp. the second view. Under the assumption that the world coordinate system equals the camera coordinate system of the first camera, the two *perspective projection* equations result to:

$$\tilde{\mathbf{m}} \cong \mathbf{A}\mathbf{P}_n\tilde{\mathbf{M}} \tag{2}$$

$$\tilde{\mathbf{m}}' \cong \mathbf{A}'\mathbf{P}_n\mathbf{D}\tilde{\mathbf{M}}\,, \tag{3}$$

where $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{m}}'$, resp. $\tilde{\mathbf{M}}$ symbolize the two 2D image points, resp. the 3D space point in homogeneous notation and the symbol $\cong$ denotes 'equality up to a non-zero scale-factor'.[4,18] The 4×4 matrix $\mathbf{D}$ contains the rotation $\mathbf{R}$ and the translation $\mathbf{t}$ that transform the 3D point from the world coordinate system into the camera coordinate system of the second view and the 3×3 matrices $\mathbf{A}$ and $\mathbf{A}'$ specify the intrinsic parameters of the first-, resp. the second camera. Finally, the 3×4 identity matrix $\mathbf{P}_n$ designates the so-called normalized perspective projection matrix.

Rearranging Eq. (2) gives an affine representation of the 3D space point M that is, however, still dependent on its depth value Z:

$$\mathbf{M} = Z\mathbf{A}^{-1}\tilde{\mathbf{m}}\,. \tag{4}$$

Substituting Eq. (4) into (3) then leads to the classical affine *disparity equation*, which defines the depth-dependent relation between corresponding points in two perspective views of the same 3D scene:

$$Z'\tilde{\mathbf{m}}' = Z\mathbf{A}'\mathbf{R}\mathbf{A}^{-1}\tilde{\mathbf{m}} + \mathbf{A}'\mathbf{t}\,. \tag{5}$$

This disparity equation can also be considered as a 3D image warping formalism, which can be used to generate an arbitrary novel view from a known reference image. This only requires the definition of the position and orientation of a "virtual" camera relative to the reference camera as well as the declaration of the "virtual" camera's intrinsic parameters. Then, if the depth values of the corresponding 3D space points are known for every pixel of the original image, the "virtual" view can be synthesized by applying Eq. (5) to all original image points. (The "real-world" problems of image resampling, resolving the visibility problem and handling of disocclusions in the novel view will be considered later in Sect. 5.)

4. STEREOSCOPIC IMAGE CREATION

On a stereoscopic- or autostereoscopic 3D-TV display, two slightly different perspective views of a 3D scene are reproduced (quasi-)simultaneously on a joint image plane. The horizontal differences between these left- and right-eye views, the so-called *screen parallax* values, are interpreted by the human brain and the two images are fused into a single, three-dimensional percept. In the ATTEST approach on 3D-TV, such stereoscopic images are not captured directly with a conventional stereo camera system, rather they are synthesized in real-time from monoscopic color video and associated per-pixel depth information.

4.1. Shift-Sensor Algorithm

The following section explains in detail how stereoscopic images can efficiently be generated from monoscopic color video and associated per-pixel depth information using a slightly altered variant of the already-introduced 3D warping equation (5). The best way to start with this explanation is to look at the two different configurations that are usually utilized in "real", high-quality stereo cameras.[19] As can be seen from Fig. 3, both setups consist of a pair of cameras – one for the left-eye view and one for the right-eye view – that are separated by the so-called *interaxial distance* t_c. The major difference between both designs, however, lies in the method that is used to establish the so-called *zero-parallax setting (ZPS)*, i. e. the part of the 3D scene that is going to be reproduced exactly on the display screen. In the "toed-in" approach, shown on the left side of the graphic (a), a *point of convergence* at Z_c is chosen by a joint inward-rotation of the left- and right camera of the setup. In the so-called *shift-sensor* approach, shown on the right side of the graphic (b), a *plane of convergence* at Z_c is established by a shift h of the camera's CCD sensors.

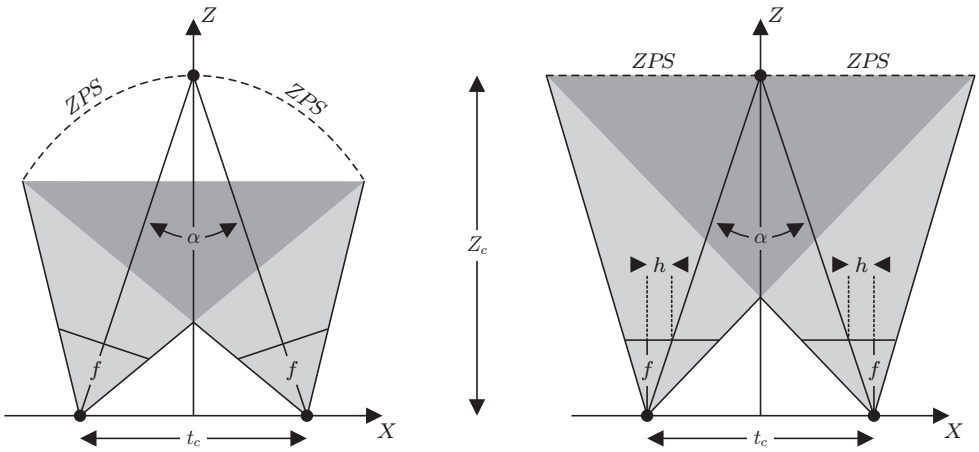

Figure 3. Different stereoscopic camera setups. (a) In the "toed-in" camera setup, a point of convergence at Z_c is established by a joint inward-rotation of the two cameras; (b) In the shift-sensor camera setup, a plane of convergence at Z_c is established by a shift h of the camera's CCD sensors.

While, technically, the "toed-in" approach is easier to realize in "real" stereo cameras, the shift-sensor approach is usually preferred because it doesn't introduce unwanted vertical differences – which are known to be a potential source of eye-strain – between the left- and the right-eye view.[20] Fortunately, this method is actually easier to implement with depth-image-based rendering (DIBR) as the required signal processing is only one-dimensional. All that is needed is the definition of two "virtual" cameras – one for the left eye and one for the right eye. With respect to the original view, these cameras are symmetrically displaced and their CCD sensors are shifted relative to the position of the lenses. Mathematically, this sensor shift can be formulated as a displacement of a camera's principal point c.[18] The *intrinsic parameters* of the two "virtual" cameras are therefore chosen to exactly correspond to the intrinsic camera parameters of the original view except for the horizontal shift h of the respective principal point. This can also be written as:

$$\mathbf{A}^* = \begin{bmatrix} \alpha_u & 0 & u_0 + h \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{A} + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} , \tag{6}$$

where the asterix symbol, which is used as a superscript here and in the following, should be substituted by either a single- or a double dash, e. g. $\mathbf{A}^*$ means either $\mathbf{A}'$ or $\mathbf{A}''$, to denote that the equation specifies the intrinsic parameters of either the left- or the right "virtual" camera.

Using the expression in Eq. (6) and taking into account that the movement of the two "virtual" cameras is restricted to be only translational with respect to the reference camera, i. e. $\mathbf{R} = \mathbf{I}$, where $\mathbf{I}$ is the 3×3 identity matrix, the following simplifications can be made in the general 3D warping equation (5):

$$\mathbf{A}^*\mathbf{R}\mathbf{A}^{-1} = \mathbf{A}^*\mathbf{A}^{-1} = \mathbf{I} + \begin{bmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} . \tag{7}$$

Inserting the simplified expression (7) into Eq. (5) yields the following reduced form of the general 3D warping equation:

$$Z^*\tilde{\mathbf{m}}^* = Z\left(\tilde{\mathbf{m}} + \begin{bmatrix} h \\ 0 \\ 0 \end{bmatrix}\right) + \mathbf{A}^*\mathbf{t} . \tag{8}$$

This expression can be simplified even more by taking into account that the only non-zero translational component needed to create a "virtual" shift-sensor camera setup is a horizontal translation t_x inside the focal plane of the original camera. With $t_z = 0$, it follows that the depth value of a 3D space point is the same in the world coordinate system – which was chosen to equal the camera coordinate system of the original view – and in the coordinate system of the "virtual" camera, i. e. $Z^* = Z$. Therefore Eq. (8) further reduces to:

$$\tilde{\mathbf{m}}^* = \tilde{\mathbf{m}} + \frac{\mathbf{A}^*\mathbf{t}}{Z} + \begin{bmatrix} h \\ 0 \\ 0 \end{bmatrix} \quad \text{with } \mathbf{t} = \begin{bmatrix} t_x \\ 0 \\ 0 \end{bmatrix} . \tag{9}$$

In this case, the affine pixel position (u, v) of each warped image point can simply be calculated as:

$$\begin{aligned} u^* &= u + \Delta u \qquad\qquad \text{, resp.} \quad v^* = v . \\ &= u + \frac{\alpha_u t_x}{Z} + h \end{aligned} \tag{10}$$

The horizontal camera translation t_x is defined to equal the half of the chosen interaxial distance t_c[†], with the direction of the movement given by:

$$t_x = \begin{cases} -\frac{t_c}{2} & : \quad \text{left-eye view} \\ +\frac{t_c}{2} & : \quad \text{right-eye view} \end{cases} . \tag{11}$$

As already described, the amount of the sensor shift h depends on the selected convergence distance Z_c and can be calculated by taking into account that for $Z = Z_c$ the horizontal component u^* of the simplified 3D warping equation (10) must be the same in the left- and in the right view, i. e. $u' = u''$, which leads to the following simple expression:

$$h = -t_x \frac{\alpha_u}{Z_c} , \tag{12}$$

where t_x is also defined by Eq. (11).

The expressions in Eq. (10) to (12) fully define a simplified 3D warping equation that can be used to efficiently implement a "virtual" shift-sensor stereo camera setup. Table 1 shows, how the resulting 3D reproduction is influenced by the choice of the three main system variables, i. e. by the choice of the interaxial distance t_c, the focal length f of the reference camera and the convergence distance Z_c. The respective changes in screen parallax values, perceived depth and object size are qualitatively equal to what happens in a "real" stereo camera when these system parameters are manually adjusted.

[†]This value is usually chosen to equal the average human eye separation of approximately 64 mm. However, for some 3D scenes a smaller or even larger t_c might be required to achieve the desired artistical 3D effect.

Parameter	+/−	Screen parallax	Perceived depth	Object size
Interaxial distance t_c	+	Increase	Increase	Constant
	−	Decrease	Decrease	Constant
Focal length f	+	Increase	Increase	Increase
	−	Decrease	Decrease	Decrease
Convergence distance Z_c	+	Decrease	Shift (forward)	Constant
	−	Increase	Shift (backwards)	Constant

Table 1. Effects of different stereo camera setup parameters. Qualitative changes in screen parallax values, perceived depth and object size when varying the interaxial distance t_c, the focal length f or the convergence distance Z_c of a "real" or "virtual" shift-sensor stereo camera setup (after Milgram and Krüger[21]).

5. IMPLEMENTATION DETAILS

To verify the concept of the ATTEST approach on 3D-TV, the described "virtual" shift-sensor algorithm was integrated in a real-time receiver demonstrator. The following sections provide some implementation details that are specific to the current version of this software.

5.1. Visibility

During the generation of a "virtual" view, it can happen that two different original image points are warped to the same location in the new image. This situation occurs when one of the corresponding 3D space points is occluded by the other one in the novel view. A very simple way to resolve this *visibility problem* during the 3D warp is to process the pixel of the original image in a so-called *occlusion-compatible warp order*.[16] This processing order only depends on the relative positioning of the "virtual" camera with respect to the original camera and is therefore independent of the 3D scene itself. The effect of adhering to it is that closer points are always warped later, thus automatically overwriting points further away. While in principle 18 different processing orders result for the most general form of the 3D warping equation (5), only two different orders have to be implemented for the described "virtual" shift-sensor algorithm. For the left-eye view, the columns of the original image are processed away from the right- towards the left image border. For the generation of the right-eye view, this warp order must simply be reversed.

5.2. Resampling

The warping of the original image points into the "virtual" view is only the first step of the synthesis work. Usually, the displaced points will not lie on the pixel raster of the output image and the novel view must be carefully *resampled*. Luckily, the described shift-sensor approach does not require a complex two-dimensional resampling, instead a more simple, one-dimensional processing can be used. This is a problem that has been dealt with extensively in the context of 2D image warping and/or morphing and, relatively recently, also in the context of 3D image warping. The main discriminating feature of the different approaches lies in the utilized interpolation kernel, e. g. nearest neighbor, linear, cubic convolution, cubic spline or sinc function, and thus in the resampling quality that can be achieved in case of *minifications* and *magnifications*. In the current implementation, an efficient "linear interpolation" resampling algorithm is utilized that had been developed by Wolberg for use with 2D image warping.[22] Some slight adjustments were necessary because the original algorithm doesn't allow for any *foldovers*, i. e. it requires the forward mapping function – in this case the 3D warping function – to be strictly monotonic. At the occurance of occlusions (see again Sect. 5.1), however, this requirement is not fulfilled anymore and some additional care has to be taken to ensure that the occluding edge is correctly antialiased.

5.3. Disocclusions and Hole-Filling

One major problem of the described depth-image-based rendering (DIBR) algorithm is due to the fact that areas, which are occluded in the original view, might become visible in any of the "virtual" left- and right-eye views, an event referred to as *exposure* or *disocclusion* in the Computer Graphics (CG) literature.[16,17] The question results, how these disocclusions should be treated during the view synthesis, as information about the previously occluded areas is neither available in the monoscopic color video nor in the accompanying depth-images?

One possible solution to the problem is provided by the so-called *layered depth-images* (LDIs). As an enhancement to the basic depth-image, LDIs allow to store more than one pair of associated color- and depth values for each pixel of the original image, with the number of layers typically depending on the scene complexity as well as the required synthesis quality.[23] With this data representation format, additional *hidden layer* information could be used to *fill-in* the exposed areas in the "virtual" left- and right-eye views. While this approach seems to be well suited from a synthesis quality point-of-view, it has the major disadvantage that its realization would very much complicate all parts of the concept (from content generation to coding, transmission and synthesis).

Because of the just-decribed drawback, the use of hidden layers has not been considered a viable option for the handling of disocclusions within the ATTEST project. Instead, techniques were developed and/or analyzed, which: (1) replace the missing image areas (holes) during the view synthesis with 'usefull' color information, or (2) preprocess the depth information in a way that no disocclusions appear in the "virtual" views. Synthesis examples for four different possible approaches are provided in Fig. 4 for a small display detail of the ATTEST test sequence 'Interview'. The first image (a) shows the well-known "rubber-sheet" artifacts that are typical for a linear color interpolation between scene foreground (head) and background (wall). The second image (b) visualizes the strip-like impairments that result from a simple extrapolation of the scene background. The artifacts in the third image (c) are due to a mirroring of background color information along the borders of the disocclusion. The fouth image (d) shows that visually less perceptable impairments can be achieved by preprocessing (smoothing) the depth information with a suitable Gaussian filter in a way that no disocclusions occur. While this must obviously lead to some geometric distortions of the displayed 3D space, the overall visual quality is usually higher than with the first three approaches. This method is therefore also used in the current implementation of the ATTEST 3D-TV receiver demonstrator software.

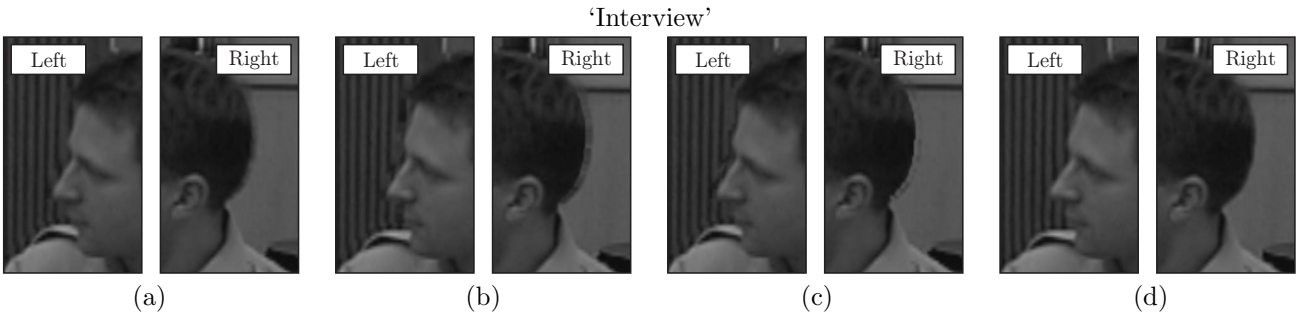

Figure 4. Synthesis artifacts resulting from the four different hole-filling algorithms. (a) A.1: Linear interpolation of foreground- and background image color; (b) A.2: Background color extrapolation; (c) A.3: Mirroring of background color information; (d) A.4: Preprocessing of depth information.

6. CODING OF 3D IMAGERY

To provide the future 3D-TV viewers with the three-dimensional content, the monoscopic color video and the associated per-pixel depth information are first compressed and then transmitted over the conventional 2D digital TV broadcast infrastructure. To ensure the required backwards-compatibility with existing 2D-TV set-top boxes, the basic 2D color video has to be encoded using the standard MPEG-2 tools currently required by the DVB (Digital Video Broadcast) project in Europe, while the supplementary depth-images can – on principle – be compressed using any of the newer, more efficient additions to the MPEG family of standards such as MPEG-4 Visual or Advanced Video Coding (AVC).[7,8]

The suitability of the different MPEG technologies for the efficient compression of depth-images was evaluated in a comparative coding experiment. The test group consisted of the following four codecs: a) the MPEG-2 reference model codec (TM-5); b) the Microsoft® MPEG-4 Visual reference model codec (MS-Ref.); c) a rate-distortion (R/D) optimized‡ MPEG-4 Visual codec developed at FhG/HHI (R/D opt.); d) the R/D optimized

‡Rate-distortion (R/D) optimization refers to the process of jointly optimizing both the objective 'image quality' and the required bitrate by systematically varying and testing different video encoder parameters.[24]

AVC reference model codec (v6.1a). The compression results for the two ATTEST test sequences 'Interview' and 'Orbi' are shown in Fig. 5 for typical broadcast encoder settings, i. e. a GOP (Group of Pictures) length equal to 12 with a GOP structure of $IBBPBBP\ldots$, by means of rate-distortion curves over a range of bitrates.

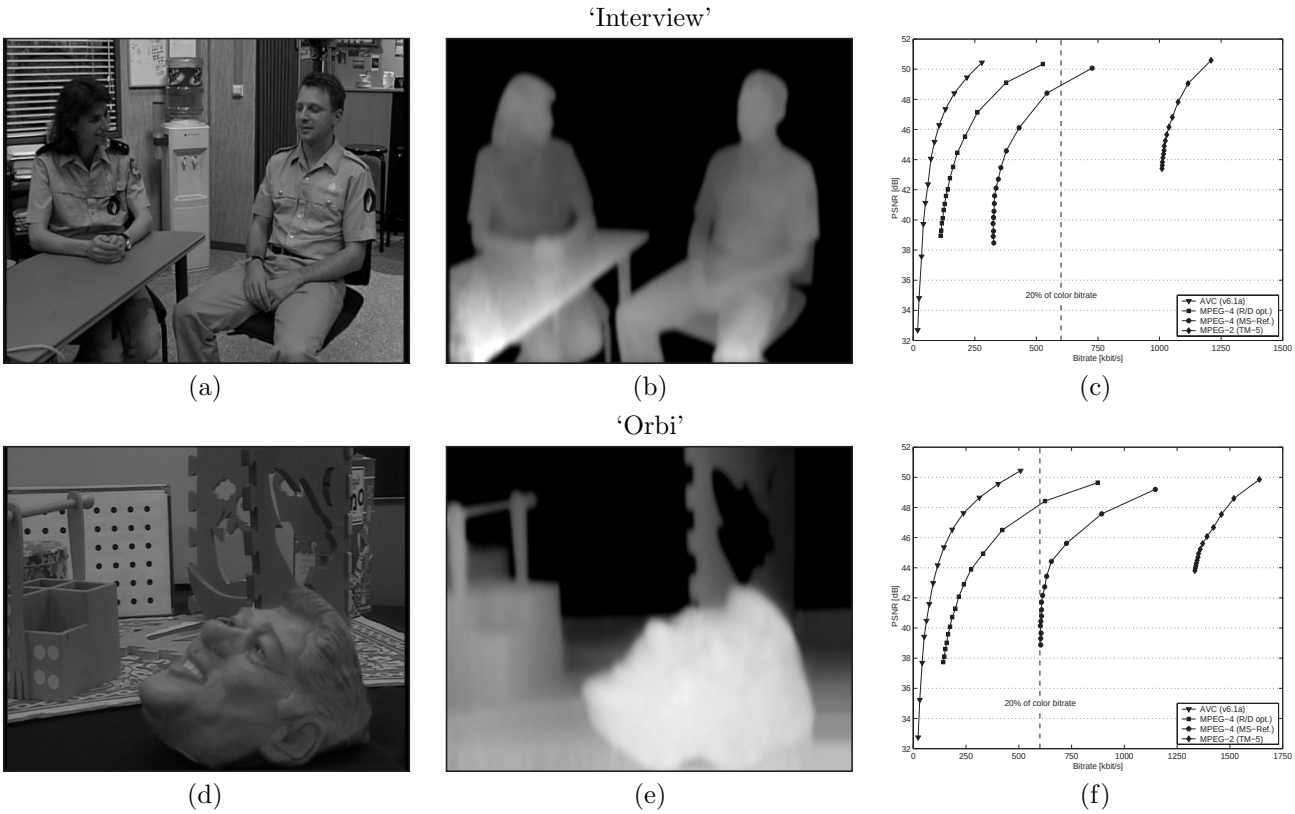

Figure 5. Test sequences 'Interview' and 'Orbi' with rate-distortion curves for four different MPEG codecs. (a,d) Monoscopic color video; (b,e) Accompanying per-pixel depth information; (c,f) Coding results shown as rate-distortion curves over a range of bitrates.

The two plots on the right side of the graphic (c,f) show, first of all, that AVC as well as MPEG-4 Visual are very well suited for the coding of per-pixel depth information (with AVC being even more efficient). The smoothness of the graylevel depth data (b,e) as well as the relatively slow camera-, resp. in-scene motion exhibited by these particular sequences lead to extremely high compression ratios. If a typical broadcast bitrate of 3 Mbit/s is assumed for the MPEG-2 encoded monoscopic color information (a,d), it can be followed from the R/D curves that the accompanying depth-images can be compressed to target rates significantly below 20% of this value. For example, AVC compression of the 'Interview' sequence at 105 kbit/s still leads to a very high PSNR of 46.29 dB. For the more complex 'Orbi' scene, this value can be reached at a bitrate of approximately 184 kbit/s. While it seems clear that these findings have to be confirmed with other, more challenging test data, the results nonetheless indicate that it will be possible to introduce the described new approach on 3D-TV with only a very minor transmission overhead compared to today's conventional 2D digital TV.

7. EXPERIMENTAL RESULTS

Figure 6 displays some further experimental results. Each of the two images on the left side of the graphic (a,d) shows AVC compressed depth information from one of the two test sequences 'Interview' and 'Orbi'. For the first scene the bitrate is equal to 105 kbit/s with a PSNR of 46.29 dB, for the second video the bitrate equals 115 kbit/s with a PSNR of 44.16 dB. The images confirm the objective distortion measures and shows that even at these very low data rates the visual quality of the displayed frames is only slightly degraded in comparison to

the original signals. The two images in the middle of the graphic (b,e) show overlaid "virtual" left- and right-eye views that were synthesized from the impaired depth-images using the before-described shift-sensor algorithm (see again Sect. 4.1). The system parameters of the "virtual" stereo camera setup were chosen such that the screen parallax values don't exceed a maximum of about 3% of the image width. The two images on the right side of the graphic (c,f) show the small synthesis errors (luminance) that are caused by the depth-image compression artifacts. Human-factors experiments conducted on a single-user, autostereoscopic 3D-TV display (lenticular lens raster) developed by FhG/HHI within the ATTEST project showed that the impaired synthesis results were visually indistinguishable from corresponding 3D sequences that were created with the original, uncompressed depth information. A more detailed description of these evaluations will be released in a separate publication.

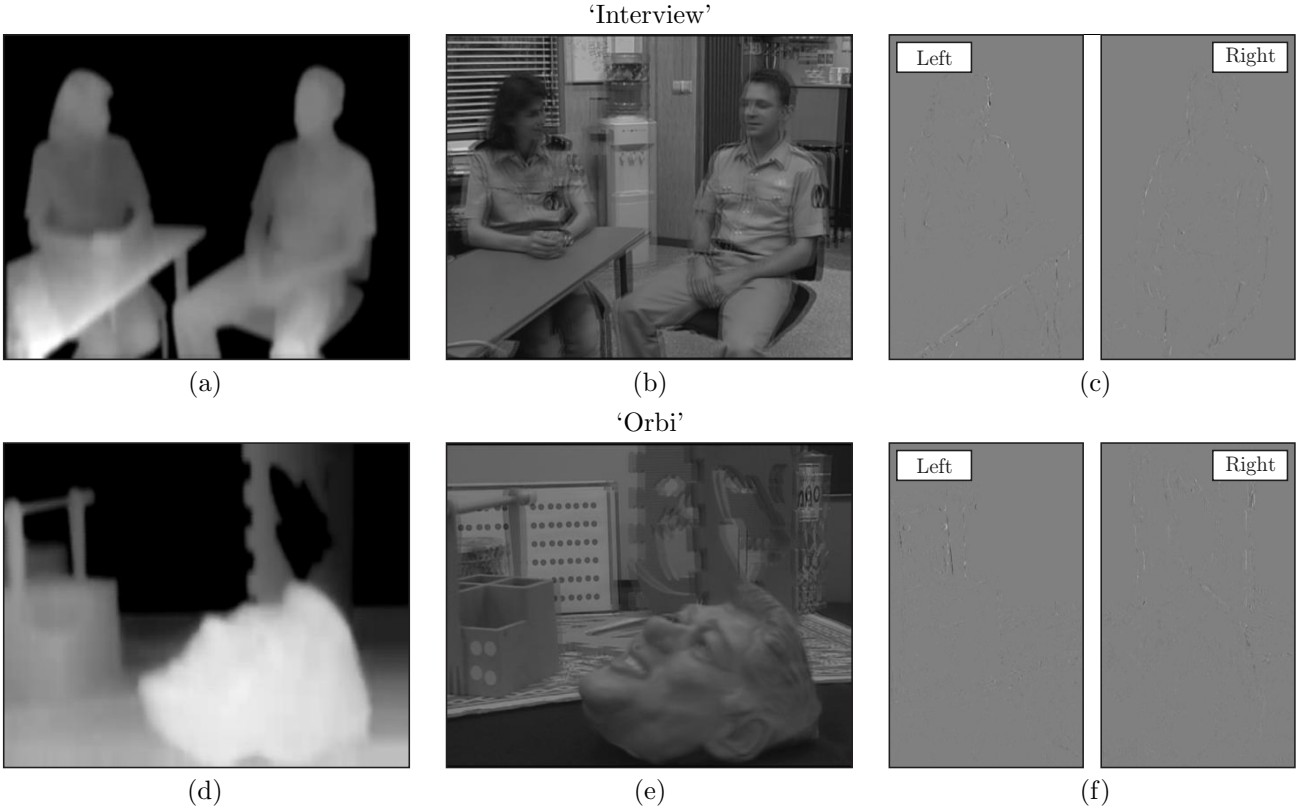

Figure 6. Coding and synthesis results for the 'Interview' and 'Orbi' test sequences. (a,d) AVC compressed per-pixel depth information; (b,e) Overlaid "virtual" left- and right-eye views; (c,f) Synthesis errors (luminance) caused by depth-image compression artifacts (5 × amplified for visualization purposes).

8. CONCLUSION

This paper provided details of a new approach on 3D-TV using depth-image-based rendering (DIBR). The given experimental results are very promissing and indicate that it would be possible to introduce the described 3D-TV scenario with only a very minor transmission overhead compared to today's conventional 2D digital TV.

ACKNOWLEDGMENTS

This work has been sponsored by the European Commission (EC) through their Information Society Technologies (IST) program under proposal No. IST-2001-34396. The author would like to thank the project officers as well as all project partners (Philips Research - The Netherlands, TU Eindhoven - The Netherlands, FhG/HHI - Germany, KU Leuven - Belgium, CERTH/ITI - Greece, 3DV Systems - Israel, De Montfort University - United Kingdom, VRT - Belgium) for their support and for their input to this publication.

REFERENCES

1. C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. A. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An Evolutionary and Optimised Approach on 3D-TV," in *Proc. of International Broadcast Conference '02*, pp. 357–365, (Amsterdam, The Netherlands), Sept. 2002.
2. C. Fehn, "A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR)," in *Proc. of Visualization, Imaging, and Image Processing '03*, pp. 482–487, (Benalmádena, Spain), Sept. 2003.
3. G. J. Iddan and G. Yahav, "3D Imaging in the Studio and Elsewhere ...," in *Proc. of SPIE Videometrics and Optical Methods for 3D Shape Measurements '01*, pp. 48–55, (San Jose, CA, USA), Jan. 2001.
4. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2000.
5. M. Pollefeys, "3D Modelling from Images." Tutorial given at European Conference on Computer Vision '00, (Dublin, Ireland), June 2000.
6. J. Mulligan and K. Daniilidis, "View-Independent Scene Acquisition for Tele-Presence," Technical Report 00-16, Computer and Information Science, University of Pennsylvania, July 2000.
7. ISO/IEC JTC 1/SC 29/WG 11, "Coding of Audio-Visual Objects – Part 2: Visual." ISO/IEC 14496-2:2001, Geneva, Switzerland, 2001.
8. ISO/IEC JTC 1/SC 29/WG 11, "Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC)." JVT Document E146d34, Geneva, Switzerland, 2002.
9. C. Fehn, E. Cooke, O. Schreer, and P. Kauff, "3D Analysis and Image-Based Rendering for Immersive TV Applications," *Signal Processing: Image Communication* **17**, pp. 705–715, Oct. 2002.
10. C. Fehn and P. Kauff, "Interactive Virtual View Video (IVVV) – The Bridge Between 3D-TV and Immersive TV," in *Proc. of SPIE Three-Dimensional TV, Video and Display '02*, pp. 14–25, (Boston, MA, USA), July 2002.
11. O. Faugeras, Q.-T. Luong, and T. Papadopoulo, *The Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications*, MIT Press, Cambridge, Massachussets, USA, 2001.
12. D. Runde, "How to Realize a Natural Image Reproduction Using Stereoscopic Displays With Motion Parallax," *IEEE Transactions on Circuits and Systems for Video Technology* **10**, pp. 376–386, Apr. 2000.
13. L. Lipton, *Foundations of the Stereoscopic Cinema – A Study in Depth*, Van Nostrand Reinhold, New York, NY, USA, 1982.
14. J. Norman, T. Dawson, and A. Butler, "The Effects of Age Upon the Perception of Depth and 3-D Shape From Differential Motion and Binocular Disparity," *Perception* **29**, pp. 1335–1359, Nov. 2000.
15. R. Gvili, A. Kaplan, E. Ofek, and G. Yahav, "Depth Keying," in *Proc. of SPIE Electronic Imaging '03*, (Santa Clara, CA, USA), Jan. 2003.
16. L. McMillan, *An Image-Based Approach to Three-Dimensional Computer Graphics*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1997.
17. W. R. Mark, *Post-Rendering 3D Image Warping: Visibility, Reconstruction, and Performance for Depth-Image Warping*. PhD thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, Apr. 1999.
18. G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*, Kluver Academic Publishers, Dordrecht, The Netherlands, 1996.
19. A. Woods, T. Docherty, and R. Koch, "Image Distortions in Stereoscopic Video Systems," in *Proc. of SPIE Stereoscopic Displays and Applications '93*, pp. 36–48, (San Jose, CA, USA), Feb. 1993.
20. W. A. IJsselsteijn, H. de Ridder, and J. Vliegen, "Stereoscopic Filming Parameters and Display Duration on the Subjective Assessment of Eye Strain," in *Proc. of SPIE Stereoscopic Displays and Virtual Reality Systems '00*, pp. 12–22, (San Jose, CA, USA), Apr. 2000.
21. P. Milgram and M. Krüger, "Adaptation Effects in Stereo Due to On-line Changes in Camera Configuration," in *Proc. of SPIE Stereoscopic Displays and Applications '92*, pp. 122–134, (San Jose, CA, USA), Feb. 1992.
22. G. Wolberg, *Digital Image Warping*, IEEE Computer Society Press, Los Alamitos, CA, USA, 1990.
23. J. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered Depth Images," in *Proc. of ACM SIGGRAPH '98*, pp. 231–242, (Orlando, FL, USA), July 1998.
24. I. E. G. Richardson, *Video CODEC Design*, John Wiley and Sons, 2002.